

III INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND NANOTECHNOLOGY - 2017



# ITNT - 2017

75th anniversary of the Samara University  
24-27 April 2017, Samara

Samara National Research University  
Image Processing Systems Institute of the RAS - branch of the Federal Scientific Research  
Centre "Crystallography and Photonics" of Russian Academy of Sciences

INFORMATION TECHNOLOGY AND NANOTECHNOLOGY (ITNT-2017)

Section: Computer Optics and Nanophotonics  
Section: Image Processing and Geoinformation Technology, and Information Security  
Section: High-Performance Computing  
Section: Data Science  
Section: Mathematical Modeling  
Section: Computer Modeling

Collection of selected papers of the III International Conference on Information Technology and  
Nanotechnology  
(Samara, 2017, 25-27 April)

978-5-6047396-2-4  
Publisher: IP Zaitsev V.D.

Samara  
2017



Collection of selected papers of the III International Conference on Information Technology and Nanotechnology (ITNT-2017). Section Computer Optics and Nanophotonics. Section Image Processing and Geoinformation Technology, and Information Security. Section High-Performance Computing. Section Data Science. Section Mathematical Modeling. Section Computer Modeling. Samara, Individual Proprietor Zaitsev V.D. 2017, 25-27 April. ISBN 978-5-6047396-2-4.

Compilers of the volume: Roman Skidanov, Vladislav Myasnikov, Vladislav Sergeyev, Victor Fedoseev, Vladimir Fursov, Vladimir Voevodin, Michael Sobolewski, Sergey Popov, Vladimir Sobolev, Dmitry Savelyev

Issuing editor: Denis Kudryashov

Publisher: Individual Proprietor Zaitsev V.D.

Copyright © 2017 for the individual papers by the papers' authors. Copying permitted only for private and academic purposes. This volume is published and copyrighted by its editors.

The conference is a forum for leading researchers from all over the world aimed to discuss the latest advances in the basic and applied research in the field of Information Technology and Nanotechnology. It is also aimed to attract young people to advanced scientific research and share the latest trends in training and research programs for future ITNT specialists.

**Table of Contents**  
Computer Optics and Nanophotonics

1. Multilayer dielectric stack Notch filter for 450-700 nm wavelength spectrum M.A. Butt, S.A. Fomchenkov, S.N. Khonina.....	1-4
DOI: 10.18287/1613-0073-2017-1900-1-4	
2. Cold mirror based on High-Low-High refractive index dielectric materials V.V. Elyutin, M.A. Butt, S.N. Khonina.....	5-9
DOI: 10.18287/1613-0073-2017-1900-5-9	
3. An algorithm for correcting X-ray image distortions caused by central projection A.V. Ustinov , N.Yu. Ilyasova, N.S. Demin.....	10-15
DOI: 10.18287/1613-0073-2017-1900-10-15	
4. Prognostic modeling of the curvilinear graphene selective hydrogenation process for the formation of optical scheme components for nanophotonics Hussein Safaa Mohammed Ridha, S.I. Kharitonov, V.S. Pavelyev.....	16-19
DOI: 10.18287/1613-0073-2017-1900-16-19	
5. Methods for creation of diffractive intraocular lenses A.V. Gornostay.....	20-27
DOI: 10.18287/1613-0073-2017-1900-20-27	
6. Wavefront aberration analysis with a multi-order diffractive optical element P.A. Khorin, S.A. Degtyarev.....	28-33
DOI: 10.18287/1613-0073-2017-1900-28-33	
7. Formation of probing radiation for investigating a uniaxial x-cut crystal with the help of an aperiodic diffractive axicon V.D. Pararin, S.V. Karpeev.....	34-37
DOI: 10.18287/1613-0073-2017-1900-34-37	
8. The elaboration of numerical simulation error light pulse propagation in a waveguide of circular cross-section A.A. Degtuarev, A.V. Kukleva.....	38-42
DOI: 10.18287/1613-0073-2017-1900-38-42	
9. Spectra and field distribution of photonic-crystal structure with inclusions of metal nanoparticles I.A. Glukhov, S.G. Moiseev.....	43-47
DOI: 10.18287/1613-0073-2017-1900-43-47	
10. Microexplosions polysterene microparticles on substrate covered by aluminum V.S. Vasilev, R.V. Skidanov.....	48-51
DOI: 10.18287/1613-0073-2017-1900-48-51	
11. Generation of regular optical pulses in VCSELs below the static threshold A.A. Krents, N.E. Molevich, D.A. Anchikov, S.V. Krestin.....	52-54
DOI: 10.18287/1613-0073-2017-1900-52-54	
12. Experimental observing of transformation Bessel beam spreading along axis of crystal during wavelength changes V.S. Vasilev1, V.V. Podlipnov.....	55-59
DOI: 10.18287/1613-0073-2017-1900-55-59	
13. Amplitude and polarization transformations of the Bessel beam as it passes through an anisotropic crystal perpendicular to the axis of the crystal A.V. Glazkova, M.V. Zablovskaya, V.V. Podlipnov.....	60-63
DOI: 10.18287/1613-0073-2017-1900-60-63	
14. Raman spectra analysis of human blood protein fractions using the projection on latent structures method A.A. Lykina, D.N. Artemyev, I.A. Bratchenko, Yu.A. Khristoforova, O.O. Myakinin, T.P. Kuzmina, I.L. Davydkin, V.P. Zakharov.....	64-68
DOI: 10.18287/1613-0073-2017-1900-64-68	
15. Intelligent learning and testing system for students training in the problem area of nanotechnology and microsystem engineering.....	69-73
D. Lyapunov, A. Yankovskaya, Y. Dementyev, K. Negodin DOI: 10.18287/1613-0073-2017-1900-69-73	



16. Sensitive detection of nitrogen dioxide using gold nanoparticles decorated single walled carbon nanotubes S. Kumar, V. Pavelyev, P. Mishra, N. Tripathi.....	74-77
DOI: 10.18287/1613-0073-2017-1900-74-77	
17. Physicochemical properties of submicron and nanoscale particles of Ga and AlGa alloy obtained by laser ablation in a liquid.....	78-83
V.S. Kazakevich, P.V. Kazakevich, P.S. Yaresko, D.A. Kamynina	
DOI: 10.18287/1613-0073-2017-1900-78-83	
18. Nanocrystalline silicon and silicon carbide optical properties D. Lizunkova, N. Latukhina, V. Chepurnov, V. Paragin.....	84-89
DOI: 10.18287/1613-0073-2017-1900-84-89	
19. The combination of Raman spectroscopy and Autofluorescence analysis for estimation of blood and urine homeostasis L.A. Shamina, I.A. Bratchenko.....	90-93
DOI: 10.18287/1613-0073-2017-1900-90-93	
20. Deposition of Zinc Oxide thin film layer with the help of modified sputtering system S.A. Fomchenkov, S.D. Poletaev.....	94-96
DOI: 10.18287/1613-0073-2017-1900-94-96	
21. Approximation of optical signals by the vortex eigenfunctions of the double finite Hankel transform M.S. Kirilenko.....	97-100
DOI: 10.18287/1613-0073-2017-1900-97-100	
22. Development of mathematical model of laser treatment heat processes using diffractive optical elements S.P. Murzin, A.Yu. Tisarev, M.V. Blokhin, S.A. Afanasiev.....	101-105
DOI: 10.18287/1613-0073-2017-1900-101-105	
23. Investigation of methods for the formation of multicolored images reconstructed with protective holograms L.A. Nayden, I.K. Tsyganov, S.B. Odinkov.....	106-107
DOI: 10.18287/1613-0073-2017-1900-106-107	
24. On model of microstructure formation during selective laser melting of metal powder bed F.Kh. Mirzade, A.V. Dubrov.....	108-116
DOI: 10.18287/1613-0073-2017-1900-108-116	
25. Research of the effect of aberrations on image quality in optical systems A.V. Kozhevnikov.....	117-121
DOI: 10.18287/1613-0073-2017-1900-117-121	
26. Current problems of development of the journal of <i>Computer Optics</i> D.V. Kudryashov.....	122-125
DOI: 10.18287/1613-0073-2017-1900-122-125	
27. Single mode ZnO/Al <sub>2</sub> O <sub>3</sub> Strip loaded waveguide at 633 nm visible wavelength M.A. Butt, E.S. Kozlova.....	126-129
DOI: 10.18287/1613-0073-2017-1900-126-129	
28. Efficient generation of a perfect optical vortex by using a phase optical element V.V. Kotlyar, A.A. Kovalev, A.P. Porfirev.....	130-135
DOI: 10.18287/1613-0073-2017-1900-130-135	
29. Symmetric encryption algorithm using “twisted” light S. A. Burlov1, A. V. Gorokhov.....	136-139
DOI: 10.18287/1613-0073-2017-1900-136-139	

# Preface

Roman Skidanov<sup>1</sup>, Dmitry Savelyev<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

Session «Computer Optics and Nanophotonics» was held at the 3rd International Conference on Information Technology and Nanotechnology - 2017 (ITNT-2017) in Samara, Russia, April 25–27, 2017 (<http://ru.itnt-conf.org/itnt17ru/>).

The goal of the ITNT-2017 Conference was to discuss problems of fundamental and applied research in information technology and nanotechnology, including:

- Computer Optics;
- Diffractive Nanophotonics;
- Image Processing;
- Computer Vision;
- Mathematical Modeling;
- Data Science.

Scientists from Austria, Belarus, Bulgaria, Denmark, Germany, Great Britain, India, Iraq, Mexico, Moldova, Russia, Spain, USA, and Finland presented over 330 reports at the ITNT-2017 Conference.

The main proceedings of the conference will published in *Procedia Engineering* (Elsevier BV). The proceedings of the seminar, not included in *Procedia Engineering*, were selected for this volume.

We are grateful to everybody who has contributed to the seminar and look forward to meeting you again at future events. Heartfelt thanks are due to all authors, reviewers and delegates. A special thank you is due to the team of organizers for making the seminar successful and this publication possible.

## Guest Editors

- Roman Skidanov, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics", Russian Academy of Sciences, Samara, Russia
- Denis Kudryashov, Samara National Research University, Samara, Russia

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov – Samara National Research University, Russia

## Vice-chairs

- Vladimir Bogatyrev – Samara National Research University, Russia
- Nikolay Kazanskiy – Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Russia
- Eduard Kolomiets – Samara National Research University, Russia
- Alexander Kupriyanov – Samara National Research University, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Russia



# Multilayer dielectric stack Notch filter for 450-700 nm wavelength spectrum

M.A. Butt<sup>1</sup>, S.A. Fomchenkov<sup>1,2</sup>, S.N. Khonina<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

In this work, a multilayer dielectric optical notch filters design is proposed based on TiO<sub>2</sub> and SiO<sub>2</sub> alternating layers. Titanium dioxide (TiO<sub>2</sub>) is selected for its high refractive index value (2.5) and Silicon dioxide (SiO<sub>2</sub>) as a low refractive index layer (1.45). These filters are conventionally envisioned for overpowering of powerful laser beams in research experiments, to obtain good signal-to-noise ratios in Raman laser spectroscopy. It is precarious that light from the pump laser should be blocked. This is attained by inserting a notch filter in the detection channel of the setup. In addition to spectroscopy, notch filters are also useful in laser-based fluorescence instrumentation and biomedical laser systems. The designed filter shows a high quality with an average transmission of more than 90% in 450-535 and 587-700 nm bandwidths. And a stop band region between 536-586 nm shows a transmission of 3% only with an optical density of greater than 3, which makes it a promising element to be used as a notch filter.

*Keywords:* Notch filter; Optical density; Distributed Bragg Reflector (DBR); visible spectrum

---

## 1. Introduction

Thin film optics is well-established technology. Many devices such as band pass filters, band-stop filters, polarizers and reflectors are realized with the help of multilayer dielectric thin films [1-4]. Thin films coatings have also been used to increase both colour and energy efficiency of glass and as reflecting mirrors coatings. However the application of single layer thin films has increased, there are a number of applications which require multilayer films that combine the attractive properties of numerous materials. Some of the important applications of multilayer films are in the design of computer disks, optical reflectors, antireflection coating, optical filters, and solar cells among others. An optical filter is an element or material which is purposefully used to change the spectral intensity distribution or the state of polarization of the electromagnetic radiation incident on it. The change in the spectral intensity distribution may or may not depend on the wavelength. The filter possibly will act in transmission, in reflection, or both. Notch filters are usually known as band-stop or band-rejection filters which are designed to transmit most of the wavelengths with the low-intensity loss while diminishing the light within a specific wavelength range to a very low level. These filters are conventionally proposed for overpowering of powerful laser beams in research experiments to obtain good signal-to-noise ratios in Raman laser spectroscopy. It is precarious that light from the pump laser should be blocked. This is attained by inserting a notch filter in the detection channel of the setup. In addition to spectroscopy, notch filters can also be used in laser-based fluorescence instrumentation and biomedical laser systems. They are also used for eye protection and as a camera accessory. These filters contain alternating layers of high (H) and low (L) refractive index materials with precise thicknesses with good knowledge about their refractive index and absorptions. Several multilayer coatings are deposited onto a transparent substrate. Both the multilayer and substrate contribute to the total performance of the filter. Layers made of oxides are, as a rule, harder than those made of fluorides, sulphides or semiconductors. Therefore, they are ideal to be used on unprotected surfaces. Semiconductor materials should be avoided in filters which have to be used over a wide range of temperatures because their optical constants can change considerably. Distributed Bragg Reflectors (DBRs) work on the principle of multiple reflections between high and low index materials interface. It has a  $\lambda/4$  thickness of the central wavelength. The high reflection region of a DBR is known as the DBR stopband and can be attained by the refractive index contrast between the constituent layers. A broad stop band can be realized by using high index contrast thin films. The schematic of the DBR is shown in figure 1.

In this work, the design of a Notch filter based on TiO<sub>2</sub>/SiO<sub>2</sub> is proposed at a central wavelength of 561 nm with an FWHM of 50 nm. Titanium dioxide (TiO<sub>2</sub>) is selected for its high refractive index value (2.5)[5] and Silicon dioxide (SiO<sub>2</sub>) as a low refractive index layer (1.45)[5]. TiO<sub>2</sub> is a vital dielectric material with a wide band-gap energy and high refractive index that can make it useful in the fabrication of multilayer thin films due to its high optical properties. For instance, its high transmittance and high refractive index in the visible region (380-760 nm) make it valuable to be employed in the production of the optical filter and window glazing [6, 7].

In the designing of optical filters, the behaviour of the entire multilayer system is anticipated on the basis of the properties of the individual layers in the stack [8]. Hence to attain the optimum performance, it is important to optically characterize and accurately determine the thickness of the individual layers. We designed this filter with a less possible number of layers with high transmission in pass band region and high reflection is obtained in the stop band. Open-source software, Open Filters, is

used in this work to design and optimize the required filter. Transmission and reflection properties of interference filters are dependent on materials refractive index and layer thickness of materials. Open filter calculates optical properties of filters. It uses transfer matrix method to calculate the transmission and reflection properties of filters based on the absorption and materials refractive indices [9]. Optimization techniques are available in this software like needle synthesis (Adding an extra

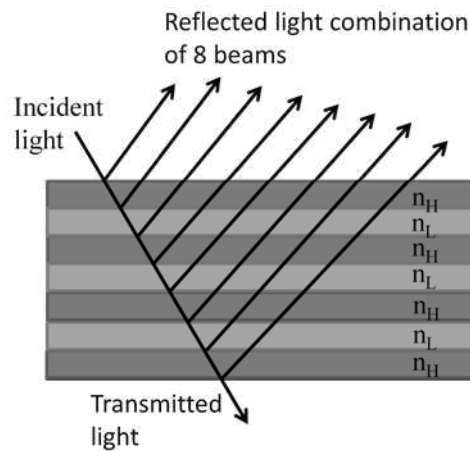


Fig. 1. Schematic of Distributed Bragg Reflector (DBR).

layer to give targeted transmission).

## 2. Optical density of the notch filter

A filter plate made of an isotropic material with smooth and parallel surfaces, the transmittance depends on the thickness, optical constants of the material, the angle of incidence and polarization state of the incident light, and the degree of coherence between multiple reflected waves [10, 11]. Optical density (OD) is used to see the blocking specification of a filter and is associated with the amount of energy transmitted through it. It uses a logarithmic scale to describe the transmission of light through a highly blocked optical filter, particularly useful when the transmission is extremely small. A high optical density value indicates very low transmission of light and low optical density indicates high transmission. For instance,  $OD=1$  relates to a transmittance value of 0.1, and  $OD=8$  corresponds to a transmittance value of  $10^{-8}$ . It can be expressed as [12]:

$$T(\text{transmission}) = 10^{-OD} \times 100$$

$$OD = -\log\left(\frac{T}{100}\right) \dots \dots \dots \text{eq} \quad (1)$$

For the filters having  $OD \geq 3$  the effects of multiple reflections are insignificant because of the low reflectance and strong absorption of the filter.

## 3. Filter design and discussion

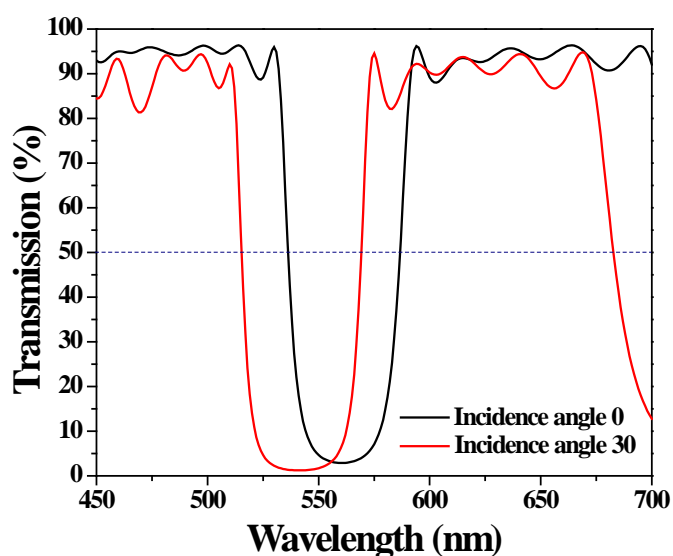
Multilayer thin films have an extensive wavelength tunability which gives an optical response that is desired for a specific application. Distributed Bragg Reflectors (DBRs)[13,14] consisting of alternating high and low refractive index material pairs are the most commonly used mirrors in FP filters, due to their high reflectivity. However, DBRs have high reflectivity for a selected range of wavelengths known as the stop band of the DBR. Its reflectance usually depends on the constructive or destructive interference of light reflected at consecutive boundaries of different layers of the stack. The performance of the multilayer devices highly depends on the interface formed between the alternating layers. Therefore an appropriate sequencing of the layers of suitable dielectric materials and their thicknesses is critical for achieving the desired spectral response and application. Therefore, it is important to optimize the coating conditions in the designing process [15, 16]. In our previous work, we proposed multilayer dielectric filter based on  $TiO_2$  and  $SiO_2$  materials because of their excellent optical properties [17]. Therefore,  $TiO_2$  and  $SiO_2$  are chosen as high and low refractive index materials, respectively. The choice of materials is made on the basis of low absorption and high index contrast in the wavelengths of interest. The notch filter is designed for visible spectrum ranges from 450-700 nm with FWHM of 50 nm. The optimized thickness of the layers is shown in table 1. The total thickness of the filter is estimated to be 3627 nm with a total of 27 alternating layers of  $TiO_2$  and  $SiO_2$  deposited on a substrate.



Table 1. Layer thickness of Notch filter based on  $\text{TiO}_2/\text{SiO}_2$ .

Layer no.	Layer name	Thickness (nm)	Layer no.	Layer name	Thickness (nm)
1	$\text{SiO}_2$	548	15	$\text{SiO}_2$	147
2	$\text{TiO}_2$	11	16	$\text{TiO}_2$	164
3	$\text{SiO}_2$	28	17	$\text{SiO}_2$	149
4	$\text{TiO}_2$	280	18	$\text{TiO}_2$	127
5	$\text{SiO}_2$	153	19	$\text{SiO}_2$	165
6	$\text{TiO}_2$	124	20	$\text{TiO}_2$	42
7	$\text{SiO}_2$	151	21	$\text{SiO}_2$	25
8	$\text{TiO}_2$	164	22	$\text{TiO}_2$	116
9	$\text{SiO}_2$	148	23	$\text{SiO}_2$	80
10	$\text{TiO}_2$	123	24	$\text{TiO}_2$	16
11	$\text{SiO}_2$	153	25	$\text{SiO}_2$	39
12	$\text{TiO}_2$	52	26	$\text{TiO}_2$	120
13	$\text{SiO}_2$	153	27	$\text{SiO}_2$	227
14	$\text{TiO}_2$	122	-	-	-

The transmission spectrum of the designed notch filter shows a stop band at 536 nm to 586 nm with a central wavelength at 561 nm. The line width which is measured at half of the maximum transmission is around 50 nm. The transmission in pass band regions 450-536nm and 586-700nm is more than 90% as shown in figure 2. The transmission of such filters can be improved by increasing the number of the layers. Whereas this designed filter has only 27 layers which can be implemented economically.

Fig. 2. The transmission spectrum of a notch filter at  $0^\circ$  and  $30^\circ$  of incidence light.

The designed filter has maximum transmission of 3% in the stop band. The OD of the filter is calculated by using an eq. (1) which provides a value greater than 3.5 (Transmission is 0.0003%). It shows a promising result for the notch filter. The optical density of the notch filter is plotted in figure 3.

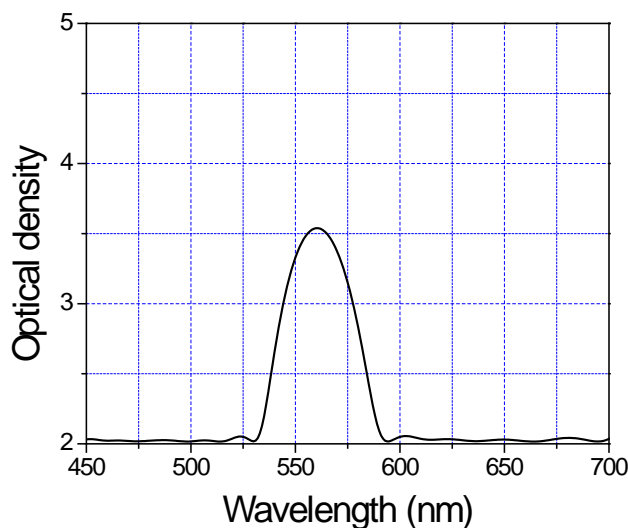


Fig. 3. The optical density of the designed notch filter.

#### 4. Effect of the angle of incidence of light on the central wavelength and FWHM

In all dielectric stack filters, the transmission depends on the angle of incidence. The central wavelength of the blocking region shifts to shorter wavelengths and FWHM increases as the angle of incidence is increased. It can be seen from figure 2, when the angle of incidence of light increases, a noticeable increase in the FWHM of the bandwidth of stop band is seen which shifts towards smaller wavelength. And an increase in the OD is also noticed which is around 3.9 with a slight decrease in the transmission of the band-pass region. Table 2 summarizes the effect of the incidence angle of light on the filters FWHM and central wavelength.

Table 2. Central wavelength and FWHM of the notch filter at different incident angles.

Angle of Incidence (Degrees)	Central wavelength (nm)	FWHM (nm)
0	561	50
30	542	53

#### 5. Conclusion

In this work, a multilayer dielectric optical notch filter design is presented which is based on  $\text{TiO}_2/\text{SiO}_2$  alternating layers. These filters provide an average transmission of more than 90% in region 450-535nm and 587-700 nm. The transmission of the stop band 536-586 nm is around 3%. The OD of this filter is greater than 3.5 which shows the high blocking specification of a filter and is associated with the amount of energy transmitted through it. With an increase in the incident angle of light, the central wavelength of the notch filter shifts toward smaller wavelength.

#### Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation and the Russian Foundation for Basic Research (grant No. 16-29-11698-ofi\_m, 16-29-11744-ofi\_m).

#### References

- [1] Macloed HA. Thin film optical filters. McGraw-Hill, 1989.
- [2] Kazanskiy NL, Serafimovich PG, Popov SB, Khonina SN. Using guided-mode resonance to design nano-optical spectral transmission filters. *Computer Optics* 2010; 34(2): 162–168.
- [3] Kazanskiy NL, Kharitonov SI, Khonina SN, Volotovskiy SG, Strelkov YuS. Simulation of hyperspectrometer on spectral linear variable filters. *Computer Optics* 2014; 38(2): 256–270.
- [4] Kazanskiy NL, Kharitonov SI, Khonina SN, Volotovskiy SG. Simulation of spectral filters used in hyperspectrometer by decomposition on vector Bessel modes, *Proc. of SPIE* 2015; 9533: 95330L – 7 pp.
- [5] Weber MJ. Handbook of Optical Materials. CRC Press: Boca, Raton, London, New York, Washington, 2003.
- [6] Hasan MM, Malek ABM, Haseeb ASMA, Masjuki HH. Investigations on  $\text{TiO}_2$  and Ag based single and multilayer films for window glazings. *ARNP Journals of Engineering and Applied Sciences* 2010; 5: 22–28.
- [7] Butt MA, Fomchenkov SA. Thermal effect on the optical and morphological properties of  $\text{TiO}_2$  thin films obtained by annealing a Ti metal layer. *J. Korean Phys. Soc.* 2017; 70(2): 169–172.
- [8] Hinczewski DS, Hinczewski M, Tepehen FZ, Tepehen GG. Optical filters from  $\text{SiO}_2$  and  $\text{TiO}_2$  multi-layers using sol-gel spin coating method. *Solar Energy Materials and Solar Cells* 2005; 87(1-4): 181–196.
- [9] Larouche S, Martinu L. Optical filters: Open-source software for the design, optimization, and synthesis of optical filter. *Appl. Opt.* 2008; 47(13): C219–C230.
- [10] Eckerle KL, Hsia JJ, Mienlenz KD, Weidner VR. Regular spectral transmittance. *NBS special Publication* 1987; 250(6): 1–59.
- [11] Zhang ZM. Optical properties of layered structures for partially coherent radiation. *Heat Transfer. Proceedings of the Tenth International Heat Transfer Conference. G.F. Hewitt* 1994; 2: 177–182.
- [12] Zhang ZM, Gentile TR, Migdall AL, Datta RU. Transmittance measurements for filters of optical density between one and ten. *Appl. Opt.* 1997; 36(34): 8889–8895.
- [13] Butt MA, Fomchenkov SA, Ullah A, Verma P, Khonina SN. Biomedical bandpass filter for fluorescence microscopy imaging based on  $\text{TiO}_2/\text{SiO}_2$  and  $\text{TiO}_2/\text{MgF}_2$  dielectric multilayers. *J. Physics. Conf. Series* 2016; 741: 012136.
- [14] Ullah A, Butt MA, Fomchenkov SA, Khonina SN. Indium phosphide air-gap Fabry-Perot filters for near Infrared spectroscopic applications. *J. Physics. Conf. Series* 2016; 741: 012135.
- [15] Kheraj VA, Panchal CJ, Desai MS, Potbhare V. Simulation of reflectivity spectrum for non-absorbing multilayer optical thin films. *Pramana-Journal of Physics* 2009; 72(6): 1011–1022.
- [16] Richter F, Kupfer H, Schlott P, Gessner T, Kaufmann C. Optical properties and mechanical stress in  $\text{SiO}_2/\text{Nb}_2\text{O}_5$  multilayers. *Thin Solid Films* 2001; 389(1-2): 278–283.
- [17] Butt MA, Fomchenkov SA, Ullah A, Habib M, Ali RZ. Modelling of multilayer dielectric filters based on  $\text{TiO}_2/\text{SiO}_2$  and  $\text{TiO}_2/\text{MgF}_2$  for fluorescence microscopy imaging. *Computer Optics* 2016; 40(5): 674–678. DOI: 10.18287/2412-6179-2016-40-5-674-678.

# Cold mirror based on High-Low-High refractive index dielectric materials

V.V. Elyutin<sup>1</sup>, M.A. Butt<sup>1</sup>, S.N. Khonina<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

## Abstract

In this paper, a design for a multilayer dielectric cold mirror based on TiO<sub>2</sub>/ SiO<sub>2</sub> and TiO<sub>2</sub>/MgF<sub>2</sub> alternating layers is presented. A cold mirror is a specific dielectric mirror that reflects the complete visible light spectrum whereas transmitting the infrared wavelengths. These mirrors are designed for an incident angle of 45°, and are modeled with multilayer dielectric coatings similar to interference filters. Our designed mirror based on TiO<sub>2</sub>/SiO<sub>2</sub> shows an average transmission of less than 5 % in the spectrum range of 425- 610 nm whereas it has an average transmission of 95 % in the spectrum range of 710-1500 nm.

**Keywords:** Cold mirror; TiO<sub>2</sub>; MgF<sub>2</sub>; SiO<sub>2</sub>; dielectric materials

## 1. Introduction

Thin film optics is a well-developed technology and many devices such as passband filters, stopband filters, polarizers and reflectors are successfully developed with the help of multilayer dielectric thin films [1-4]. These optical elements comprise of alternating layers of high and low refractive index materials with specific thicknesses and awareness of their refractive index and absorption. Multilayer dielectric filters are based on the principle of multiple reflections that takes place between the interfaces of high and low index materials. Distributed Bragg Reflectors (DBRs) are one of the widely used filters which are quarter wave thick of the center wavelength. The high reflection region of a DBR is known as the DBR stopband and can be obtained by the refractive index contrast between the constituent layers [5]. A cold mirror is a specific dielectric mirror that reflects the visible light spectrum while transmits the infrared wavelengths. These mirrors work on the principle of multiple reflections between high and low index material interface. The visible spectrum of light spans ~380-770 nm and the region beyond 770 nm in the near infrared, which is heat. Radiations from a tungsten lamp contain at least six times as much heat as useful light in the visible spectrum. The term cold light defines the radiation in which the IR spectrum is removed [6].

A hot mirror is just the opposite of cold mirror which is designed to reflect infrared region while transmits the visible portion of the beam. These mirrors can separate visible light from UV and NIR which helps in separating the heat from the system as shown in figure 1. Cold mirrors have many practical applications such as in projectors, copy machines, medical instruments and fibre optical illuminations [6, 7].

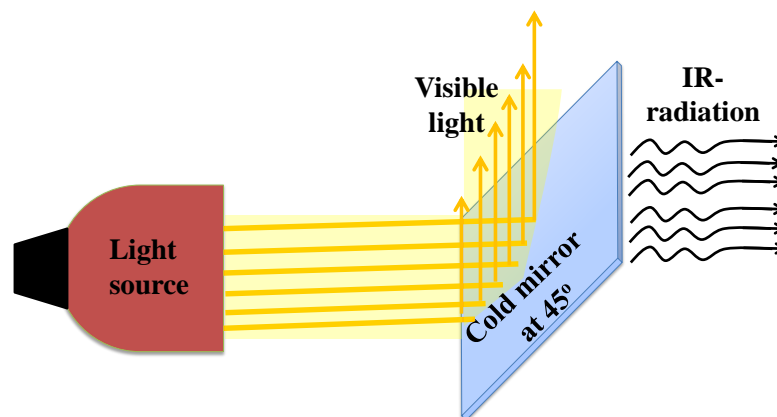


Fig. 1. Schematic of a cold mirror.

## 2. Theoretical basis of multilayer structure

Consider a multilayer dielectric system surrounded by an environment. Light from the source falls on the system at an angle  $\alpha_0$ . For this purpose, wave front can be considered as planar. To calculate the spectral transmittance and reflectance intensity for the p- and s-polarized light, matrix method is used:

$$T_s(\lambda) = \frac{n_m}{n_o} |t_s|^2, \quad R_s(\lambda) = |r_s|^2, \quad (1)$$

$$T_p(\lambda) = \frac{n_m}{n_o} |t_p|^2, \quad R_p(\lambda) = |r_p|^2, \quad (2)$$

where  $t_s$ ,  $r_s$  — amplitude transmission and reflection coefficients of the multilayer interference system for s-polarized light whereas  $t_p$ ,  $r_p$  — transmission and amplitude reflection coefficients for p-polarized light. Now, we will only consider s-polarization because equations for both s and p polarization are related till equation (7). Amplitude coefficients are determined from the following equations:

$$t_s = \frac{2n_o}{n_o m_{11s} + i n_o n_m m_{12s} + i m_{21s} + n_m m_{22s}}, \quad (3)$$

$$r_s = \frac{n_o m_{11s} + i n_o n_m m_{12s} - i m_{21s} - n_m m_{22s}}{n_o m_{11s} + i n_o n_m m_{12s} + i m_{21s} + n_m m_{22s}}, \quad (4)$$

where  $n_o$ ,  $n_m$  – the effective refractive indices of the substrate and the environment, respectively;  $m_i$ ,  $j_s$  – elements of the characteristic matrix  $M_s$  for s-polarized light:

$$M_s = \begin{pmatrix} m_{11s} & i m_{12s} \\ i m_{21s} & m_{22s} \end{pmatrix} = M_{1s} M_{2s} M_{3s} \dots M_{q-2s} M_{q-1s} M_{qs}, \quad (5)$$

q – Number of layers.

In the expression (5) matrices  $M_k (k = \overline{1, q})$  determine the properties of each individual layer of the optical filter. Filter design needs layers with high and low refractive indices. Therefore, the spectral characteristics are described by matrices multiplying:

$$M_1 = \begin{pmatrix} \cos(\phi_1) & \frac{i}{n_1} \sin(\phi_1) \\ i n_1 \sin(\phi_1) & \cos(\phi_1) \end{pmatrix}, \quad M_2 = \begin{pmatrix} \cos(\phi_2) & \frac{i}{n_2} \sin(\phi_2) \\ i n_2 \sin(\phi_2) & \cos(\phi_2) \end{pmatrix},$$

$$M_{q-1} = \begin{pmatrix} \cos(\phi_{q-1}) & \frac{i}{n_{q-1}} \sin(\phi_{q-1}) \\ i n_{q-1} \sin(\phi_{q-1}) & \cos(\phi_{q-1}) \end{pmatrix}, \quad M_q = \begin{pmatrix} \cos(\phi_q) & \frac{i}{n_q} \sin(\phi_q) \\ i n_q \sin(\phi_q) & \cos(\phi_q) \end{pmatrix}, \quad (6)$$

where  $\phi_k$  – phase thickness for s- polarized light, which is calculated by the following equations:

$$\phi_1 = \frac{2\pi}{\lambda} n_1 h_1 \cos(\alpha_1), \quad \phi_2 = \frac{2\pi}{\lambda} n_2 h_2 \cos(\alpha_2),$$

$$\phi_{q-1} = \frac{2\pi}{\lambda} n_{q-1} h_{q-1} \cos(\alpha_{q-1}), \quad \phi_q = \frac{2\pi}{\lambda} n_q h_q \cos(\alpha_q), \quad (7)$$

where  $h_k$  – the physical thickness of the layers,  $n_m$ ;  $\alpha_k$  – the angles of refraction in the layers;  $n_k$  – effective indexes refractive of the layers which depends on the wavelength. In this case, the angle of refraction in the layers is 45 degrees, relative to the normal.

The main difference in calculations between s- and p- polarized light is specified in (8) and (9) equations.

$$n_{1s} = n_1 \cos(\alpha_1), \quad n_{2s} = n_2 \cos(\alpha_2), \quad (8)$$

$$n_{1p} = n_1 / \cos(\alpha_1), \quad n_{2p} = n_2 / \cos(\alpha_2), \quad (9)$$

The angle of refraction in the layers is calculated by the equations (10).



$$\alpha_1 = \arccos\left(\sqrt{1 - \frac{n_o^2}{n_1^2} \sin^2(\alpha_o)}\right), \quad \alpha_2 = \arccos\left(\sqrt{1 - \frac{n_o^2}{n_2^2} \sin^2(\alpha_o)}\right), \quad (10)$$

Transmission of an unpolarized light is calculated as an average of  $T_s$  and  $T_p$ :

$$T = \frac{1}{2}(T_s + T_p), \quad (11)$$

By using these equations, the transmission spectrum of the multilayer  $\text{TiO}_2/\text{MgF}_2$  filter was plotted with the help of Java programing along with the transmission spectrum generated by commercially available open source filter Open filter. Their response is fairly comparable as shown in figure 2.

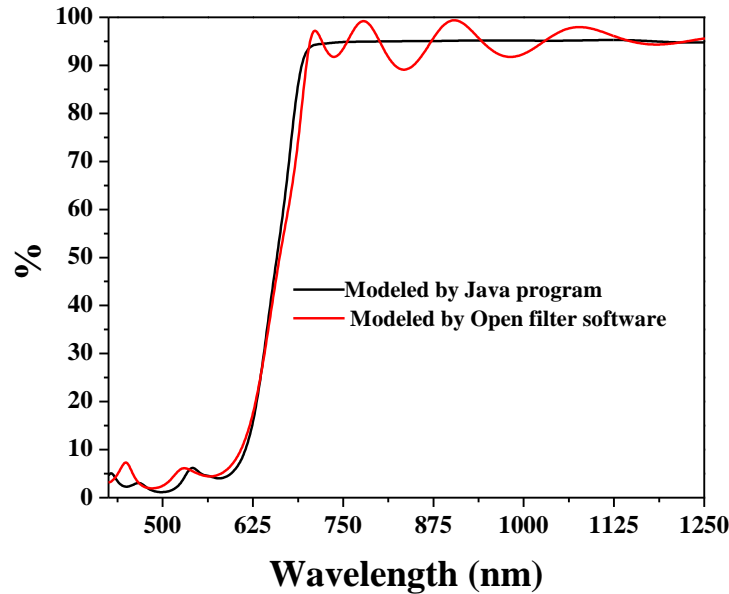


Fig. 2. Transmission spectrum of cold mirror modeled by Java programming and open source software: Open filter.

### 3. Filter design

In the designing of optical filters, the behaviour of the total multilayer system is estimated on the basis of the properties of the individual layers in the stack [8]. Therefore to achieve the optimum performance, it is significant to optically characterize and accurately determine the thickness of the individual layers. In this work, cold mirrors are designed in the wavelength range of 425-1500nm by using open source software Open Filter to selectively pass the wavelengths of interest and rejecting the undesired wavelengths in the visible spectrum.  $\text{TiO}_2$ ,  $\text{SiO}_2$  and  $\text{MgF}_2$  materials are carefully selected based on their high and low refractive indices, respectively.  $\text{TiO}_2$  is a vital dielectric material with a wide band-gap energy and high refractive index that can make it useful in the fabrication of multilayer thin films due to its high optical properties. For instance, its high transmittance and high refractive index in the visible region (380-760 nm) make it valuable to be employed in the production of the optical filter and window glazing [9, 10]. Layers made of oxides are harder than those made of fluorides, sulphides or semiconductors. Thus, they are ideal to be used on exposed surfaces. Semiconductor materials should be avoided in filters which have to be used over a wide range of temperatures because their optical constants can change considerably. The open filter uses transfer matrix method to analyze the transmission and reflection of light from layers based on thickness and type of materials. Designs are optimized to maximum the transmission required at wavelengths using needle synthesis method (addition of thin layers called needle and analyze transmission till the best results obtained) [11]. The thicknesses of the layers for cold mirror based on  $\text{TiO}_2/\text{MgF}_2$  and  $\text{TiO}_2/\text{SiO}_2$  are shown in table 1. Both mirrors have 20 layers with almost comparable total thickness. Special attention has been given to keep the thickness of the filters within economic limits.

Assuming the incident angle of un-polarized light equals  $45^\circ$ , these mirrors have reflective properties in the spectral range from 425-610 nm and 710-1500 nm up to 95 % and 5 %, respectively as shown in figure 3. For all dielectric stack filters, the transmission depends on the angle of incidence. The central wavelength of the FP filter shifts toward the smaller wavelengths as the angle of incidence is increased. When the incident angle of light decreases from  $45^\circ$  to  $0^\circ$  the transmission spectrum shifts from 710 to 750 nm.

#### 4. Conclusion

In this work, we presented the modeling results of cold mirrors based on  $\text{TiO}_2/\text{MgF}_2$  and  $\text{TiO}_2/\text{SiO}_2$  for  $45^\circ$  of un-polarized incident light by using java programming and commercially available Open source software Open filter. Both mirrors show the reflection of 95% in the spectral range of 425-610 nm and 95% of transmission in the spectral range of 710-1500 nm. The designs are optimized to maximum the transmission required at wavelengths using needle synthesis method. We observed a right shift in a spectrum when the angle of incidence of light was reduced from  $45^\circ$  to  $0^\circ$ .

Table 1. Layer thicknesses of  $\text{TiO}_2/\text{MgF}_2$  and  $\text{TiO}_2/\text{SiO}_2$  based Cold mirrors.

Layer no.	Material	Thickness (nm)	Layer no.	Material	Thickness (nm)
1	$\text{TiO}_2$	14	1	$\text{TiO}_2$	25
2	$\text{MgF}_2$	114	2	$\text{SiO}_2$	121
3	$\text{TiO}_2$	45	3	$\text{TiO}_2$	55
4	$\text{MgF}_2$	84	4	$\text{SiO}_2$	82
5	$\text{TiO}_2$	62	5	$\text{TiO}_2$	55
6	$\text{MgF}_2$	71	6	$\text{SiO}_2$	65
7	$\text{TiO}_2$	43	7	$\text{TiO}_2$	44
8	$\text{MgF}_2$	87	8	$\text{SiO}_2$	99
9	$\text{TiO}_2$	44	9	$\text{TiO}_2$	51
10	$\text{MgF}_2$	107	10	$\text{SiO}_2$	110
11	$\text{TiO}_2$	66	11	$\text{TiO}_2$	71
12	$\text{MgF}_2$	90	12	$\text{SiO}_2$	92
13	$\text{TiO}_2$	68	13	$\text{TiO}_2$	72
14	$\text{MgF}_2$	130	14	$\text{SiO}_2$	123
15	$\text{TiO}_2$	48	15	$\text{TiO}_2$	54
16	$\text{MgF}_2$	118	16	$\text{SiO}_2$	106
17	$\text{TiO}_2$	87	17	$\text{TiO}_2$	93
18	$\text{MgF}_2$	54	18	$\text{SiO}_2$	53
19	$\text{TiO}_2$	79	19	$\text{TiO}_2$	80
20	$\text{MgF}_2$	228	20	$\text{SiO}_2$	218
Total thickness		1639	Total thickness		1669

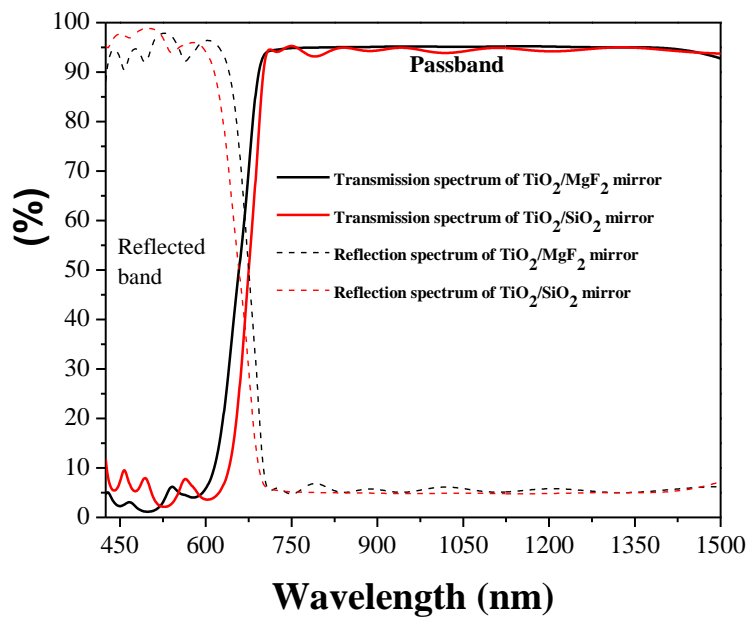


Fig. 3. Transmission and reflection spectrum of the cold mirror in the wavelength range of 425-1500 nm.

#### Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation and the Russian Foundation for Basic Research (grant No. 16-29-11698-ofi\_m, 16-29-11744-ofi\_m).

#### References

- [1] Macloed HA. Thin film optical filters. McGraw-Hill, 1989.
- [2] Kazanskiy NL, Serafimovich PG, Popov SB, Khonina SN. Using guided-mode resonance to design nano-optical spectral transmission filters. Computer Optics 2010; 34(2): 162–168.
- [3] Kazanskiy NL, Kharitonov SI, Khonina SN, Volotovskiy SG, Strelkov YuS. Simulation of hyperspectrometer on spectral linear variable filters. Computer Optics 2014; 38(2): 256–270.

- [4] Kazanskiy NL, Kharitonov SI, Khonina SN, Volotovskiy SG. Simulation of spectral filters used in hyperspectrometer by decomposition on vector Bessel modes, Proc. of SPIE 2015; 9533: 95330L-7pp.
- [5] Butt MA, Fomchenkov SA, Ullah A, Habib M, Ali RZ. Modelling of multilayer dielectric filters based on TiO<sub>2</sub>/SiO<sub>2</sub> and TiO<sub>2</sub>/MgF<sub>2</sub> for fluorescence microscopy imaging. Computer Optics 2016; 40(5): 674–678. DOI: 10.1109/ICECUBE.2016.7495230.
- [6] Baumeister PW. Optical coating technology. SPIE Press book, 2004.
- [7] Guenther BD. Modern Optics. Oxford University Press, 2015.
- [8] Hinczewski DS, Hinczewski M, Tepehen FZ, Tepehen GG. Optical filters from SiO<sub>2</sub> and TiO<sub>2</sub> multi-layers using sol-gel spin coating method. Solar Energy Materials and Solar Cells 2005; 87(1-4): 181–196.
- [9] Hasan MM, Malek ABM, Haseeb ASMA, Masjuki HH. Investigations on TiO<sub>2</sub> and Ag based single and multilayer films for window glazings. ARPN Journals of Engineering and Applied Sciences 2010; 5: 22–28.
- [10] Butt MA, Fomchenkov SA. Thermal effect on the optical and morphological properties of TiO<sub>2</sub> thin films obtained by annealing a Ti metal layer. J. Korean Phys. Soc. 2017; 70(2): 169–172.
- [11] Larouche S, Martinu L. Optical filters: Open-source software for the design, optimization, and synthesis of optical filter. Appl. Opt. 2008; 47(13): C219–C230.

# An algorithm for correcting X-ray image distortions caused by central projection

A.V. Ustinov<sup>2</sup>, N.Yu. Ilyasova<sup>1,2</sup>, N.S. Demin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoye shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, 151 Molodogvardeyskaya street, 443001, Samara, Russia

## Abstract

We propose an algorithm for correcting X-ray image distortions caused by central projection. Relationships between coordinates of X-ray image points obtained using parallel and central projection are derived. We describe two correction techniques using which the original image can be made similar to the one based on parallel projection. In this way, the three-dimensional heart vessel model can be essentially simplified and diagnostic parameters can be assessed with higher accuracy, resulting in a more accurate early disease diagnosis.

*Key words:* X-ray image; distortion correction; central projection

## 1. Introduction

Roentgenology is an extensively used and dynamic branch of medicine that uses X-ray imagery for the diagnosis of a variety of diseases. By way of illustration, X-ray angiography is utilized for diagnosis of cardiovascular diseases [1]. As a rule, the diagnosis is made based on visual assessment of angiograms, however, its accuracy essentially depends on the projection angle. A three-dimensional model of heart vessels serves to visualize three-dimensional geometric and topological information, thus enabling the diagnosis to be made with higher accuracy [2-6].

The initial data is a sequence of DICOM frames [7, 8]. The projection procedure is affected by a number of technical limitations, resulting in the imagery characterized by a variety of distortions. Various techniques for distortion correction were described in Ref. [9,10,13]. The X-ray imagery is obtained using specialized clinical equipment, such as an operating-room X-ray unit C-ARM. Examples of such equipment are illustrated in Figure 1. The unit is composed of an X-ray source and receiver connected by an arc-shaped holder freely moving on a support. With such a design, the X-ray camera has two degrees of freedom. The camera can also move relative to the holder in the longitudinal direction, enabling the image to be scaled. The spatial position of the camera is described by two angles: primary and secondary angles of the camera position.

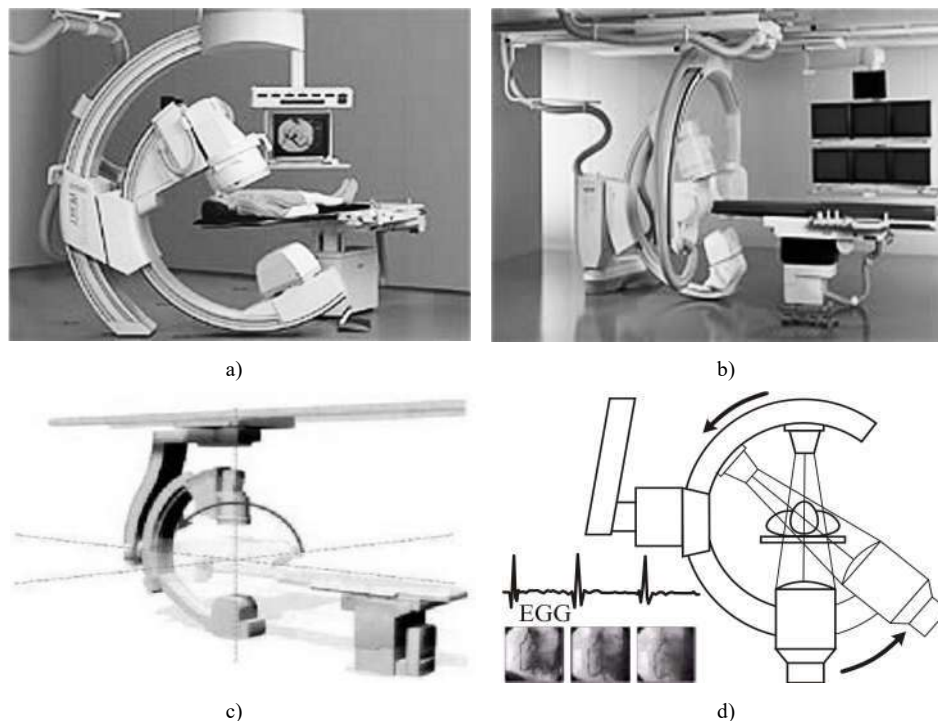


Fig. 1. Specialized clinical equipment: a) AXIOM Multista, b) AXIOM Artis BC, c) primary angle, d) secondary angle.

The primary angle (denoted as  $\alpha$ ) is provided by rotating the camera holder and the support as a whole relative to the mount beam. The secondary angle  $\beta$  (analogous to the geographic latitude) is provided by sliding the arc-shaped holder on the support guide, with the camera and X-ray source moving on a circular arc. The imaging is done by a diverging X-ray, with the divergence angle defined by technical characteristics of the scanning device and usually found within 10-12°. As a result, the X-



ray image is observed in the central projection. Based on such projections, it is not possible to reconstruct a model of the original object without additional information concerning imaging conditions, which is not available. At the same time, the original image can be reconstructed from parallel projections without use of any additional information [7,8]. Our idea is that effects caused by central projection can partially be compensated for by making the X-ray image look as if it were built using parallel projections.

By correcting X-ray image distortions caused by central projection, the three-dimensional heart vessel model can be essentially simplified [12] and diagnostic parameters can be assessed with higher accuracy [5], resulting in a more accurate early disease diagnosis.

## 2. Mathematical model

Because the refractive index for X-rays for all substances is very close to one, it is impossible to make a collimator that converts a divergent beam from a source close to a point beam into a parallel beam. This leads to an object image deformation even at the planar object. At straight rays falling (the primary and secondary angles are zero), the distortion reduces to a change in scale - the image remains similar to the image obtained by parallel projection. At inclined falling an additional distortion appears: the scale becomes different on image area, which leads to a violation of similarity (a change in the proportions of the object in the case of oblique incidence also occurs in parallel projection). We propose an algorithm for compensating of central projection influence (divergent rays). It should be noted that full compensation is possible only for a planar object, because otherwise it is principally impossible to indicate precise values of some parameters. Nevertheless, the achieved compensation is enough for more precision of three-dimensional trace process described in Ref. [7, 12].

In order that to compensate the central projection influence we are need to get a relation of image point coordinates at parallel and central projection. For this purpose a calculation of world and planar coordinates of projection point is required. Let a point  $S$  is a light source and distance  $OS = H$  is given. A direction to the light source is defined as  $OS = nH$ . Point  $A = (x_0; y_0; h)$  is an intersection of projection plane with straight-line having a directing vector  $SA$ , where  $h$  is distance from image plane to the object at zero angles. Now we obtain the projection of point  $A = (x; y; z)$ . Planar coordinates of point are  $u = OAu = xu_x + yu_y + zu_z, v = OAv = xv_x + yv_y + zv_z$ , where  $u, v$  are basic vectors of planar coordinates system. If the plane is defined by an equation  $Ax + By + Cz + D = 0$  and straight-line is defined by formulas  $x = x_0 + lt; y = y_0 + mt; z = z_0 + nt$ , then coordinates of point A are obtained at substitution of value  $t$  derived from equation  $(Al + Bm + Cn)t + Ax_0 + By_0 + Cz_0 + D = 0$ . In our case:

$$D = 0; (A; B; C) = (n_x; n_y; n_z); (x_0; y_0; z_0) = (x_0; y_0; h); (l; m; n) = AS = nH - (x_0; y_0; h) = (Hn_x - x_0; Hn_y - y_0; Hn_z - z_0).$$

At the parallel projection:  $(l; m; n) = n = (A; B; C)$ . Taking in account these values we get the projection point coordinates:

$$\begin{aligned} x &= \frac{x_0(B^2 + C^2) - y_0AB - hAC}{1 - (Ax_0 + By_0 + Ch) / H}, \\ y &= \frac{-x_0AB + y_0(A^2 + C^2) - hBC}{1 - (Ax_0 + By_0 + Ch) / H}, \\ z &= \frac{-x_0AC - y_0BC + h(A^2 + B^2)}{1 - (Ax_0 + By_0 + Ch) / H}. \end{aligned} \tag{1}$$

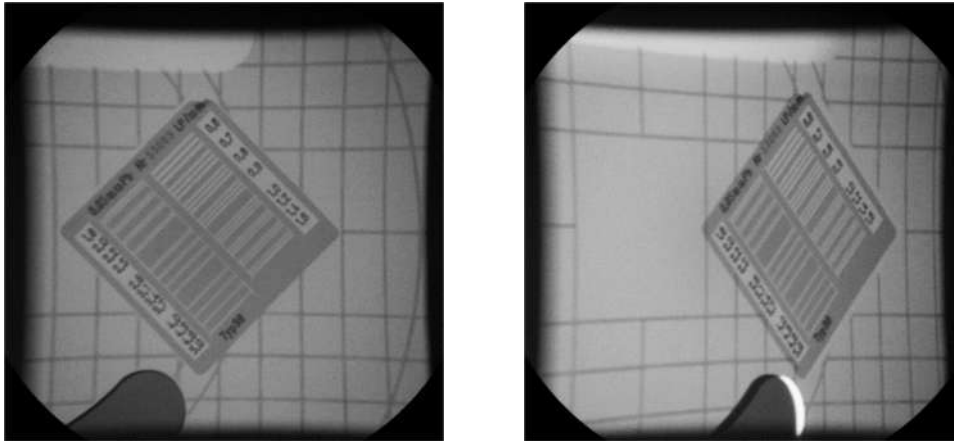


Fig. 2. Effect of divergent beam. Left - straight falling ( $\alpha=\beta=0$ ), right - inclined ( $\alpha=45^\circ, \beta=0$ ). The second frame has larger crosshairs on left part in image than in right part.

Numerators are projection point coordinates at the parallel projection and  $L = Ax_0 + By_0 + Ch$  in denominators is signed distance from point  $A$  to the projection plane.

Taking in account formulas for planar coordinates we get:

$$u = u_{par} / (1 - L / H),$$

$$v = v_{par} / (1 - L / H).$$

In some cases value of  $H$  we can get from DICOM-file header, else we can evaluate it with aid of mira image - special etalon image (grid in fig.2). In item 3 an algorithm invented for evaluation of  $H$  is described. About value of  $h$ : if the object is not planar then  $h$  can not be obtain in principle, since a reconstruction of three-dimensional object is need to this. For approximate solution of correction task we propose three approaches:

- 1) If object length in  $OZ$  axis direction compare with  $H$  is small and object is near to receiver then we use  $h=0$ .
- 2) If object length is small but object is not near to receiver there we use some average value, for example, a heart size.
- 3) Assuming that distortions are not large (three-dimensional tracing process is not failed) we build three-dimensional tree ignoring the central projection influence. That  $h$  equal to  $z$ -coordinate of corresponding tree point. After correction the three-dimensional tree building is repeated. But this variant increases calculation volume.

### 3. Determination of distance from the source to the receiver with usage of mira

For the determination of distance from the source to receiver with usage of mira, we preliminarily consider an auxiliary case. Let we have straight falling rays (at neglecting distortion of central projection) and a task is planar: source, object and coordinates origin lie in one plane perpendicular to the plane of the receiver. The distance from point source to receiver plane (the origin) is  $H$ , the distance from point object to the origin is  $h$ . Then a following relation holds:  $x_c = (H/H - h)x_p$ , it is connects coordinates of the object image on the receiver plane with the parallel projection and the central one.

Now we consider our case. The task is still planar - the source and the non-point object passing through the origin lie in one plane perpendicular to the plane of the receiver. Let he object is segment of length  $2a$  lying horizontally in this plane and the origin is its center, i.e.  $h=0$  for segment center. A left point will be the first, and a right point will be the second. Rays fall bias at angle  $\phi$  to vertical (the angle is counter on clockwise). The distance of the point source to the origin is  $H$ . Under this scheme of observation the central point of segment-object retains its zero coordinate on segment-image, but it will not be a center of segment-image. For the parallel projection, image stays symmetrical: if we use a length (instead of coordinate with sign) then  $x_{p1} = x_{p2} = a \cos \phi$ .

Because of the segment is not lie in in receiver plane, value  $h \neq 0$  for its ending points. A module of this value is equal for the first and the second points:  $|h_1| = h_2 = a \sin \phi$ .

But for the first (left) point it has minus sign. For absolute values we have final result:

$$x_{c1} = \frac{H}{H + a \sin \phi} a \cos \phi;$$

$$x_{c2} = \frac{H}{H - a \sin \phi} a \cos \phi.$$
(2)

This means that if we know angle  $\phi$  and the lengths of segments measured on the image which are equal on the object, then we can calculate distance  $H$  and the length of segment by formulas:

$$a = \frac{2x_{c1}x_{c2}}{(x_{c1} + x_{c2}) \cos \phi};$$

$$H = \frac{2x_{c1}x_{c2}}{x_{c2} - x_{c1}} \operatorname{tg} \phi.$$
(3)

A really used mira image contains many segments. Since mira grid on object is rectangular then it is desirable to use of images photographed when one of camera orientation angle is zero. This give two advantages: angle  $\phi$  is calculated easily (it is equal to modulus of the second orientation angle) and lines of equal scale stay be straight and parallel to image sides. If  $\beta=0$ , these lines are vertical, if  $\alpha=0$ , ones are horizontal. For definiteness, we shall consider case of  $\beta=0$ . At this condition we have the following algorithm.

Step 1. We do search of mira grid crosshair nearest to the image center.

Step 2. We do step back on some squares to left (not necessary on line containing center). A difference of horizontal coordinates of central and new crosshair is value  $x_{c1}$ .

Step 3. We do step back on some squares to right (not necessary on line containing center). A difference of horizontal coordinates of new and central crosshair is value  $x_{c2}$ .

Step 4. The sought distance  $H$  is calculated by the second formula (3). If it is negative then modulus is used - minus sign means that right segment is shorter of left one that takes place of rays falling left from vertical. At obtaining of formulas (2) we assumed rays falling right from vertical.

#### 4. Correction distortion of central projection with cycle on source image

The first method correction distortion of central projection consists in following: for points on source image we search corresponding points on corrected image.

If  $x_0, y_0$  are found then correction is implemented by formulas:

$$u_{par} = u \left( 1 - \frac{Ax_0 + By_0 + Ch}{H} \right)$$

$$v_{par} = v \left( 1 - \frac{Ax_0 + By_0 + Ch}{H} \right)$$

Solution of equation set for values  $x_0, y_0$  is following:

$$x_0 = \frac{a_4(u + b_1h) - a_2(v + b_2h)}{a_1a_4 - a_2a_3}$$

$$y_0 = \frac{a_1(v + b_2h) - a_3(u + b_1h)}{a_1a_4 - a_2a_3}$$
(4)

Knowing expressions for vectors dependence on primary and secondary angles

$$n = (-\sin \beta; \sin \alpha \cos \beta; \cos \alpha \cos \beta)$$

$$u = (\cos \beta; \sin \alpha \sin \beta; \cos \alpha \sin \beta)$$

$$v = (0; \cos \alpha; -\sin \alpha)$$

we get expressions for coefficients  $a_1, a_2, a_3, a_4, b_1, b_2$ :

$$a_1 = -(u/H) \sin \beta + \cos \beta,$$

$$a_2 = (u/H) \sin \alpha \cos \beta + \sin \alpha \sin \beta,$$

$$a_3 = -(v/H) \sin \beta,$$

$$a_4 = (v/H) \sin \alpha \cos \beta + \cos \alpha,$$

$$b_1 = -(u/H) \cos \alpha \cos \beta - \cos \alpha \sin \beta,$$

$$b_2 = -(u/H) \cos \alpha \cos \beta + \sin \alpha.$$
(5)

But if we want to form the correct image this method is undesirable: because of discretization part of output image points (in area of stretching) will not have a prototype. There is a grid from points without the prototype on fig. 3 that usually is not suited for us.

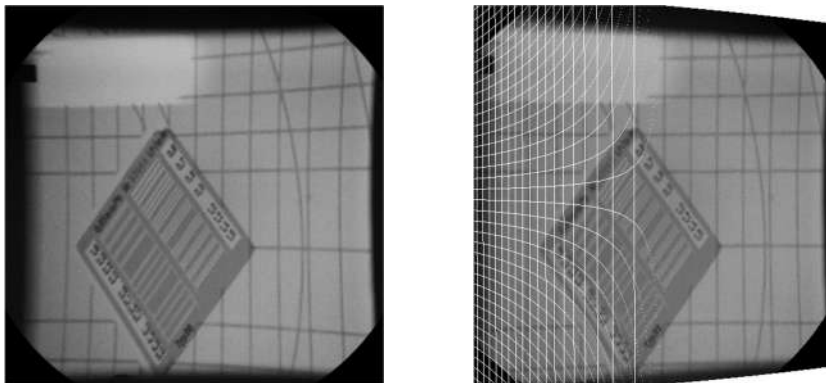


Fig. 3. Correction with cycle on source image. Left – source image, right – corrected one. A white grid is points without prototype.

## 5. Correction distortion of central projection with cycle on corrected image

For avoidance of empty points we employ the second method correction effects of central projection. Here the standard approach of image spatial transformation (for example, rotation or reflecting in non-planar mirror) is used - a cycle employs on output image and the prototype of current point is found on the source image.

If  $x_0, y_0$  are found then the prototype point has coordinates:

$$u = u_{par} / \left( 1 - \frac{Ax_0 + By_0 + Ch}{H} \right),$$

$$v = v_{par} / \left( 1 - \frac{Ax_0 + By_0 + Ch}{H} \right).$$

Coordinates of point  $x_0, y_0$  is calculated by formulas:

$$x_0 = \frac{a_4(u_{par} + b_1h) - a_2(v_{par} + b_2h)}{a_1a_4 - a_2a_3},$$

$$y_0 = \frac{a_1(v_{par} + b_2h) - a_3(u_{par} + b_1h)}{a_1a_4 - a_2a_3}. \quad (6)$$

Coefficients are equal to  $a_1 = \cos \beta$ ;  $a_2 = \sin \alpha \sin \beta$ ;  $a_3 = 0$ ;  $a_4 = \cos \alpha$ ;  $b_1 = -\cos \alpha \sin \beta$ ;  $b_2 = \sin \alpha$ , where  $\alpha, \beta$  are primary and secondary angles of camera rotation, and  $h$  is distance from the object to the projection plane.

After a substitution of these values we find simpler formulas:

$$x_0 = \frac{a_4(u_{par} + b_1h) - a_2(v_{par} + b_2h)}{a_1a_4},$$

$$y_0 = \frac{v_{par} + b_2h}{a_4}. \quad (7)$$

On fig. 4 we see that the grid from points without prototype is absent, i.e. the correction is more precise.

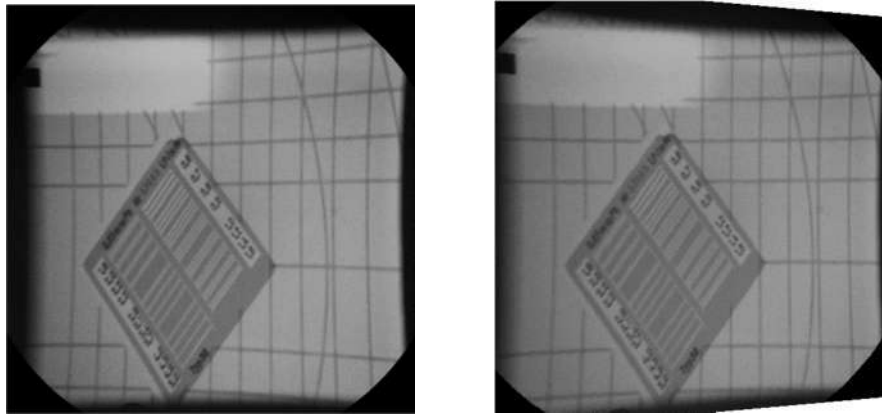


Fig. 4. Correction with cycle on corrected image. Left – source image, right – corrected one.

## 6. Conclusion

In this work, we propose an algorithm enabling X-ray image distortions caused by central projection to be corrected for. Considering that it is not possible to reconstruct the original object from a three-dimensional heart vessel model without additional information concerning the imaging conditions, which is not available in most cases, the process gets more complicated. At the same time, the reconstruction based on parallel projections requires no additional information. We propose a relation of image point coordinates at parallel and central projection. We describe two correction techniques using which the original image can be made similar to the one based on parallel projection. In this way, the three-dimensional heart vessel model can be essentially simplified and diagnostic parameters can be assessed with higher accuracy, resulting in a more accurate early disease diagnosis.

## Acknowledgements

The work was carried out with the partial support of the Ministry of Education and Science of the Russian Federation within the framework of the activities of the Program for Enhancing the Competitiveness of the SSAU among the world's leading scientific and educational centers for 2013-2020; Grants of the Russian Foundation for Basic Research No. 14-07- 97040, No. 15-29- 03823, No. 15-29- 07077, No. 16-57-48006, No. 16-41-630761; Program № 6 of fundamental research Department of Nanotechnologies and Information Technologies of the Russian Academy of Sciences "Bioinformatics, modern information technologies and mathematical methods in medicine" 2016 -2017.

## References

- [1] Chandra T, Pukenas B, Mohan S, Melhem E. Contrast-Enhanced Magnetic Resonance Angiography. *Magnetic Resonance Imaging Clinics of North America* 2012; 20(4): 687–698.
- [2] Ilyasova N. Computer Systems for Geometrical Analysis of Blood Vessels Diagnostic Images. *Optical Memory and Neural Networks (Information Optics)* 2014; 23(4): 278–286.
- [3] Ilyasova N. Methods to evaluate the three-dimensional features of blood vessels. *Optical Memory and Neural Networks (Information Optics)* 2015; 24(1): 36–47.
- [4] Ilyasova N. Evaluation of Geometric Characteristics of the Spatial Structure of Vessels. *Pattern Recognition and Image Analysis* 2015; 25(4): 621–625.
- [5] Ilyasova NYu. Estimating the geometric features of a 3D vascular structure. *Computer Optics* 2014; 38(3): 529–538.
- [6] Ilyasova NYu. Methods for digital analysis of human vascular system. Literature review. *Computer Optics* 2013; 37(4): 511–535.
- [7] Ilyasova NYu, Kupriyanov AV, Khramov AG. Information technologies of image analysis in medical diagnostics. *Radio and communication*, 2012.
- [8] Soifer VA. *Computer Image Processing. Part II: Methods and algorithms: Appendix A2. Biomedical Images Processing*. VDM Verlag, 2009.
- [9] Hong C, Lee D-H, Han BS. Characteristics of geometric distortion correction with increasing field-of-view in open-configuration MRI. *Magnetic Resonance Imaging* 2014; 32(6): 786–790.
- [10] Cardoso PL, Dymerska B, Bachratá B, Fischmeister FPhS, Mahr N, Matt E, Trattng S, Beisteiner R, Robinson SD. The clinical relevance of distortion correction in presurgical fMRI. *NeuroImage* 2016.
- [11] Menga C, Zhang J, Zhou F, Wang T. New method for geometric calibration and distortion correction of conventional C-arm. *Computers in Biology and Medicine* 2014; 52: 49–56.
- [12] Ilyasova NYu, Kazanskiy NL, Korepanov AO, Kupriyanov AV, Ustinov AV, Khramov AG. Computer technology for reconstructing the 3D structure of coronary arteries from angiographic projections. *Computer Optics* 2009; 33(3): 281–318.
- [13] Moravec J, Hub M. Automatic correction of barrel distorted images using a cascaded evolutionary estimator. *Information Sciences* 2016; 366: 70–98.

# Prognostic modeling of the curvilinear graphene selective hydrogenation process for the formation of optical scheme components for nanophotonics

Hussein Safaa Mohammed Ridha<sup>1,2</sup>, S.I. Kharitonov<sup>1,3</sup>, V.S. Pavelyev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoe shosse, 443086, Samara, Russia

<sup>2</sup>University of Karbala, 56001, Karbala, Iraq

<sup>3</sup>Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, 151 Molodogvardejskaya Street, 443001, Samara, Russia

---

## Abstract

The article deals with modeling of curvilinear graphene hydrogenation process. During the hydrogen atoms addition, the maximum stresses shift from the region of the edge atoms to the central region of the structure. The ionization potential of curvilinear graphene begins to increase even with an insignificant hydrogen atom concentration on its surface. To vary the energy gap value of the curvilinear graphene spectrum, a high concentration of hydrogen atoms is necessary.

*Keywords:* graphene; graphane; nanocarbon structures; optical properties

---

## Introduction

Nowadays, one of the promising directions in the field of nano- and bioelectronics is the development of new devices based on functionalized graphene. Today, the functionalization of graphene is one of the most effective ways to manage the properties of graphene material in order to expand the boundaries of its possible application in electronics and optics. The production of graphene nanostructures, functionalized with hydrogen, is an intensively developing direction of the modern nanoindustry, as well as the study of their properties. The hydrogen-functionalized graphene layer is a promising material for nanoelectronics and has received the name graphane in the literature. For the first time, graphane was experimentally obtained by the staff of the laboratory of Manchester University with the participation of Geim and Novoselov in 2009 by placing the graphene monolayer in hydrogen plasma. Graphane has two-dimensional, hexagonal crystalline structure. Hydrogen atoms are attached on both sides of the carbon atom plane by chemical bonds. According to the graphane sizes, graphite nanoparticles and nanobelts should be distinguished. The sizes of nanoparticles differ no more than in 3 times and do not exceed 100 nm in various directions.

The discovery of graphane have created the background for research its properties and searching for possible applications. In particular, graphane has unique optical properties. It has been established that the dielectric constant of graphane nanobelts does not depend on the shape of the belt edges and its width. In addition, as a result of studying the graphane optical properties, it has been shown that there is a moderate anisotropy with respect to the type of light polarization. The attention of researchers is also attracted to the study of the graphane thermal properties. For example, in [1] the authors investigate the graphane nanobelts thermal conductivity using the nonequilibrium Green's function method. In this paper it has been shown that the graphane thermal conductivity can be effectively controlled by the edge shape, the width, and also by the hydrogen vacancy concentration. In particular, the ballistic graphane nanobelts thermal conductivity usually increases with the belt width.

A promising area for graphane studies is the research of the graphane magnetic properties. It have been established in [2] that hydrogenated graphene demonstrates weak ferromagnetism in a wide range of temperatures down to room temperature, the nature of which is determined by the features of the graphane atomic structure itself.

As consequence of its unique physicochemical properties, graphane finds applications in a wide variety of scientific and technical fields. In particular, this material can be used in hydrogen economy. It has been found that heating of graphane leads to the atomic hydrogen release. Consequently, graphane can be considered as one of the most effective ways of storing hydrogen. Another important application of graphane will be its use in nanoelectronics as a basis for printed circuits with conductive and non-conductive areas on a sheet of graphane. One more possible application of graphane is the field of biosensorics. In the experimental work [3], the possibility of using graphane for electrochemical detection by applying graphan biomarkers is considered.

In one of the recent papers [4], the authors have studied theoretically the possibility of the existence of a two-dimensional doped graphane superconducting state. According to the results presented in the paper, doped graphane is a promising candidate for the creation of superconductors with a critical temperature higher than that of copper oxides. Another promising field of graphane application is its use as thermoelectric materials for thermionic devices. It is predicted that disordered armchair graphane nanobelts with low thermal conductivity can become a basis for creating thermoelectric materials.

## Methods of investigation

In the work, the ionization potential and the energy gap, determined from the electronic spectrum, have been considered as the electron-energy characteristics of curvilinear graphene adsorbing hydrogen. The composite electron spectrum has been calculated with the close coupling method. Fig. 1 shows the energy levels diagram with an indication of the energy gap and the



ionization potential. The ionization potential is determined by the last filled energy level (HOMO), and the energy gap is the interval between the last filled (HOMO) and the first vacant level (LUMO).

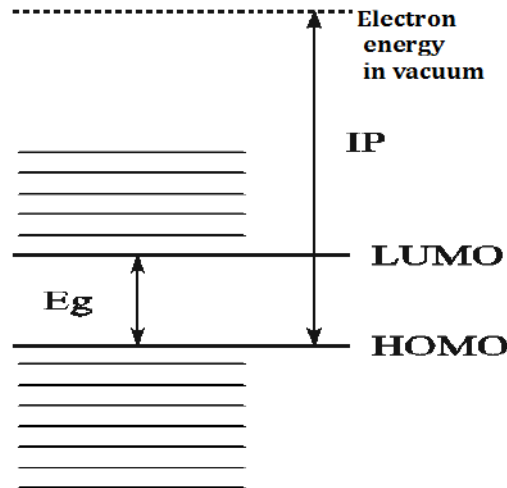


Fig. 1. An approximate diagram for the energy levels arrangement in the electronic spectrum with an indication of the HOMO and LUMO levels.

The change in the ionization potential and the energy gap of the electron spectrum during the process of curvilinear graphene selective hydrogenation is shown in the graphs represented by Fig. 2 and Fig. 3.

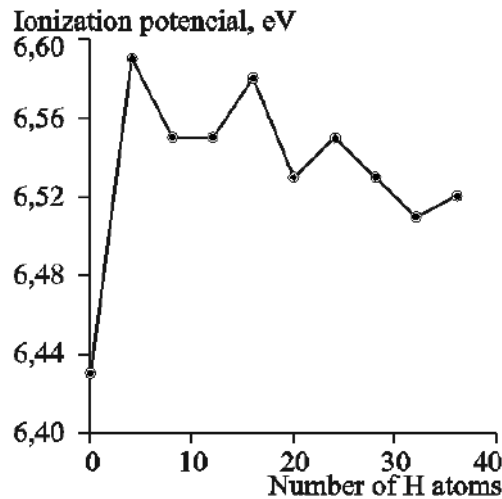


Fig. 2. The change in the ionization potential of the graphene electron spectrum during the hydrogen atoms addition.

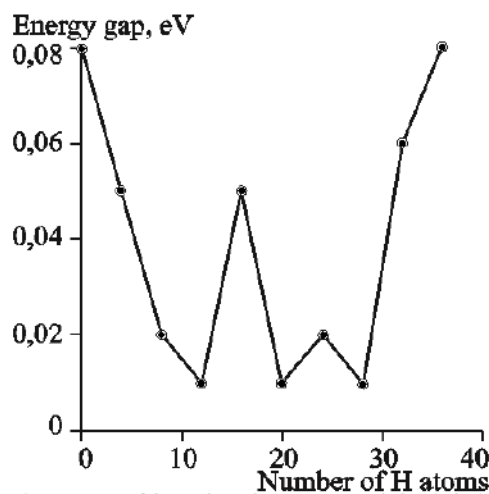


Fig. 3. The change in the energy gap of the graphene electron spectrum during the hydrogen atoms addition.

It can be seen from the graphs, even at the moment of addition of the first group of hydrogen atoms, the ionization potential of the structure increases steeply from 6.43 to 6.59 eV, and then changes in small limits near the value of 6.55 eV. Nevertheless, the obtained results indicate that the work function value, conclusions about which can be made from the ionization potential value, generally increases during the chemical adsorption of hydrogen by graphene, and hence the emissivity of such graphene structures decreases. The energy gap of the graphene electron spectrum, as seen from the graph in Fig. 3, varies discontinuously,

linearly decreasing during the addition of the first groups of hydrogen atoms, and then changing in an alternating manner, increasing or decreasing. However, the range of values in which the gap varies evidence that the hydrogen atom concentration for the graphene fragment of given sizes considered in the work is insufficient to change the curvilinear graphene conductivity type from a semimetal to a semiconductor, or even to a dielectric.

Further, we have investigated the change in the density-of-states (DOS) distribution of curvilinear graphene with a gradually increasing of the hydrogen atom concentration. The energy spectrum of curvilinear graphene in which the energy of each molecular orbital was represented as a spectral line has been constructed in order to calculate DOS. The intensities of all the lines have been set to one. After that each line has been replaced by a Gaussian distribution with a half-width at a given half-height of 0.1 eV. The intensities of all distributions for each energy value have been added up.

In constructing the partial electron density of atomic orbitals  $x$ , the intensity of each line corresponding to the molecular orbital  $y$  has been assumed to be equal to the sum of the squared coefficients of the atomic orbitals  $x$  in the expansion of method of linear combinations of atomic orbitals (MO LCAO) of orbital  $y$ . Further, the algorithm for the partial density of states was analogous to the algorithm for constructing the total density of states.

The results of calculating the distribution of the  $\pi$ -electronic states density of the initial curvilinear graphene are shown in Fig. 4. The vertical line in the figure indicates the HOMO level. The figure shows that there are two characteristic symmetrical peaks of approximately equal intensity in the DOS distribution, one of which is in the valence band, the other is in the conduction band, and also a small cluster of electronic states near the HOMO level.

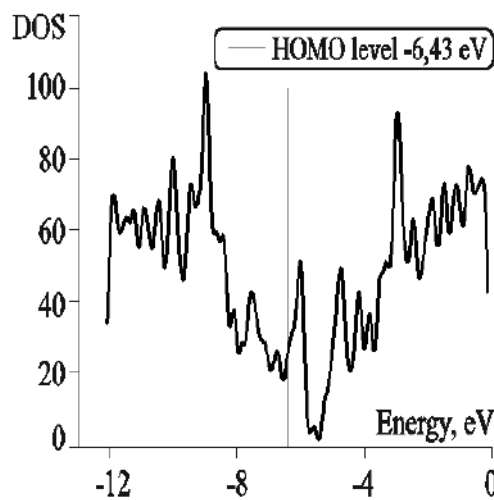


Fig. 4. The DOS distribution for the  $\pi$ -electrons of curvilinear graphene.

Further similar calculations and constructions were performed for curvilinear graphene with a different number of hydrogen atoms. The change in the DOS distribution for  $\pi$ -electrons of curvilinear graphene at each of the stages of hydrogen atom addition is shown in Fig. 5. It can be seen from the graph that the hydrogen addition causes a shift of the DOS characteristic towards the conduction band. The general character of the arrangement of peaks with maximum intensity also changes. The most significant changes are observed near the HOMO level and along the edges of the valence band as well as the conduction band. Near the HOMO level, the level density increases at each of the stages of hydrogen addition, while along the edges of the conduction and valence bands the density of states decreases in an oscillating manner.

## Main results

The aim of this work is to determine the patterns of hydrogen atom chemical adsorption on curvilinear graphene using computer simulation methods.

In the course of studying the process of selective hydrogenation of curvilinear graphene, new physical patterns were revealed:

- From the energy point of view, chemical addition of hydrogen atoms to curvilinear graphene atoms with the greatest stress will be beneficial;
- During the addition of hydrogen atoms, the maximum stresses shift from the region of the edge atoms to the central region of the structure;
- The ionization potential of curvilinear graphene begins to increase even with an insignificant concentration of hydrogen atoms on its surface;
- To vary the energy gap value of the curvilinear graphene spectrum, a high concentration of hydrogen atoms is required;

The chemical addition of even a small number of hydrogen atoms leads to a shift in the DOS distribution toward the conduction band and the peak intensity redistribution near the HOMO level and along the band edges.

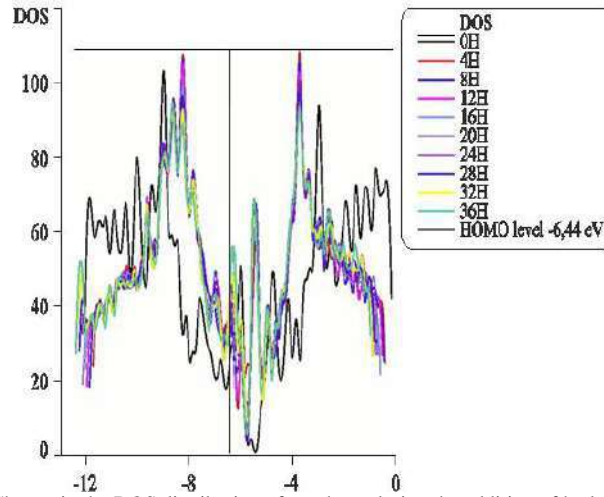


Fig. 5. Change in the DOS distribution of graphene during the addition of hydrogen atoms.

## Conclusion

Thus, on the ground of the obtained results, an energetically advantageous mechanism of selective hydrogenation of curvilinear graphene has been proposed to control the charge carrier motion in the structure. The proposed mechanism can be used to form conductive areas in modern electronic circuits and to produce components of optical elements.

## References

- [1] Li D, Xu Y, Chen X, Li B, Duan W. Tunable anisotropic thermal conduction in graphane nanoribbons. *Applied Physics Letters* 2014; 104: 143108.
- [2] Eng AYS, Poh HL, Sanek F, Marysko M, Matejkova S, Sofer Z, Pumera M. Searching for Magnetism in Hydrogenated Graphene: Using Highly Hydrogenated Graphene Prepared via Birch Reduction of Graphite Oxides. *ACS Nano* 2013; 7: 5930-5939.
- [3] Peng Q, Dearden AK, Crean J, Han L, Liu S, Wen X, De S. New materials graphyne, graphdiyne, graphone, and graphane: review of properties, synthesis, and application in nanotechnology. *Nanotechnology, Science and Applications* 2014; 7: 1-29.
- [4] Durajski AP. Influence of hole doping on the superconducting state in graphane. *Superconductor Science and Technology* 2015; 28: 035002-8 pp.

# Methods for creation of diffractive intraocular lenses

A.V. Gornostay

Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

Hybrid intraocular lenses (IOLs) are diffractive-refractive lenses with pseudoaccommodation. They can imitate the optical system of healthy eye more precisely, than other types. These lenses have 2-3 or more stable focuses. I proposed a comparison of hybrid IOLs: their shape, their processing, their aberrations and diffractive efficiency. I showed a review of new methods for correction of chromatism and I proposed a method of using volume holograms. I discussed features of possible materials.

*Keywords:* computer optics; diffractive optics; focusator; intraocular lens; holography; digital holography

---

## 1. Introduction

Some of aged peoples have cataract. Affected crystalline in this case must be removed. As an eye loses an ability of focusing an image on retina, it is necessary to use artificial IOLs (intraocular lenses). Some basic types of IOL exist: monofocal, accommodative and hybrid diffractive-refractive IOLs. Monofocal lenses [1] are the simplest in processing, but after implantation patients can't live without glasses. This problem was solved by making accommodative lenses [2]. But these lenses have small opportunities in correction of aberrations. Patients can receive the most natural sight after implantation of hybrid IOLs. Investigation of multifocal lenses, and, particularly, IOLs is object of interest in many countries from the 80th to our days [3-11]. A review of the most interesting hybrid IOLs, their processing is a subject of this paper. Also there is proposed a method of processing IOLs by using volume holograms.

## 2. Review of Russian and foreign lenses

### 2.1. Foreign lenses

There are well known American lenses «AcrySof ReSTOR» of the «Alcon» corporation from the USA, lenses «AcriLisa» of the German company «AcriTec», lenses «Tecnis ZM900» of «AMO» [12-15].

These lenses correspond to the next **standard claims**:

- Lenses must be soft for the implantation through the small section;
- Material must be hydrophobic for minimization of the treats and of the appearance of biological concretions on lenses;
- Additive optical power, formed by diffractive structure, is near +4 diopters – for reducing the intensity of the defocused image;
- Lens must adsorb UV radiation because it can treat retina.

Advantage of lenses AcrySof ReSTOR is an opportunity of the far sight in different illumination. In case of diameter of pupil equal 3.5 mm the refraction part of IOL begins working and the energy moves to the far focus. The effect of blinding by headlights in case of night driving was eliminated by reducing the size of the central zone. Despite of the small increasing of the number of zones, the profit in energy is not significant. Also the small size of the diffractive zone is a source of increasing a sensitivity of the pupil's center relatively to axis.

The redistribution of the light energy in lenses ReSTOR is made by the reduction of the depth of diffractive relief. The radius of the central zone is

$$r_0 = \sqrt{2\lambda_0 f} \quad (1)$$

where  $\lambda_0$  is constructive wavelength,  $f$  is a focus of lens in the 1<sup>st</sup> order of diffraction. The radius of the central zone can be reduced by using a phase shift

$$r_k^2 = r_0^2 + 2k\lambda f \quad (2)$$

But the next condition have place

$$r_2^2 - r_1^2 = r_k^2 - r_{k-1}^2 \quad (3)$$

It means that the number of zones can't be increased more than on one zone and there is no any significant effect [12-20].

Hybrid multifocal IOL AcriLisa (Acri. Tec GmbH, Германия) is a monolithic aspherical bifocal IOL with the correction of aberrations, which is processed of hydrophobic acryl. 65% of intensity comes to the far focus and 35% - to the near focus. Bifocal work is independent from the size and function of pupil, because the diffractive structure is on full light diameter. The refractive component has aspherical form.

Diffractive IOL of silicon with three components Tecnis ZM900 (Advanced Medical Optics, Inc., США) has a diameter of optical part equal 6 mm. Diffractive structure on back surface has the additional optical power +4 diopters, the incident light is distributed homogenously free of size and function of pupil. The first surface have aspherical form.

## 2.2. Russian lenses

A group of scientists in Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences has developed the first Russian IOL [16-20]. A great quality of far and near sight, IOL's independence on pupil. The function of correction of eye's and IOL's aberrations has added. Reverse slopes for decreasing the risk of appearance of concretions has added. An optical part of lens has plano-convex shape with the triangle profile and ring microstructure on back surface (fig. 1). The IOLs «MIOL-Accord» are processed in Hizhniy Novgorod by the company «Reper-NN» (fig. 2). The developing has done by the Institute of Automation and Electrometry SB RAS, Novosibirian branch of the MNTK «Eye microsurgery» and private corporation «Intra OL».

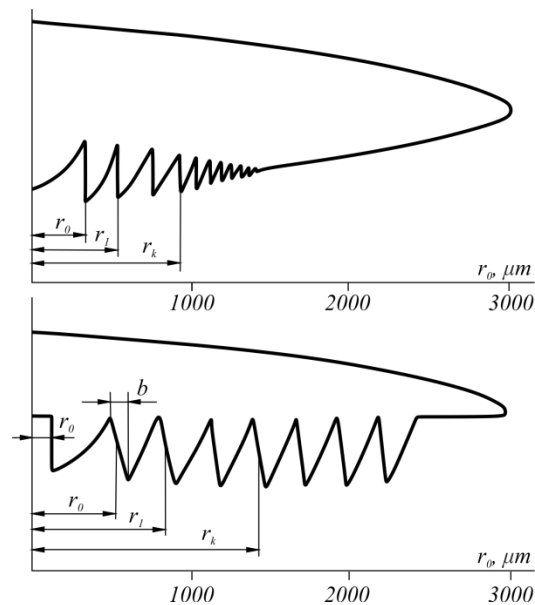


Fig. 1. (a) ReSTOR; (b) MIOL-Accord.  $r$  is radial coordinate,  $r_0$ ,  $r_l$ ,  $r_k$  are radiuses of central diffractive zones.

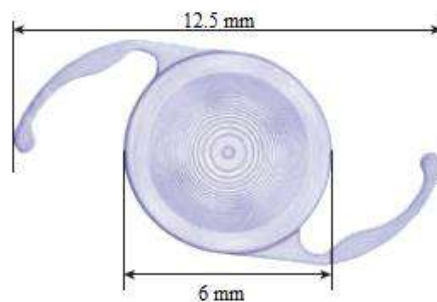


Fig. 2. Diffractive-refractive lens MIOL-Accord.

Forming of IOL can be made by photo-solidification of liquid oligomeres, which can be polymerized. Polymerizing lasts as crystal growth. Structure of polymer and absence of mechanical processing decrease the risk of appearance of concretions. The material has good bio-compatibility. The mold is a matrix of quartz with the diffractive micro-structure, processed by the method of direct laser recording, where the shape of the beam can be changed (fig. 3, 4).

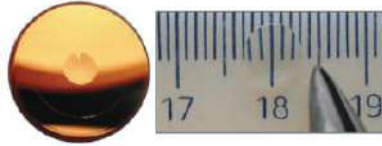


Fig. 3. Diffractive matrix and a final lens.

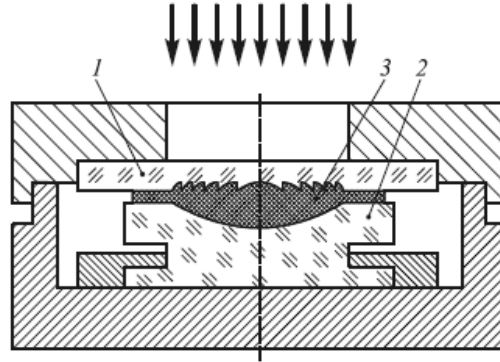


Fig. 4. Pressing of IOL by diffractive mold. where 1 is a diffractive matrix, 2 is another part of mold, 3 is a lens.

For compensating of decreasing of the diffraction efficiency the height of peaks was increased. The diffractive structure is situated on the entire light diameter. This construction effectively spread light equally between the focuses independent of pupil. If the IOL is decentered, the cutting of diffraction zones can't take place. For minimizing the blinding when the pupil diminishing the mini-zone has added. The curvature is the same as the curvature of the main base with diffractive structure. IOL has the ability of correction of refractive components of retina, IOL and vitreous body.

In comparison with lenses ReSTOR lenses MIOL-Accord have increasing square of every diffractive zone

$$M = c\lambda / f \quad (4)$$

where  $c$  is a non-dimensional aberration coefficient. Increasing of zones is a result of the correction of aberrations. The model of eye is Lotmar's.

Later, in company «Reper-NN» hybrid lenses with rectangle profile of peaks were developed. It can give the possibility of using three orders of diffraction. A maximum of -1<sup>st</sup> order can be used for forming images of far objects, a 0<sup>th</sup> maximum – for average distances (500 mm) and a 1<sup>st</sup> order – for near objects (250 mm). For comparison: triangle profile, which is more widespread, can give only focusing in two orders of diffraction, also the diffraction efficiency is higher. Thus increasing the number of focuses improves vision on any distances [21-22].

IPSI RAS has many investigations in counting [23-36] and processing [37-55] of diffractive optics. IPSI RAS and Laser Center of Hannover have investigated the two photon polymerization technology for processing microdevices such as IOLs [56-58]. The element is three-focal, their diameter equals 2.7 mm, focal lengths is between 27 and 34 mm. The size of the element's section is less than the wavelength. In comparison with the method of diamond turning it is more economic. It gives the possibility to focus complicated 3D structures. A binary structure has processed and their features were analyzed. A height of stair of microrelief can be counted as

$$h = \frac{\text{mod}_{2\pi}(\varphi)}{k(n-1)} \quad (5)$$

where  $k$  is a wavenumber,  $n$  is a refraction coefficient,  $\text{mod}_{2\pi}(\varphi)$  is an excess of division eikonal to  $2\pi$ .

Complex amplitudes of given and modified waves

$$W_0(\rho) = \sqrt{I_0(\rho)} \exp[i\varphi_0(\rho)] \quad (6)$$

$$W(\rho) = \sqrt{I_0(\rho)} \exp\left(-\frac{ik\rho^2}{2f_1} + i\Phi[\text{mod}_{2\pi}\left(\frac{k\rho^2}{2f_2}\right)]\right)$$



where  $\rho$  is a current radius of an element,  $f_1$  and  $f_2$  are focal lengths. Focuses of the element are

$$F_{-1} = \frac{f_1 f_2}{f_1 - f_2} \quad (7)$$

$$F_0 = f_1$$

$$F_1 = \frac{f_1 f_2}{f_1 + f_2}$$

The distribution of intensity is

$$I(r, z) = \left| \frac{k}{z} \int_0^R W(\rho) \exp \left\{ i \left[ \varphi_{mf}(\rho) + \frac{k\rho^2}{2z} \right] \right\} J_0 \left( \frac{kr\rho}{z} \right) \rho d\rho \right|^2 \quad (8)$$

where  $z$  is a longitudinal coordinate,  $\varphi_{mf}(\rho)$  is eikonal,  $r$  is a radial coordinate. The graph is on the fig. 5.

Later these organizations considered an ability of constructing of diffraction relief with sub-micron height and sine profile. The possibility of processing of three-focal hybrid IOLs with the help of nanoprint technique was considered. The distribution of intensity between focuses can be counted before. Focal powers are -3, 0 and 3 diopters. In comparison with the method of two-photon polymerization this method is more precise and fast. The theoretical results are agreed with the experimental [59].

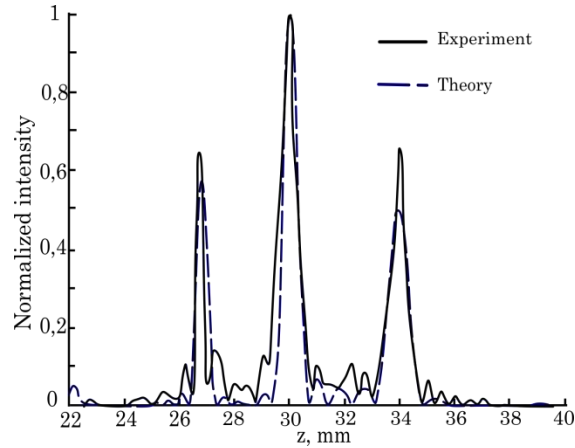


Fig. 5. The distribution of intensity along the optical axis.

### 3. Methods of eliminating chromatism

#### 3.1 Multi-order diffractive lens

In Kyiv in 2015 an IOL with decreased chromatism has investigated [60-61]. It is a multi-order lens. These lenses has a diminished chromatism. These lenses have an increased in  $p$  times thickness.

A matrix of focal lengths can be determined as

$$f_N = \frac{p f_0 \lambda_0}{N \lambda} \quad (9)$$

where  $f_0$  is a focal length for the main wavelength  $\lambda_0$ ;  $N$  is a main order of diffraction;  $\lambda \neq \lambda_0$ ;  $p$  is a parameter. The meaning of the equation is that if  $p \lambda_0 / N \lambda = 1$ , some of the wavelengths can be focused in 1 point with the big diffraction efficiency. It can be determined as

$$\eta_N = \text{sinc}^2(\alpha \mu p - N) \quad (10)$$

where  $\alpha = \frac{\lambda_0[n(\lambda)-1]}{\lambda[n(\lambda_0)-1]}$  is a relative phase retard when  $\lambda \neq \lambda_0$ ,  $\mu = t'/t$  is a thickness coefficient, where  $t'$  and  $t$  are real and counted thicknesses of profile, respectively. A  $\mu$  coefficient can't influence on the location of focuses, but it can change the distribution of energy between them. In article [ ] it equals 1. When the  $p$  increase, the number of wavelengths, which can satisfy the condition of appearance of maximum, increase. For the main maximum  $p = N$  and the wavelength is  $\lambda_0$ . The good meaning of for visible light is  $p=6$ . On the fig. 6 the relation between the diffraction efficiency and the wavelength is shown. When  $p=20$ , chromatism is as bigger as in spherical lenses. The dispersion for three wavelengths can be determined as

$$V_\lambda = \frac{N_3 \lambda_3}{N_c \lambda_c - N_k \lambda_k} \tag{11}$$

On the fig. 7 the relation between the diffraction efficiency and focal length is shown. The numbers of orders of diffraction cannot be agreed. Two groups of wavelengths are considered:  $\lambda_b = 485$  nm,  $\lambda_g = 573$  nm,  $\lambda_r = 700$  nm and  $\lambda_b = 420$  nm,  $\lambda_g = 485$  nm,  $\lambda_r = 573$  nm. For the near and far focuses the dispersions are -32, 74 and 38.8 respectively. But the dispersion of the usual lens equals -3.5. Thus, chromatism of multi-ordered lens is significantly less than chromatism of refractive lens. These lenses have infinite accommodation.

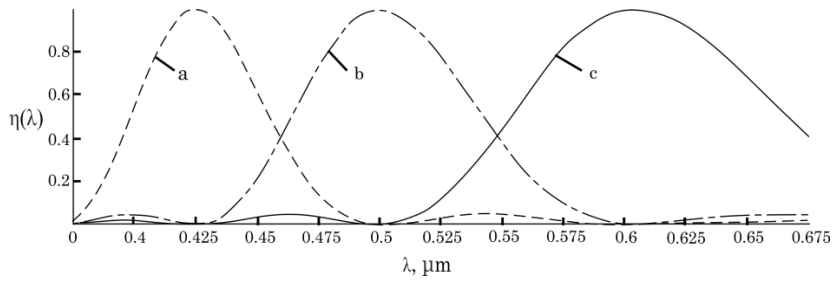


Fig. 6. Relation between the diffraction efficiency and the wavelength. a)  $N = 7$ ; b)  $N = 6$ ; c)  $N = 5$ .

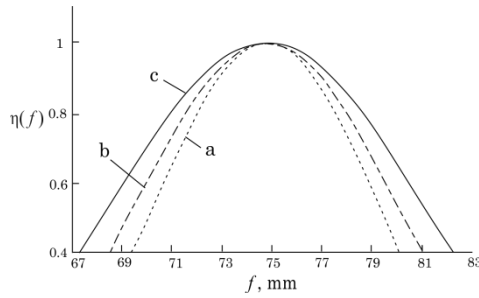


Fig. 7. Distribution of light along the optical axis: a)  $N = 7$  (синий); b)  $N = 6$  (зелёный); c)  $N = 5$  (красный).

For proposed IOL the parameters are:  $f = 100$  mm,  $p=6$ , the material is PMMA,  $\lambda_0 = 525$  nm. In light diameter  $D = 7$  mm 19 diffractive zones are situated, maximum depth of the groove is 6  $\mu$ m. The anterior surface of lens is spherical, the model of eye is taken from Gullstrand [24].

### 3.2 Holographic approach for the creating of intraocular lenses

It is possible to use volume holograms for clearance of chromatism. The direction of rays is significant and the recording can be made by composing object and referent waves in the photosensitive layer [62-64]. The recording is comparatively fast. The required meanings of aberrations can be created by the methods of computer optics.

Holograms receive the features of volume holograms with the height of the layer near  $\sim 7$   $\mu$ m. In this case holograms have only the virtual image. Changing the scheme of recording give us a real image instead of virtual. The holograms are also phase, but their surface is smooth. These holograms can't have significant chromatism because the Bragg's condition haves place:

$$2d \cdot \sin \theta = n\lambda \tag{12}$$

where  $d$  is a period of grating,  $\theta$  is an angle between the ray and the normal to surface,  $n$  is the order of diffraction,  $\lambda$  is a wavelength.

The good materials are bichromated gelatina, silver holograms like «PFG-01», «PFG-03», «Ultimate-08» [65], «Ultimate U-04». Photo-thermo-refractive glasses from the university IFMO are interesting medium (or matrix) for photosensitive matter, but

glass is solid material [66]. Instead of glass nanoporous acryl can be used.

Bichromated gelatina can be used for any surface, the holograms have great resolution and good spectral selectivity [67]. The material needs a protective coating. But the hologram can change features at the temperature 34 °C so this material can't be used.

Silver holograms are more convenient. The methods of increasing their diffractive efficiency are investigated by IFMO [68-69]. The material with the great color transfer Ultimate-08 is developed by Alkiss Lembessis.

The scheme of recording and reconstructing of a point source is on fig. 8. For receiving a real image the source must be imaginary. It means that the recording wave is divergent. Using optical elements or phase holograms for forming desired aberrations in IOL gives an optical system, which aberrations are compensated.

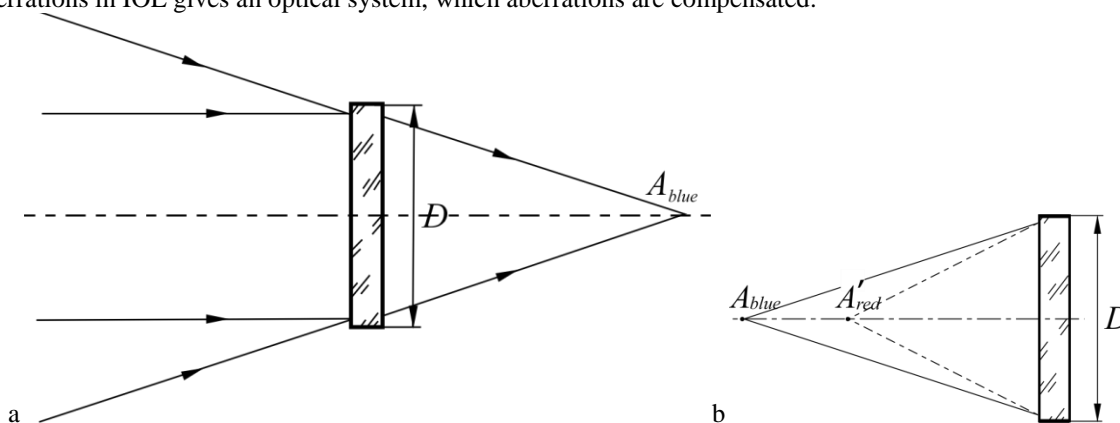


Fig. 8. (a) Recording with blue light (for example). (b) Change of the focus while the wavelength changes (blue – record, red - reconstruction).

Phase holograms (not volume) are another way for making of intraocular lenses. Image, which is made by using the hologram, can be stretched and the position can be changed. The photosensitive medium has the possibility to record three or more holograms and for switching between them. The medium is polydimethylsiloxane with gold nanorods [70]. Using of these surfaces can make the possibility of decreasing errors of an eye's aberrations.

#### 4. Conclusion

A review of existing intraocular lenses (IOLs) has shown that lenses can imitate crystalline with good quality, they can't be perfect in all directions like monochromatic aberrations, chromatic aberrations, diffractive efficiency, quantity of focuses simultaneously. Perhaps, the best criteria of the IOL's quality is optical performance, what means that IOL must give an image on retina, which must be as close to real image as possible. Some problems like chromatism or diffractive efficiency are actual now, also the decisions are exist. A proposed method of processing volume holograms as a diffractive part of hybrid intraocular lenses in comparison with others methods is faster and it can be done without complicate devices. The proposed method is interesting because intraocular lens has no significant chromatism. The diffraction efficiency is good for intraocular optics. Other aberrations can be reproduced by methods of computer optics. Traditional optics like lenses and plates can be used for forming aberrations, too. Thus, the analysis shows that the method of using volume holograms is perspective for further investigations.

#### References

- [1] Winther-nielsen A, Gyldenkerne G, Corydon L. Contrast sensitivity, glare, and visual function: Diffractive multifocal versus bilateral monofocal intraocular lenses. *Journal of cataract and refractive surgery* 1995; 21(2): 202–207.
- [2] Tunç Z. Developments in accommodating intraocular lenses. *Türk Oftalmoloji Dergisi* 2012; 42(4): 288–293.
- [3] Futhey JA. Diffractive bifocal intraocular lens. *SPIE* 1989; 1052: 142.
- [4] Cosoburd T, Kedmi J, Grossinger Israel, Levy U. Diffractive multi-focal lens 1998; US 5760871.
- [5] Grossinger I, Golub M. Simultaneous multifocal lens and method of utilising same for treating visual disorders US 6364483, Acc. Apr. 2, 2002.
- [6] Michael Morris G, Dale A. Buralli, Richard J. Federico. Bifocal multiorder diffractive lenses for vision correction US 2005/0057720 A1, Accepted Mar. 17, 2005.
- [7] Greisukh GI, Ezhov EG, Stepanov SA. Comparative analysis of chromatism of diffractive and refractive lenses. *Computer Optics* 2005; 28: 60–65.
- [8] Morris GM, Buralli DA, Federico RJ. Bifocal multiorder diffractive lenses for vision correction. No 7093938 B2, 2006.
- [9] Morris GM, Buralli DA, Federico RJ. Diffractive lenses for vision correction US 7 156 516 B2, Accepted Jan. 2, 2007.
- [10] Plainis S, Atchison DA, Charman WN. Power Profiles of Multifocal Contact Lenses and Their Interpretation. *Optometry and Vision Science* 2013; 90(10): 1066–1077.
- [11] Zolotarev AV, Karlova EV, Kotova SP, Patlan' VV, Russkov KN, Samagin SA, Sapsina TN. Depth of Focus of Intraocular Lenses with Higher-Order Aberrations. *Bulletin of the Lebedev Physics Institute* 2013; 12. DOI: 10.3103/S1068335613120038.
- [12] Takhtaev YuV, Balashevich LI. Surgical correction of hypermetropia and presbyopia by using refractive-diffractive pseudoaccommodative lenses AcrySof Restor. *Ophthalmology* 2005; 3: 12–16.
- [13] Kohnen T, Allen D, Boureau C. European multicenter study of the AcrySof ReSTOR apodized diffractive intraocular lens. *Ophthalmology* 2006; 113: 578–584.
- [14] Souza CE, Muccioli C, Soriano ES. Visual Performance of AcrySof ReSTOR apodized diffractive IOL: a prospective comparative trial. *Am J Ophthalmol* 2006; 141: 827–832. DOI: 10.1016/j.ajo.2005.12.031.

- [15] Alfonso JF, Fernández-Vega L, Señaris A, Montés-Micó R. Prospective study of the AcriLISA bifocal intraocular lens. *J Cataract Refract Surg* 2007; 33: 1930–1935. DOI: 10.1016/j.jcrs.2007.06.067.
- [16] Lenkova GA. The Effect of Phase Profile Depth on Intensity Distribution in Diffraction Orders of a Bifocal Element, *Optoelectr. Instrum. Data Process* 1995; 5: 15–22.
- [17] Lenkova GA, Myznik MM. Spherochromatic Aberrations of a Model Eye with Bifocal Hybrid Intraocular Lens Refraction, *Optoelectr. Instrum. Data Process* 2001; 5: 73–81.
- [18] Iskakov IA, Koronkevich VP, Lenkova GA, Korol'kov VP. Russian bifocal diffractive-refractive IOL structure: design, optical properties. *Vestnik OGU* 2007; 12: 85–88.
- [19] Koronkevich VP, Lenkova GA, Korol'kov VP, Poleshchuk AG, Iskakov IA, Gutman A, Treushnikov VM. Bifocal intraocular lens instead of crystalline. *Photonics* 2008; 1: 10–13.
- [20] Koronkevich VP, Lenkova GA, Korol'kov VP, Poleshchuk AG, Iskakov IA, Gutman A. New-generation bifocal diffractive-refractive intraocular lenses. *Computer Optics* 2008; 32(1): 50–58.
- [21] Lenkova GA. Chromatic aberrations of diffractive-refractive intraocular lenses in an eye model. *Optoelectronics, Instrumentation and Data Processing: April 2009*; 45(2): 171–183. DOI: 10.3103/S8756699009020113.
- [22] Lenkova GA. Specific Features of Measuring the Optical Power of Artificial Refractive and Diffractive-Refractive Eye Lenses. *Optics and Spectroscopy* 2016; 121(2): 335–347.
- [23] Golub MA, Karpeev SV, Prokhorov AM, Sisakyan IN, Soifer VA. Focusing light into a specified volume by computer synthesized holograms. *Soviet Technical Physics Letters* 1981; 7(10): 264–266.
- [24] Golub MA, Kazanskiy NL, Sisakyan IN, Soifer VA, Uspleneyev GV, Yakunenkova DM. Multigradation Fresnel Lens. *Soviet Technical Physics* 1991; 61(4): 195–197.
- [25] Golub MA, Doskolovich LL, Kazanskiy NL, Kharitonov SI, Soifer VA. Computer generated diffractive multi-focal lens. *Journal of Modern Optics* 1992; 39(6) 1245–1251. DOI: 10.1080/713823549.
- [26] Soifer VA, Doskolovich LL, Kazanskiy NL. Multifocal diffractive elements. *Optical Engineering* 1994; 33(11): 3610–3615. DOI: 10.1117/12.179890.
- [27] Doskolovich LL, Kazanskiy NL, Kharitonov SI, Tzaregorodtzev AY. A method for estimating the DOE's energy efficiency. *Optics and Laser Technology* 1995; 27(4): 219–221.
- [28] Doskolovich LL, Golub MA, Kazanskiy NL, Khramov AG, Pavelyev VS, Seraphimovich PG, Soifer VA, Volotovskiy SG. Software on diffractive optics and computer generated holograms. *Proceedings of SPIE* 1995; 2363: 278–284. DOI: 10.1117/12.199645.
- [29] Doskolovich LL, Kazanskiy NL, Soifer VA, Perlo P, Repetto P. Design of DOEs for wavelength division and focusing. *Journal of Modern Optics* 2005; 52(6) 917–926. DOI: 10.1080/09500340512331313953.
- [30] Golovashkin DL, Kasanskiy NL. Solving Diffractive Optics Problems using Graphics Processing Units. *Optical Memory and Neural Networks (Information Optics)* 2011; 20(2) 85–89. DOI: 10.1134/S1063776110120095.
- [31] Kazanskiy NL, Skidanov RV. Diffractive beam splitter. *Computer Optics* 2011; 35(3): 329–335.
- [32] Khonina SN, Ustinov AV, Skidanov RV. Binary lens: investigation of local focuses. *Computer optics* 2011; 35(3): 339–346.
- [33] Kazanskiy N, Skidanov R. Binary beam splitter. *Applied Optics* 2012; 51(14): 2672–2677. DOI: 10.1364/AO.51.002672.
- [34] Serafimovich PG. Diffraction analysis of focusing optical elements. *Proc. SPIE* 2013; 9156. DOI: 10.1117/12.2054492.
- [35] Kazanskiy NL, Khonina SN, Skidanov RV, Morozov AA, Kharitonov SI, Volotovskiy SG. Formation of images using multilevel diffractive lens. *Computer Optics* 2014; 38(3): 425–434.
- [36] Soifer VA, Doskolovich LL, Golub MA, Kazanskiy NL, Kharitonov SI, Perlo PP. Multifocal and combined diffractive elements (invited paper). *Proceedings SPIE* 1993; 1992: 226–234.
- [37] Soifer VA, Kotlyar VV, Kazanskiy NL, Doskolovich LL, Kharitonov SI, Khonina SN, Pavelyev VS, Skidanov RV, Volkov AV, Golovashkin DL, Solovyev VS, Uspleneyev GV. *Methods for Computer Design of Diffractive Optical Elements*. John Wiley & Sons, Inc., New York, 2002.
- [38] Golovashkin DL, Kotlyar VV, Soifer VA, Doskolovich LL, Kazanskiy NL, Pavelyev VS, Khonina SN, Skidanov RV. *Computer Design of Diffractive Optics*. Cambridge Inter. Scien. Pub. Ltd & Woodhead Pub. Ltd., 2012.
- [39] Popov VV. Materials and methods for flat optical elements. *Computer Optics* 1989; 1(1): 125–128.
- [40] Berezny AE, Karpeev SV, Uspleneyev GV. Computer-generated holographic optical elements produced by photolithography. *Optics and Lasers in Engineering* 1991; 15(5): 331–340.
- [41] Volkov AV, Kazanskiy NL, Moiseev OYu, Soifer VA. A method for the diffractive microrelief formation using the layered photoresist growth. *Optics and Lasers in Engineering* 1998; 29(4-5) 281–288. DOI: 10.1016/s0143-8166(97)00116-4.
- [42] Kazanskiy NL, Uspleneyev GV, Volkov AV. Fabricating and testing diffractive optical elements focusing into a ring and into a twin-spot. *Proceedings of SPIE* 2000; 4316: 193–199. DOI: 10.1117/12.407678.
- [43] Kazanskiy NL, Kolpakov VA, Kolpakov AI. Anisotropic etching of SiO<sub>2</sub> in high-voltage gas-discharge plasmas. *Russian Microelectronics* 2004; 33(3): 169–182. DOI: 10.1023/B:RUMI.0000026175.29416.eb.
- [44] Doskolovich LL, Kazanskiy NL, Repetto P, Tyavin YeV. Design and investigation of colour separation diffraction gratings. *Journal of Optics* 2007; 9(2): 123–127. DOI: 10.1088/1464-4258/9/2/001.
- [45] Pavelyev VS, Borodin SA, Kazanskiy NL, Kostyuk, GF, Volkov AV. Formation of diffractive microrelief on diamond film surface. *Optics & Laser Technology* 2007; 39(6): 1234–1238. DOI: 10.1016/j.optlastec.2006.08.004.
- [46] Kazanskiy NL, Murzin SP, Osetrov YeL, Tregub VI. Synthesis of nanoporous structures in metallic materials under laser action. *Optics and Lasers in Engineering* 2011; 49(11) 1264–1267. DOI: 10.1016/j.optlaseng.2011.07.001.
- [47] Abul'khanov SR, Kazanskiy NL, Doskolovich LL, Kazakova OY. Manufacture of diffractive optical elements by cutting on numerically controlled machine tools. *Russian Engineering Research* 2011; 31(12): 1268–1272. DOI: 10.3103/S1068798X11120033.
- [48] Kazanskiy NL. Research & education center of diffractive optics. *Proceedings of SPIE* 2012; 8410: 84100R. DOI: 10.1117/12.923233.
- [49] Kazanskiy NL, Kolpakov VA, Podlipnov VV. Gas discharge devices generating the directed fluxes of off-electrode plasma. *Vacuum* 2014; 101: 291–297. DOI: 10.1016/j.vacuum.2013.09.014.
- [50] Kazanskiy NL, Moiseev OYu, Poletayev SD. Microprofile Formation by Thermal Oxidation of Molybdenum Films. *Technical Physics Letters* 2016; 42(2): 164–166. DOI: 10.1134/S1063785016020085.
- [51] Porfirev AP, Khonina SN. Experimental investigation of multi-order diffractive optical elements matched with two types of Zernike functions. *Proceedings of SPIE* 2016; 9807: 98070E. DOI: 10.1117/12.2231378.
- [52] Podlipnov VV, Kolpakov VA, Kazanskiy NL. Etching of silicon dioxide in off-electrode plasma using a chrome mask. *Computer Optics* 2016; 40(6): 830–836. DOI: 10.18287/2412-6179-2016-40-6-830-836.
- [53] Verma P, Zaman KK, Khonina SN, Kazanskiy NL, Gopal R. Ultraviolet-LIGA-based fabrication and characterization of a nonresonant drive-mode vibratory gyro/accelerometer. *Journal of Micro/ Nanolithography, MEMS, and MOEMS* 2016; 15(3): 035001.
- [54] Kazanskiy NL, Stepanenko IS, Khaimovich AI, Kravchenko SV, Byzov EV, Moiseev MA. Injectional multilens molding parameters optimization. *Computer Optics* 2016; 40(2): 203–214. DOI: 10.18287/2412-6179-2016-40-2-203-214.
- [55] Kazanskiy NL, Kolpakov VA. *Optical Materials: Microstructuring Surfaces with Off-Electrode Plasma*. CRC Press, 2017.
- [56] Osipov V, Doskolovich LL, Bezus EA, Cheng W, Gaidukeviciute A, Chichkov B. Fabrication of three-focal diffractive lenses by two-photon polymerization technique. *Applied Physics A: Materials Science and Processing* 2012; 107(3): 525–529. DOI: 10.1007/s00339-012-6903-9.

- [57] Osipov V, Doskolovich LL, Bezus EA, Drew T, Zhou K. Application of nanoimprinting technique for fabrication of trifocal diffractive lens with sine-like radial profile. *Journal of Biomedical Optics* 2015; 20(2): 025008. DOI: 10.1117/1.JBO.20.2.025008.
- [58] Hinze U, El-Tamer A, Doskolovich LL, Bezus EA, Reiß S. Additive manufacturing of a trifocal diffractive-refractive lens. *Optics Communications* 2016; 372: 235–240.
- [59] Bezus EA, Doskolovich LL, Kazanskiy NL. Evanescent-wave interferometric nanoscale photolithography using guided-mode resonant gratings. *Microelectronic Engineering* 2011; 88(2): 170–174. DOI: 10.1016/j.mee.2010.10.006.
- [60] Kolobrodov VG, Tymchik GS, Siryi IaA. Quality assessment of multifocal diffractive lens images. *Devices and Methods of Measurements* 2014; 8(1): 115–118.
- [61] Kolobrodov VG, Tymchik GS, Kuchugura IO. Design of the multiorder intraocular lenses. *Devices and Methods of Measurements* 2015; 8(1): 204–210.
- [62] Collier RJ, Burkhardt CB, Lin LH. *Optical Holography*. Academic Press, New York, 1971.
- [63] Ostrovsky YuI. *Holography and Its applications* (Translated from Russian by G. Leib). Moscow: Mir Publ., 1977.
- [64] Bobrov ST, Greysukh GI, Turkevich YuG. *Optics of diffractive elements and systems*. Leningrad: Mashinostroenie, 1986.
- [65] Sarakinos A, Lembessis A, Zervos N. OptoClones and HoLoFoS: advances in colour display holography, *Holography. Science and practice, Proceedings of the 10th International conference HoloExpo 2013*; 124–125.
- [66] Nikonorov NV. New photo-thermo-refractive glasses for volume holograms recording: properties, technologies and applications, *Holography. Science and Practice, 13-th International Conference HoloExpo 2016*; 68–70.
- [67] Gornostay AV, Odinkov SB. A method to design a diffractive laser beam splitter with color separation based on bichromated gelatine. *Computer Optics* 2016; 40(1): 45-50. DOI: 10.18287/2412-6179-2016-40-1-45-50.
- [68] Andreeva OV, Andreeva NV, Kuzmina TB. Plasmonic particles of colloidal silver in high-resolution recording media. *Optics and Spectroscopy* 2017; 122(1): 52–58.
- [69] Pshenova AS, DA, Klyukin, Nashchekin AV, Sidorov AI. Migration of silver on the nanoporous glasses surface under the action of an electric field. *Applied Optics* 2017; 56(10): 2821–2825.
- [70] Malek SC, Ee H-S, Agarwal R. Strain Multiplexed metasurface holograms on a stretchable substrate. *Nano Letters, Article ASAP*, Publication date (web), 2017. DOI: 10.1021/acs.nanolett.7b00807.

# Wavefront aberration analysis with a multi-order diffractive optical element

P.A. Khorin<sup>1</sup>, S.A. Degtyarev<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

## Abstract

In this paper we show an ability to use a multi-order (multi-channel) diffractive optical element for wavefront relief function expansion in terms of Zernike polynomials. This approach can be successfully used for small meanings of wavefront aberrations when the wavefront relief can be represented as a linear superposition of Zernike polynomials. Unfortunately, linear approximation is becoming unworkable with increasing of aberration meanings. In this work we study an applicability of this Zernike expansion method.

**Keywords:** Zernike polynomials; measurements of wavefront aberrations; multi-order diffractive optical element

## 1. Introduction

Light field phase retrieval is one of the fundamental problems of signal analysis processing [1]. Nowadays we do not have a direct method of phase registration; consequently phase can be measured indirectly through light intensity analysis. For example, wavefront can be retrieved with interferometrical methods accompanied by subsequent calculation algorithms. Once more method is implemented in Shack-Hartmann wavefront sensor. Usually, it consists of an array of pinholes or microlenses; each of them plots a tilt of analyzed wavefront into sensor matrix [2]. Phase retrieval can be also implemented with diffractive optical elements (DOEs) which can expand analyzed light field into an orthogonal basis [3-5].

Commonly used wavefront representation is an expansion into Zernike polynomials [6]. Weight coefficients of the wavefront expansion make it possible to calculate root-mean-square deviation from the ideal spherical wavefront. In this case every weight coefficient is associated with a certain aberration. Thus, the weight coefficient with high value automatically points into an aberration which mostly deforms the wavefront.

In this work we propose a diffractive optical element (DOE) which is matched with Zernike polynomials basis [7, 8, 9]. This element has been successfully used for weakly aberrated wavefront analysis [10-11].

Each Zernike polynomial corresponds to a certain optical wavefront aberration. Conventional wavefront aberration type can be described as a certain Zernike polynomial with orders  $n$  and  $m$  like it is shown in Table 1.

Table 1. Correspondence between Zernike polynomials and conventional wavefront aberration types.

No	$n$	$m$	Zernike polynomial	Aberration type
1	0	0	1	Constant
2	1	-1	$2r \sin(\theta)$	Tilt
3	1	1	$2r \cos(\theta)$	Tilt
4	2	-2	$\sqrt{6}r^2 \sin(2\theta)$	Astigmatism
5	2	0	$\sqrt{3}(2r^2 - 1)$	Defocus
6	2	2	$\sqrt{6}r^2 \cos(2\theta)$	Astigmatism
7	3	-3	$2\sqrt{2}r^3 \sin(3\theta)$	(Trefoil)
8	3	-1	$2\sqrt{2}(3r^3 - 2r) \sin(\theta)$	Pure coma
9	3	1	$2\sqrt{2}(3r^3 - 2r) \cos(\theta)$	Pure coma
10	3	3	$2\sqrt{2}r^3 \cos(3\theta)$	(Trefoil)
11	4	-4	$\sqrt{10}r^4 \sin(4\theta)$	Quadrofoil
12	4	-2	$\sqrt{10}(4r^4 - 3r^2) \sin(2\theta)$	2 <sup>th</sup> order Astigmatism
13	4	0	$\sqrt{5}(6r^4 - 6r^2 + 1)$	Spherical
14	4	2	$\sqrt{10}(4r^4 - 3r^2) \cos(2\theta)$	2 <sup>th</sup> order Astigmatism
15	4	4	$\sqrt{10}r^4 \cos(4\theta)$	Quadrofoil

In this work we assume that Zernike polynomials appear as follows:

$$\Psi_{nm}(r, \varphi) = \sqrt{\frac{n+1}{\pi r_0^2}} R_n^m(r) \begin{cases} \cos(m\varphi) \\ \sin(m\varphi) \end{cases}, \quad (1)$$

here  $R_n^m(r)$  is radial Zernike polynomials:

$$R_n^m(r) = \sum_{p=0}^{(n-m)/2} \frac{(-1)^p (n-p)!}{p! \left(\frac{n+m}{2} - p\right)! \left(\frac{n-m}{2} - p\right)!} \left(\frac{r}{R}\right)^{n-2p}$$

Wavefront aberrations  $W(r, \varphi)$  are commonly described in terms of Zernike polynomials in the following way:

$$W(r, \varphi) = \exp[iw(r, \varphi)], \tag{2}$$

$$w(r, \varphi) = \sum_{n=0}^N \sum_{m=-n}^n c_{nm} \Psi_{nm}(r, \varphi). \tag{3}$$

There is so-called ‘‘Zernike pyramid’’ in figure 1. 2D patterns of few Zernike polynomials form this pyramid. In a vertical direction radial number  $n$  varies from 0 to 4 ( $n = 0$  to 4) and in a horizontal direction azimuthal number  $m$  varies from  $-n$  to  $n$  ( $m = -n$  to  $n$ ).

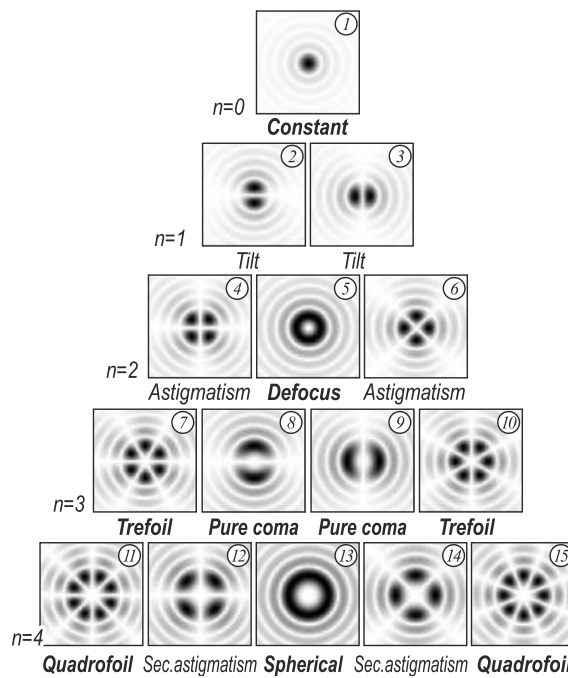


Fig. 1. Zernike pyramid: 2D pattern of few first Zernike polynomials.

For plotting the point-spread function we use a model of Fourier optical correlator (figure 2). This simple optical arrangement is simulated with Zemax ray-tracing software [12].

Control of optical image quality involves measurements optical system image photometric parameters such as spread function of optical system (for example, point-spread function). Practically, these image photometric parameters quantitatively characterize optical system image quality. State-of-the-art theory has been carefully developed, thus, a certain set of parameters completely describes the quality of an optical system image.

Experimentally received point-spread function characterizes the quality of an optical system. It describes quite thoroughly optical system characteristics including wavefront and optical surfaces microrelief.

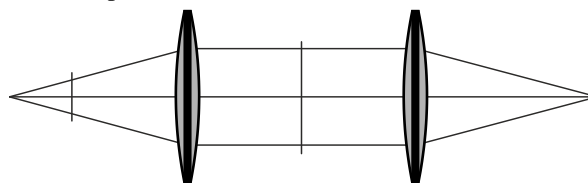


Fig. 2. Fourier-correlator optical arrangement.

Optical elements work together and create an image. Unfortunately, created with an optical system image can not be ideal. Lenses and mirrors have their own aberrations; in addition, aberrations can arise due to inaccuracies of fabrication, misalignments, and so on. Partial error compensation can be provided after precise measurements of wavefront aberrations.

In our simulations we introduce aberrations into the optical system through the adding to the relief of the first surface of the second lens. These approach and algorithm are discussed in detail in paper [13]. 2D patterns of point-spread function are shown in figure 3. This patterns are arranged as a pyramid corresponding to Zernike pyramid in figure 1.

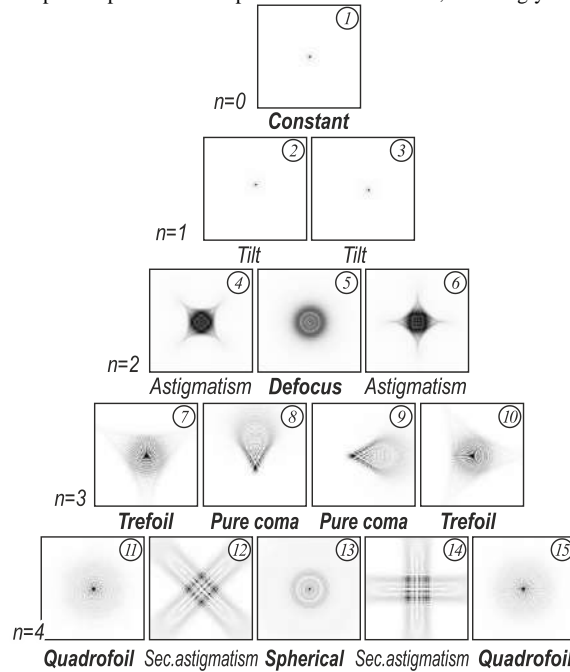


Fig. 3. Point-spread functions 2D patterns that are corresponding to conventional aberration types.

## 2. Wavefront aberration analysis

In the paper [14] authors investigate an ability of using multi-order diffractive optical element for analysis of human eye optical system aberrations [15, 16]. Note that, this way may be used for other optical systems, including telescopes [17-19].

It should be noticed that Zernike bases can differ not only in indexing and normalization but in the angular dependence shape. In particular, in [20] exponential and trigonometric angular dependences are used for designing multi-order diffractive optical elements for wavefront analysis.

Optical Zernike analyzer is a combination of diffractive optical element and lens; DOE is matched with basis function and lens does Fourier transform. Achieved Fourier spectrum at the center has the value which means scalar product of input light function and the basis function. Experimental testing of the same devise is described in [20, 21].

Certain basis Zernike function is coded in each order of multi-order diffractive optical element. DOE transmission function looks as follows:

$$\tau(x, y) = \sum_{p=0}^P \sum_{q=0}^Q \Psi_{pq}^*(x, y) \exp[i(\alpha_{pq}x + \beta_{pq}y)] \quad (4)$$

Amplitude and phase 2D patterns of 8-order DOE transmission function are shown in fig. 4.

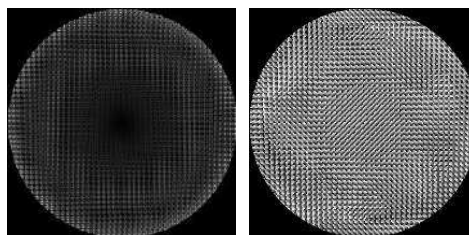


Fig. 4. Amplitude and phase 2D patterns of 8-order DOE transmission function.

There is an illustration of 8-order diffractive optical filter working in figure 5. Plane wave (wavefront is flat  $w(x, y) = 1$ ) illuminates the element and diffracts on it. As can be seen, all orders are holed. It means that analyzed wavefront does not have any aberrations.

In this work we propose a physical model of optical device (figure 6) for wavefront aberration analysis. This device is based on multi-order Zernike-matched diffractive optical element. Analyzed wavefront  $w(x, y)$  passes through the element. The lens  $O$  with focal distance  $f$  makes Fourier transform. Sensor matrix should be posed at the focal distance from the lens. Sensor plane has  $u$  and  $v$  coordinate axes.



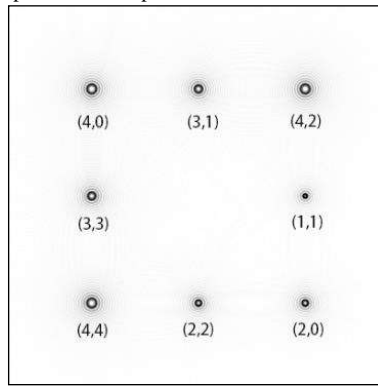


Fig.5. Plane wave filtering with 8-order diffractive optical filter.

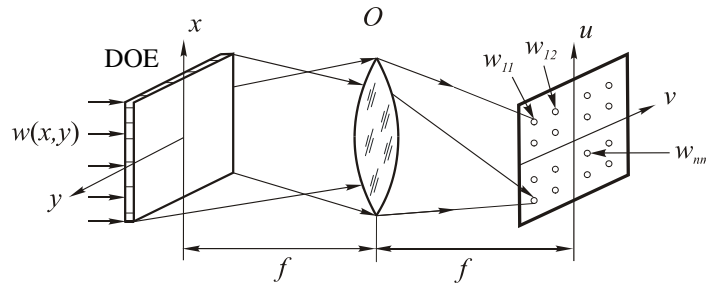


Fig.6. Optical scheme of proposed aberration analyzer.

Another testing simulation of proposed Zernike analyzer model is provided for initial beam wavefront  $w(x,y)=\psi_{2,0}(x,y)+\psi_{2,2}(x,y)$ . Resulting pattern is shown in fig. 7. Simulation shows that 8-order filter detects defocusing and astigmatism aberrations as it is set for initial beam. Detected coefficient meanings are  $C_{20}=0.995$ ,  $C_{22}=0.996$  but in initial beam they are equal to 1.

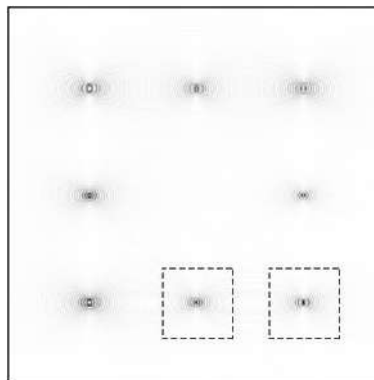


Fig. 7. Resulting amplitude patterns which is formed with 8-order analyzer after filtering the initial wavefront  $w(x,y) = \psi_{2,0}(x,y)+\psi_{2,2}(x,y)$ .

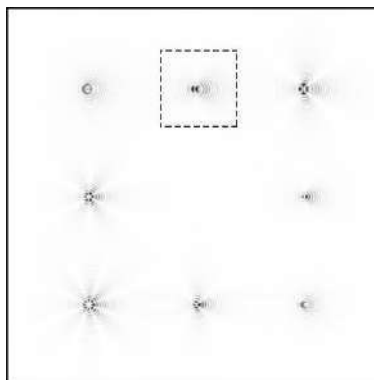


Fig. 8. Resulting amplitude patterns which is formed with 8-order analyzer after filtering the initial wavefront  $w(x,y)=\exp[i\alpha c_{31}\psi_{3,1}(x,y)]$ .

Once more example of working of proposed model is considered if the incident wavefront is  $w(x,y) = c_{31}\psi_{3,1}(x,y)$ ,  $c_{31}=2\sqrt{2} \approx 2.8284$ . The analyzer detects com-a aberration with coefficient  $c_{31}^{\text{detected}} = 2.7489$ . Resulting 2D amplitude patterns are shown in fig. 8.

It should be noticed that aberrations detection can be successful only for weak aberrations of the wavefront. With increasing of aberrations linear approximation of the wavefront as a sum Zernike polynomials is becoming nonapplicable and false detection happens.

In the table 2 a dependence of detection quality is shown. Incident wavefront is constant and equal  $w(x,y)=\exp[i\alpha c_{31}\psi_{3,1}(x,y)]$ ,  $c_{31}=2\sqrt{2} \approx 2.8284$ . Parameter  $\alpha$  is varied from 0 to 2.0; we observe a varying of detected value  $c_{31}^{\text{detected}}$ . From the table 2 it is seen that for too small and too high meanings of parameter  $\alpha$  the detection is mistaken.

Table 2. Comparison of detected and initial aberration if initial wavefront has aberration  $c_{31}=2.8284$  and parameter  $\alpha$  varies from 0 to 2.0.

$\alpha$	$c_{31}$	$c_{31}^{\text{detected}}$
0	2,8284	0,0104
0,25	2,8284	1,6585
0,5	2,8284	2,2136
0,75	2,8284	2,5905
1,0	2,8284	2,7489
1,25	2,8284	2,6919
1,5	2,8284	2,4603
1,75	2,8284	2,1192
2,0	2,8284	1,7398

### 3. Conclusions

Every conventional aberration can be expressed in terms of decomposition into Zernike basis. Evidently, to compensate the most prominent aberration it is enough to modify the incident field by adding a phase which is complex-conjugated to revealed aberration. It can be done with diffractive optics methods including etching relief to the lenses surfaces or adding DOE to the optical system.

It is worth to notice that the size of proposed DOE from Zernike analyzer is equal about 5 mm. 8-channel filter sensor has 512x512 pixels. Therefore it is easy to calculate resolution power of the device, it means 12.4  $\mu\text{m}$ .

However, resolution power of a diffractive optical element nowadays is limited by the 1  $\mu\text{m}$  meaning. Resolution of proposed 8-order DOE is far from technical limit. Thus, it says that the diffractive optical element can be easily produced with modern etching machines.

In this work provided simulation has shown that proposed Zernike analyzer can successfully detect aberration types and the model also reveals limitation of incident field aberration to be detected with the device. Defined tolerance range of aberration coefficient  $\alpha$  is  $\{0.5;1.75\}$ . Optimal meaning is 1. If  $\alpha$  goes out of tolerance range, aberration can not be detected or false detection can happens.

### 4. Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation and Russian Foundation for Basic Research grant No. 15-29-03823ofi\_m.

### References

- [1] Goncharsky AV, Popov VV, Stepanov VV. Introduction to computer optics. Moscow: MSU Publishing House, 1991.
- [2] Lane RG, Tallon M. Wave-front reconstruction using a Shack-Hartmann sensor. Appl. Opt. 1992; 31(32): 6902–6908.
- [3] Soifer VA, Golub MA. Laser beam mode selection by computer generated holograms. CRC Press, Boca Raton, 1994.
- [4] Kotlyar VV, Khonina SN, Soifer VA. Light field decomposition in angular harmonics by means of diffractive optics. Journal of Modern Optics 1998; 45(7): 145–150.
- [5] Khonina, SN, Almazov AA. Design of multi-channel phase spatial filter for selection of Gauss-Laguerre laser modes, Proceedings of SPIE 2002; 4705: 30–39.
- [6] Wolf E, Born M. Principles of Optics. Moscow: Science, 1973.
- [7] Ha Y, Zhao D, Wang Y, Kotlyar VV, Khonina SN, Soifer VA. Diffractive optical element for Zernike decomposition. Proceedings of SPIE 1998; 3557: 191–197.
- [8] Khonina SN, Kotlyar VV, Soifer VA, Wang Y, Zhao D. Decomposition of a coherent light field using a phase Zernike filter. Proc. SPIE 1998; 3573: 550–553.
- [9] Khonina SN, Kotlyar VV, Wang Y. Diffractive optical element matched with Zernike basis. Pattern Recognition and Image Analysis 2001; 11(2): 442–445.
- [10] Kotlyar VV, Khonina SN, Soifer VA, Wang Y, Zhao D. Coherent field phase retrieval using a phase Zernike filter. Computer Optics 1997; 17: 43–48.
- [11] Khonina SN, Kotlyar VV, Kirsh DV. Zernike phase spatial filter for measuring the aberrations of the optical structures of the eye. Journal of Biomedical Photonics Engineering 2013; 1(2): 146–153.
- [12] Zemax® User's Guide. Zemax Development Corporation, 2005.
- [13] Khorin PA, Khonina SN, Karsakov AV, Branchevskiy SL. Analysis of corneal aberration of the human eye. Computer Optics 2016; 40(6): 810–817. DOI: 10.18287/0134-2452-2016-40-6-810-817.
- [14] Kirilenko MS, Khorin PA, Porfirev AP. Wavefront analysis based on Zernike polynomials. CEUR Workshop Proceedings 2016; 1638: 66–75. DOI: 10.18287/1613-0073-2016-1638-66-75.

- [15] Lombardo M, Lombardo G. Wave aberration of human eyes and new descriptors of image optical quality and visual performance. *Journal of Cataract and Refractive Surgery* 2010; 36: 313–331.
- [16] Westheimer G, Liang J. Influence of ocular light scatter on the eye's optical performance. *Journal of the Optical Society of America A*. 1995; 12: 1417–1424.
- [17] Tokovinin A, Heathcote S. DONUT: measuring optical aberrations from a single extrafocal image. *Publications of the Astronomical Society of the Pacific* 2006; 118(846): 1165–1175.
- [18] Booth MJ. Wavefront sensorless adaptive optics for large aberrations. *Optics Letters* 2007; 32(1): 5–7.
- [19] Klebanov IM, Karsakov AV, Khonina SN, Davydov AN, Polyakov KA. Wave front aberration compensation of space telescopes with telescope temperature field adjustment. *Computer Optics* 2017; 41(1): 30-36. DOI: 10.18287/0134-2452-2017-41-1-30-36.
- [20] Porfirev AP, Khonina SN. Experimental investigation of multi-order diffractive optical elements matched with two types of Zernike functions. *Proc. Optical Technologies for Telecommunications*. SPIE 2016; 9807: 98070E-9 pp.
- [21] Degtyarev SA, Porfirev AP, Khonina SN. Zernike basis-matched multi-order diffractive optical elements for wavefront weak aberrations analysis. *Laser Physics and Photonics XVII; and Computational Biophysics and Analysis of Biomedical Data III*. *Proc. SPIE* 2017; 10337: 103370Q. DOI:10.1117/12.2269218.

# Formation of probing radiation for investigating a uniaxial x-cut crystal with the help of an aperiodic diffractive axicon

V.D. Paragin<sup>1</sup>, S.V. Karpeev<sup>1,2</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

<sup>2</sup>*Image Processing Systems Institute – Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, 151 Molodogvardeyskaya st., Samara 443001, Russia*

---

## Abstract

The paper presents an experimental study of transformation of a laser beam formed by an aperiodic diffractive axicon (fracxicon) in a lithium niobate x-cut. The beam is shown to undergo astigmatic rhomboidal transformation induced by the crystal birefringence. The output beam intensity distribution is measured at various distances from the crystal. The effects analyzed make it possible to extend the range of measuring thickness or birefringence of solid, liquid or gaseous media with uniaxial optical anisotropy.

*Keywords:* fracxicon; uniaxial crystal; birefringence

---

## 1. Introduction

Bessel laser beams [1-4] possessing non-diffractive properties are an efficient tool in various metrological [5,6], diagnostic [7,8] and testing [9-13] applications. Beams of this kind are also useful for investigating optical anisotropy and birefringence. In [14-22] it is shown that Bessel beam propagation in birefringent crystals of various cuts is accompanied with the transformation of the beam order or kind. In [23-25] the influence of the position and parameters of certain elements of optical scheme (laser wavelength, illuminating beam wavefront curvature, crystal temperature) on the Bessel beam characteristics at the crystal output is analyzed. In [26, 27] the thicknesses of z- and x-cuts of uniaxial crystals are measured by an optical method using the effects mentioned. Similar results can be expected for media with various refractive index distributions: linear, parabolic etc. This makes Bessel beams promising means for remote control of anisotropic films, metamaterials, birefringent crystals and ceramics.

As a rule, conical and diffractive axicons are used to form Bessel beams. The diameter of an axicon (laser beam) amounts to 200-300 mm, while its numerical aperture assumes specified values within a wide range in the long-wave part of the visible spectrum and shortwave infrared. Therefore, Bessel beams can be used for studying both submicron films and air routes many kilometers long.

There are many other axisymmetrical optical elements that form beams with non-diffractive properties, among them a logarithmic axicon [28-30], a generalized axicon [31], an axilens [32], and an aperiodic (fractional) axicon [33]. Linear diffractive axicons [28, 29] are used to produce Bessel laser modes, whereas the analogue of a logarithmic axicon is used to form hypergeometric modes of laser radiation [34, 35] that retain their mode properties longer than Bessel beams. The tandem of a lens and an axicon – a lensacon that makes it possible to form conical axial distributions - also possesses interesting properties. The aperiodic (fractional) axicon also referred to as a fracxicon [33] presented in the paper comprises an axicon and a parabolic lens as special cases.

The theoretical models developed and the experimental results obtained do not limit the use of non-diffractive beams to solids only. It is also possible to measure distributions of optical parameters of liquid and gaseous anisotropic media on their basis. For example, we can analyze the state of disturbed atmosphere, the properties of gas-plasma flows, distribution of ionic solution concentrations. The advantages of special beams including singular and vector ones for the purpose of atmospheric data transmission are described in the review [36]. We should also mention the possibility of producing tunable diffractive elements based on the electro-optic effect for the purpose of fast data transmission and three-dimensional addressing [37, 38].

A relatively small range of measuring due to periodic transformation of the beam order in a crystal [20, 27] is one of the problems of measuring thicknesses of z-cut birefringent crystals. Astigmatic beam transformation in x- and z-cut crystals leads to the splitting of the beam into separate intensity maxima [18, 19]. The angular dimension of the maxima decreases with the increase of the crystal thickness and birefringence [18, 19]. In the case of crystals of considerable thickness it leads to the problem of insufficient spatial resolution of the video camera.

Both problems are solved by using a multizone axicon or a variable-period axicon (fracxicon, lensacon etc.). An element of this kind makes it possible to choose a site of diffractive microrelief the spatial period of which conforms to the optical and dimensional parameters of the tested sample.

The aim of the study was to analyze the transformation of a laser beam formed by an aperiodic diffractive axicon (fracxicon) in a uniaxial x-cut crystal.

## 2. Experimental study

An optical setup was assembled to investigate astigmatic beam transformation. The scheme of the setup is presented in fig. 1. The setup includes a helium-neon laser, a spatial filter – beam expander, a polarizer, a fracxicon, a lithium niobate x-cut crystal, a CCD matrix. The spatial filter consists of a microlens 20x, a pinhole aperture with the diameter of 15  $\mu\text{m}$ , a collimator lens

with the focal distance of 200 mm. The setup allowed the formation and investigation of a sufficiently extended fracxicon beam observed at a distance up to 600 mm.

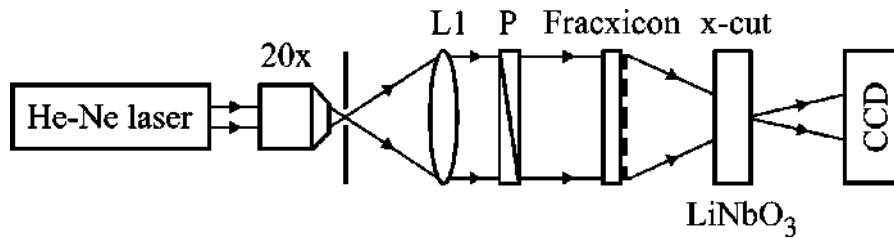


Fig. 1. Scheme of the experimental setup.

A polished uniaxial x-cut crystal of lithium niobate with  $842 \pm 2 \mu\text{m}$  thickness was used as the beam converter. The optical axis of the crystal was aligned parallel to one of its sides and its direction was marked. The axis of the polarizer and that of the crystal were parallel in the experiments. A phase diffractive variable-period axicon (fracxicon) shown in fig. 2 was used to form the beam. The fracxicon was made on a fused-silica substrate by plasma-chemical etching. The diameter of the fracxicon was 20 mm, the period of the diffraction microrelief -  $7 \mu\text{m}$  in the central part of the optical element and  $70 \mu\text{m}$  at the edge of the element.

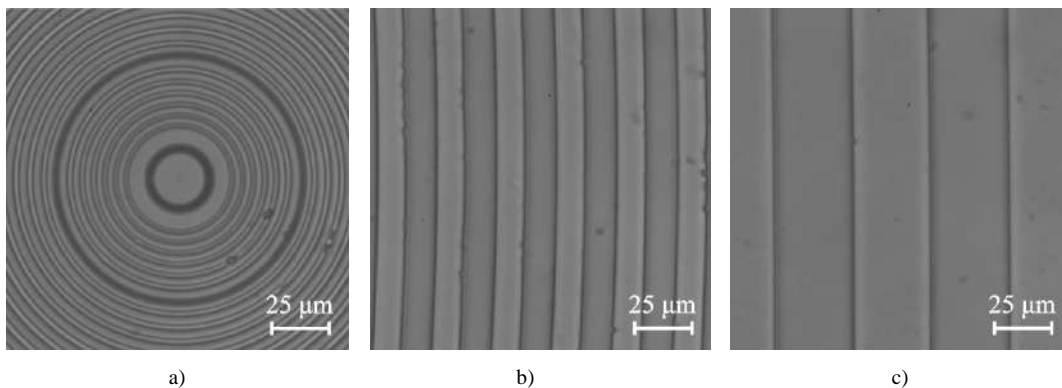


Fig. 2. Photos of the diffractive fracxicon microrelief: a) central part, b) middle part, c) edge part.

The distance between the crystal and the fracxicon was 75 mm. The image of the output beam was formed directly by the CCD matrix without the use of imaging optics. The images of the beams observed for various distances  $L$  between the CCD array and the crystal are presented in fig. 3.

The distance between the crystal and the fracxicon was 75 mm. The image of the output beam was formed directly by the CCD matrix without the use of imaging optics. The images of the beams observed for various distances  $L$  between the CCD array and the crystal are presented in fig. 3.

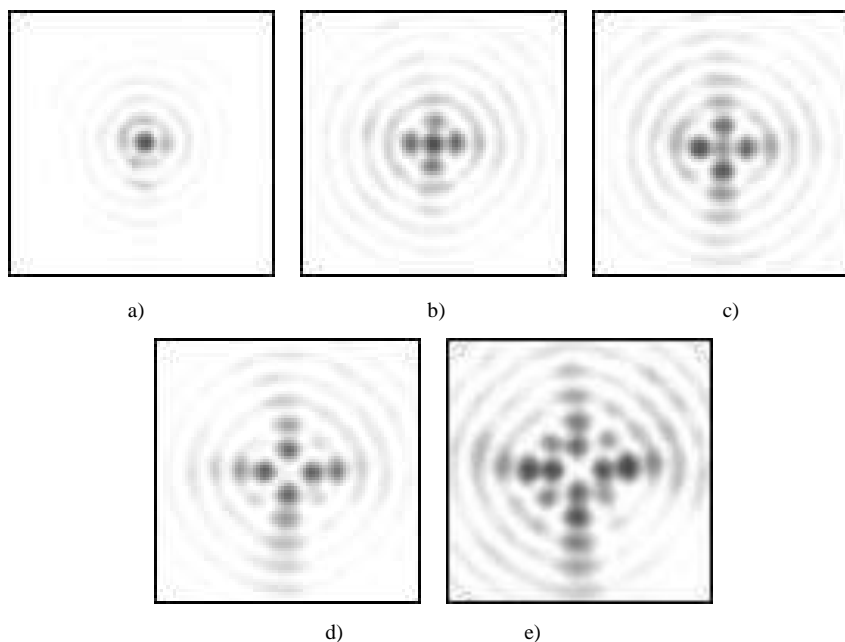


Fig. 3. Output beams for various distances  $L$  "crystal - CCD array": a)  $L=250 \text{ mm}$ , b)  $L=300 \text{ mm}$ , c)  $L=350 \text{ mm}$ , d)  $L=400 \text{ mm}$ , e)  $L=450 \text{ mm}$ .

Increasing the distance between the crystal and the camera makes astigmatic beam transformation more complicated due to the inclusion of fracxicon regions with increased angular aperture. This makes the use of several variable-period diffractive axicons unnecessary. One variable-period diffractive element is sufficient for reliable measurement of the thickness or birefringence of a plane-parallel crystal.

The results obtained in this work are in good agreement with earlier studies [18, 19]. The approach proposed has the advantage of a simpler optical measurement scheme that does not comprise an analyzer. The simplification is made possible due to parallel orientation of the polarizer and the crystal optical axis.

### 3. Conclusion

The results obtained confirm the validity of using an aperiodic diffractive axicon with specified radial period distribution. It is possible to select the part of the microrelief conforming to the test specimen parameters and the characteristics of the optical measurement system on the basis of this element. Measurement with the use of several parts of fracxicon improves the accuracy of determining the thickness or birefringence of the crystal.

### Acknowledgements

The work was supported by the Ministry of Education and Science of the Russian Federation and the Russian Foundation for Basic Research (RFBR grants 16-07-00825, 16-29-11698-ofi\_m).

### References

- [1] Bereznyi AE, Prokhorov AM, Sisakyan IN, Soifer VA. Bessel optics. DAN SSS 1984; 274(4): 802–805.
- [2] Bereznyi AE, Sisakyan IN. Binary elements of Bessel-optics. Computer Optics 1987; (1): 132–133.
- [3] Dumin, J. Exact solutions for nondiffracting beams. I. The scalar theory. J. Opt. Soc. Am. A 1987; 4(4): 651–654.
- [4] Durmin J, Miceli JJ, Eberly JH. Diffraction-free beams. Physical Review Letters 1987; 58: 1499–1501.
- [5] Kotlyar VV, Skidanov RV, Khonina SN. Contactless precision measurement of a linear displacement using DOEs forming Bessel beams. Computer Optics 2001; 21: 102–104.
- [6] Wang K, Zeng L, Yin Ch. Influence of the incident wave-front on intensity distribution of the nondiffracting beam used in large-scale measurement. Opt. Commun 2003; 216: 99–103.
- [7] Lee KS, Rolland JP. Bessel beam spectral-domain high-resolution optical coherence tomography with micro-optic axicon providing extended focusing range. Opt. Lett. 2008; 33(15): 1696–1698.
- [8] Cizmar T, Kollarov V, Tsampoula X, Gunn-Moore F, Sibbett W, Bouchal Z, Dholakia K. Generation of multiple Bessel beams for a biophotonics workstation. Optics Express 2008; 16(18): 14024–14035.
- [9] Arlt J, Dholakia K, Soneson J, Wright EM. Optical dipole traps and atomic waveguides based on Bessel light beams. Physical Review A 2001; 63: 063602-1–063602-8.
- [10] Garces-Chavez V, McGloin D, Melville H, Sibbett W, Dholakia K. Simultaneous micromanipulation in multiple planes using a self-reconstructing light beam. Nature 2002; 419: 145–147.
- [11] Khonina SN, Skidanov RV, Kotlyar VV, Soifer VA. Rotating microobjects using a DOE-generated laser Bessel beam. Proceedings of SPIE 2004; 5456: 244–255.
- [12] Khonina SN, Kotlyar VV, Skidanov RV, Soifer VA, Jefimovs K, Simonen J, Turunen J. Rotation of microparticles with Bessel beams generated by diffractive elements. Journal of Modern Optics 2004; 51(14): 2167–2184.
- [13] Fortin M, Piché M, Borra EF. Optical tests with Bessel beam interferometry. Optics Express 2004; 12(24): 5887–5895.
- [14] Khonina SN, Volotovskiy SG, Kharitonov SI. Features of nonparaxial propagation of Gaussian and Bessel beams along the axis of the crystal. Computer Optics 2013; 37(3): 297–306.
- [15] Khonina SN, Morozov AA, Karpeev SV. Effective transformation of a zero-order Bessel beam into a second-order vortex beam using a uniaxial crystal. Laser Phys 2014; 24: 056101-1–056101-5.
- [16] Khonina SN, Kharitonov SI. Comparative investigation of nonparaxial mode propagation along the axis of uniaxial crystal. Journal of Modern Optics 2015; 62(2): 125–134.
- [17] Khonina SN, Karpeev SV, Morozov AA, Pararin VD. Implementation of ordinary and extraordinary beams interference by application of diffractive optical elements. Journal of Modern Optics 2016; 63(13): 1239–1247.
- [18] Khonina SN, Pararin VD, Ustinov AV, Krasnov AP. Astigmatic transformation of Bessel beams in a uniaxial crystal. Optica Applicata 2016; 46(1): 5–18.
- [19] Khonina S, Pararin V, Degtyarev S, Savelyev D. Transformation of Bessel beams passing through uniaxial y-cut crystal. Materials of 17th International Conference on Transparent Optical Networks ICTON 2015; 1–4.
- [20] Khonina SN, Pararin VD, Karpeev SV, Morozov AA. Study of polarization transformations and interaction of ordinary and extraordinary beams in nonparaxial regime. Computer Optics 2014; 38(4): 598–605.
- [21] Pararin VD, Karpeev SV, Krasnov AP. A converter of circularly polarized laser beams into cylindrical vector beams based on anisotropic crystals. Computer Optics 2015; 39(5): 644–53. DOI: 10.18287/0134-2452-2015-39-5-644-653.
- [22] Khonina SN, Pararin VD. Electro-optical correction of Bessel beam conversion along axis of a barium niobate-strontium crystal. Computer Optics 2016; 40(4): 475–481. DOI: 10.18287/2412-6179-2016-40-4-475-481.
- [23] Pararin VD, Karpeev SV, Khonina SN. Control of the formation of vortex Bessel beams in uniaxial crystals by varying the beam divergence. Quantum Electronics 2016; 46(2): 163–168.
- [24] Pararin VD, Khonina SN, Karpeev SV. Control of the Optical Properties of a CaCO<sub>3</sub> Crystal in Problems of Generating Bessel Vortex Beams by Heating. Optoelectronics, Instrumentation and Data Processing 2016; 52(2): 174–179.
- [25] Pararin VD, Karpeev SV, Khonina SN. Transformation of Bessel beams in c-cuts of uniaxial crystals by varying the emission source wavelength. Journal of Russian Laser Research 2016; 37(3): 250–253.
- [26] Gornostay AV, Odnokov SB. A method to design a diffractive laser beam splitter with color separation based on bichromated gelatine. Computer Optics 2016; 40(1): 45–50. DOI: 10.18287/2412-6179-2016-40-1-45-50.

- [27] Pararin VD. Measuring the thickness of z-cut uniaxial crystals based on Bessel laser beams. *Computer Optics* 2016; 40(4): 594–599. DOI: 10.18287/2412-6179-2016-40-4-594-599.
- [28] Jaroszewicz Z, Sochacki J, Kołodziejczyk A, Staronski LR. Apodized annular-aperture logarithmic axicon: smoothness and uniformity of the intensity distribution. *Optics Letters* 1993; 18: 1893–1895.
- [29] Golub I, Chebbi B, Shaw D, Nowacki D. Characterization of a refractive logarithmic axicon. *Optics Letters* 2010; 35: 2828–2830.
- [30] Khonina SN, Balalaev SA. The comparative analysis of the intensity distributions formed by diffractive axicon and diffractive logarithmic axicon. *Computer Optics* 2009; 33(2): 162–174.
- [31] Sochacki J, Kołodziejczyk A, Jaroszewicz Z, Bará S. Nonparaxial design of generalized axicons. *Applied Optics* 1992; 31: 5326–5330.
- [32] Davidson N, Friesem AA, Hasman E. Holographic axilens: high resolution and long focal depth. *Optics Letters* 1991; 16(7): 523–525.
- [33] Khonina SN, Volotovskiy SG. Fracxicon – diffractive optical element with conical focal domain. *Computer Optics* 2009; 33(4): 401–411.
- [34] Kotlyar VV, Skidanov RV, Khonina SN, Soifer VA. Hypergeometric modes. *Optics Letters* 2007; 32(7): 742–744.
- [35] Khonina SN, Balalayev SA, Skidanov RV, Kotlyar VV, Paivanranta B, Turunen J. Encoded binary diffractive element to form hyper-geometric laser beams. *Journal of Optics A: Pure and Applied Optics* 2009; 11: 065702-1-065702-7.
- [36] Soifer VA, Korotkova O, Khonina SN, Shchepakina EA. Vortex beams in turbulent media: review. *Computer Optics* 2016; 40(5): 605–624. DOI: 10.18287/2412-6179-2016-40-5-605-624.
- [37] Pararin VD. Methods to control parameters of a diffraction grating on the surface of lithium niobate electro-optical crystal. *Technical Physics* 2014; 59(11): 1723–1727.
- [38] Pararin VD, Karpeev SV, Tukmakov KN, Volodkin BO. Tunable diffraction grating with transparent indium-tin oxide electrodes on a lithium niobate X-cut crystal. *Computer Optics* 2016; 40(5): 685–688. DOI: 10.18287/2412-6179-2016-40-5-685-688.

# The elaboration of numerical simulation error light pulse propagation in a waveguide of circular cross-section

A.A. Degtuarev<sup>1</sup>, A.V. Kukleva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We considered the problem of estimating the error in the solution of the wave equation recorded using infinite series Fourier-Bessel. The algorithm that adjusts the number of elements in a partial sum of infinite series, based on the assessment of the series balance. The application of the algorithm made it possible, without loss of accuracy, to substantially reduce the number of summable elements of the series in the numerical simulation of the light pulse propagation in a circular cross-section.

*Keywords:* wave equation; Fourier-Bessel series; evaluation of the residual series; numerical simulations; pulse of light; computational experiment; redundancy of partial sum components

---

## 1. Introduction

During the development of an application program for the numerical simulation of a physical process, it is important to investigate the actual error of the method used on special test cases. As test cases typically use such examples that can be resolved by an alternative method with high sufficiently precision, allowing to calculate the error of numerical method [1, 2].

This work is devoted to study the error of test value problem for the wave equation describing the propagation process of the light pulse in a waveguide in circular cross section. To elaboration the error estimate, we used remainder of the Fourier-Bessel. To check the quality of the balance assessment in the series we used the technique of computational experiment, which allows determine the degree of redundancy among several elements needed to sum to achieve the necessary precision [3].

In solving problems from numerical simulation propagation of a light pulse in a medium, various mathematical descriptions of the pulse [4-6]. In this paper, we considered two options describe different degrees of smoothness pulse function.

## 2. Mathematical model of light pulse propagation in a waveguide of circular cross-section

To describe the process of light pulse propagation we will consider the following boundary value problem:

$$\begin{cases} \frac{\partial^2 E}{\partial t^2} = \frac{c^2}{n^2} \left( \frac{\partial^2 E}{\partial r^2} + \frac{1}{r} \frac{\partial E}{\partial r} + \frac{\partial^2 E}{\partial z^2} \right), & r \in (0; R], z \in [0; L], t \in [0; T]; \\ E|_{t=0} = 0, & r \in (0; R], z \in [0; L]; \\ \left. \frac{\partial E}{\partial t} \right|_{t=0} = 0, & r \in (0; R], z \in [0; L]; \\ E|_{z=0} = \psi(r, t), & r \in (0; R], t \in [0; T]; \\ \left. \frac{\partial E}{\partial z} \right|_{z=L} = 0, & r \in (0; R], t \in [0; T]; \\ E|_{r=R} = 0, & z \in [0; L], t \in [0; T], \end{cases}$$

where  $E$  is a dielectric field intensity,  $c$  is a wave propagation speed in vacuum,  $n$  is a refractive index material of the waveguide,  $R$  and  $L$  is the radius and length of the waveguide,  $T$  is the duration of the dissemination process,  $\psi(r, t)$  is the function describing the pulse shape.

It is assumed when  $r = R$  an ideally conducting shell bound the waveguide, and the medium is not perturbed at the initial instant of time.

Here are the following two variants of kinetic moment:

$$\psi_1(r, t) = \varphi(r) \gamma(t) \sin \omega t, \quad \psi_2(r, t) = \varphi(r) \gamma(t) \sin \omega t \sin^2 \omega^* t,$$

$$\text{where } \gamma(t) = \begin{cases} 1, & t \in [0; t^*]; \\ 0, & t \in (t^*; T], \end{cases} \quad \omega = \frac{2\pi c}{\lambda}, \quad \omega^* = \frac{2\pi c}{\lambda j}, \quad t^* \text{ is the pulse duration at the entrance of the waveguide, } \lambda \text{ is the}$$

length of disturbing wave in vacuum,  $j$  a positive integer.  $\psi_1(r, t)$  a piecewise smooth function at variable  $t$ , because derivative has function jump in  $t = 0, t = t^*$ . Function  $\psi_2(r, t)$  has the smoothness of a second-order variable  $t$ .



### 3. Exact solution of boundary value problem

Application of the separation variables method [5] allows getting solution of boundary-value problem for the wave equation, it can be thought of as infinite series Fourier-Bessel. For example, when describing an impulse function  $\psi_1(r, t)$  and using  $\varphi(r) = J_0(\lambda_1 r)$  the solution would be:

$$E(r, z, t) = J_0(\lambda_1 r) \left[ \sum_{k=0}^{\infty} c_k \sin(v_k z) \frac{\omega \sin(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \omega_k \sin(\omega t)(\hat{\omega}^2 - \omega^2)}{\omega_k (\omega_k^2 - \omega^2)} + \sin(\omega t) \right], \text{ if } t \in [0; t^*];$$

$$E(r, z, t) = J_0(\lambda_1 r) \sum_{k=0}^{\infty} \sin(v_k z) \left( a_1(t^*) \cos(\omega_k(t - t^*)) + \frac{a_2(t^*)}{\omega_k} \sin(\omega_k(t - t^*)) \right), \text{ if } t \in (t^*; T].$$

When writing these formulas, we use the following notation:

$$\lambda_1 = \frac{\mu_1}{R}, \quad c_k = \frac{4}{\pi(2k+1)}, \quad v_k = \frac{\pi(2k+1)}{2L}, \quad \omega_k = \frac{c}{n} \sqrt{v_k^2 + \lambda_1^2}, \quad \hat{\omega} = \frac{c}{n} \lambda_1,$$

$$a_1(t) = c_k \left[ \frac{\omega \sin(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \omega_k \sin(\omega t)(\hat{\omega}^2 - \omega^2)}{\omega_k (\omega_k^2 - \omega^2)} + \sin(\omega t) \right],$$

$$a_2(t) = c_k \omega \left[ \frac{\cos(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \cos(\omega t)(\hat{\omega}^2 - \omega^2)}{(\omega_k^2 - \omega^2)} + \cos(\omega t) \right].$$

Graph of the cross section of a pulse by a plane  $r = 1 \mu\text{m}$  in the process of its propagation in the waveguide has shown in fig.1.

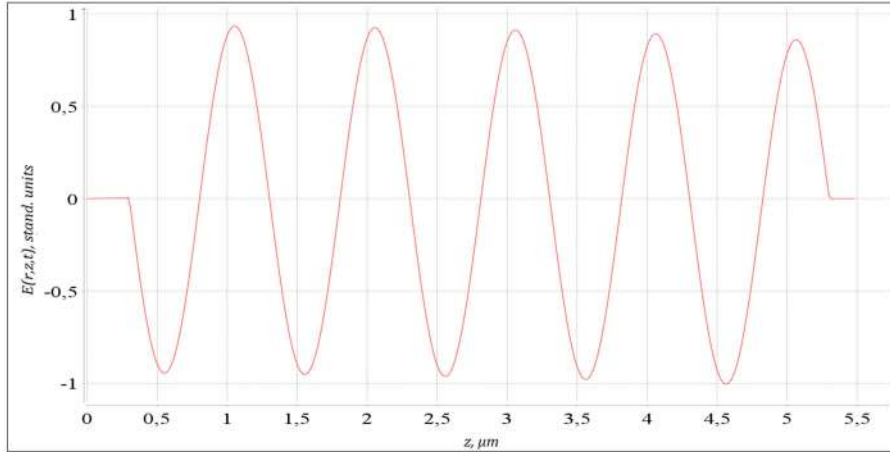


Fig. 1. Modeling the distribution piecewise smooth impulse in wave conductor, separation  $r = 1 \mu\text{m}$ .

For the case of smooth pulse described by function  $\psi_2(r, t)$  if  $\hat{\omega}^* = \frac{\omega}{10}$  and,  $\varphi(r) = J_0(\lambda_1 r)$  solution of boundary-value problem is as follows:

$$E(r, z, t) = J_0(\lambda_1 r) \left[ \sum_{k=0}^{\infty} \frac{c_k}{\omega_k} \sin(v_k z) (0.5a_3(t) + a_4(t) + a_5(t)) + \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right) \right], \text{ if } t \in [0; t^*];$$

$$E(r, z, t) = J_0(\lambda_1 r) \sum_{k=0}^{\infty} \sin(v_k z) \left( a_6(t^*) \cos(\omega_k(t - t^*)) + \frac{a_7(t^*)}{\omega_k} \sin(\omega_k(t - t^*)) \right), \text{ if } t \in (t^*; T].$$

In the last formulas, we used the following notations:

$$a_3(t) = \frac{\hat{\omega}^2 - \omega^2}{\omega^2 - \omega_k^2} (\omega \sin \omega_k t - \omega_k \sin \omega t),$$

$$a_4(t) = 5 \frac{0.16\omega^2 - 0.25\hat{\omega}^2}{16\omega^2 - 25\omega_k^2} \left( 4\omega \sin \omega_k t - 5\omega_k \sin \frac{4}{5} \omega t \right), \quad a_5(t) = 5 \frac{0.36\omega^2 - 0.25\hat{\omega}^2}{36\omega^2 - 25\omega_k^2} \left( 6\omega \sin \omega_k t - 5\omega_k \sin \frac{6}{5} \omega t \right),$$

$$a_6(t) = \frac{c_k}{\omega_k} (0.5a_3(t^*) + a_4(t^*) + a_5(t^*)) + c_k \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right),$$

$$a_7(t) = \frac{c_k}{\omega_k} \left( 0.5a_3'(t^*) + a_4'(t^*) + a_5'(t^*) \right) + c_k \left( \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right) \right),$$

$\mu_1$  is a root of an equation  $J_0(\mu R) = 0$ .

The process of propagating a piecewise-smooth pulse has shown in figure 2.

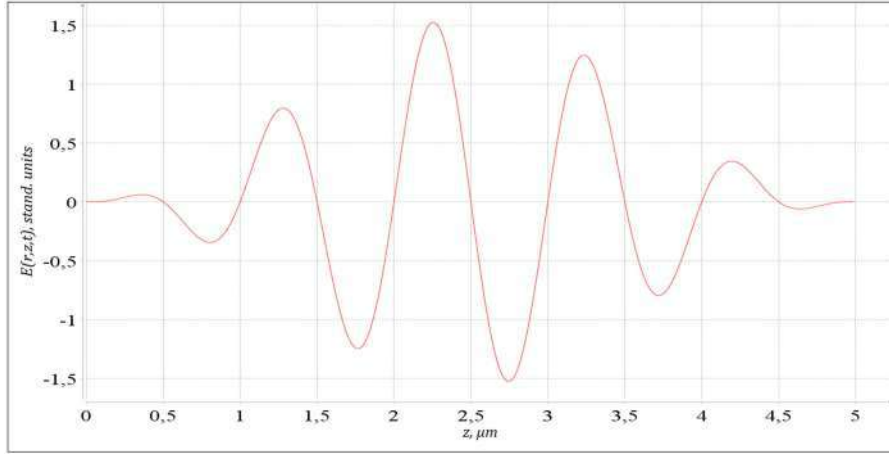


Fig.2. Modeling of smooth pulse in wave conductor, separation  $r = 1 \mu m$ .

#### 4. Series truncation error control

A computer program simulating the spread of pulse truncation of the infinite series implied above.

If we can get an estimate a balance number of  $E(r, z, t) = \sum_{k=1}^N u_k(r, z, t)$  in the form of

$$|R_N| = \left| \sum_{k=N+1}^{\infty} u_k(r, z, t) \right| \leq \Phi(N),$$

where  $\Phi(N)$  is the positive monotonically decreasing function if  $N \rightarrow +\infty$ , this assessment can be used to control the truncation error. To do this, we need only find  $N(\varepsilon)$ , is the least value  $N$ , satisfy the inequality  $\Phi(N) \leq \varepsilon$ , and for

approximate calculation of function values  $E(x, z, t)$  use a partial amount  $E_{N(\varepsilon)} = \sum_{n=1}^{N(\varepsilon)} e_n(x, z, t)$ .

In this case, the actual error of the calculated value of a function  $E$  at the selected point does not exceed the required level  $\varepsilon$ , that is

$$\varepsilon_{fact} = |E - E_{N(\varepsilon)}| = |R_{N(\varepsilon)}| \leq \Phi(N(\varepsilon)) \leq \varepsilon.$$

For the above two ways to specify the light pulse residues had been received by the relevant rows with the following  $t^*$  and  $w^*$ :

$$w^* = \frac{\pi c}{5\lambda}, \quad t^* = \frac{10\lambda}{c}.$$

In the case of piecewise smooth impulse, that described function  $\psi_1(r, t)$ , assessed takes the following form:

$$|E_N(r, z, t)| \leq \frac{8Ln}{\pi(2N+1)} \left( \frac{1.003}{\lambda} + \frac{2nL(\omega^2 - \hat{\omega}^2)}{\pi^2 c^2 \left( 3 + 2N - \frac{4nL}{\lambda} \right)} \right),$$

as for the case of smooth pulse, that described function  $\psi_2(r, t)$ , assessed takes the following form:

$$|E_N(r, z, t)| \leq \frac{0.16n^2 L^2 \omega^2}{c^2 \pi^3 (2N+1)^2}.$$

It should be noted that recorded higher truncation error estimates infinite series are uniform for all independent variables.

## 5. The method of refinement of the number of summable elements of a series using a computational experiment

Proposed evaluation are not ideal because they are using strict inequalities, and also they are uniform for all independent variables. That is why using of estimates results in adding more elements than is necessary to achieve the required accuracy. In this case, it is advisable to apply a technique, which reduces the degree of redundancy terms in the partial sum, and in so doing guarantees the achievement of required accuracy [3].

Let  $N$  positive integer, satisfies the inequality  $N \leq N(\varepsilon_1)$ , where  $\varepsilon_1 < \varepsilon$ , number  $N(\varepsilon_1)$  found by the rule described in paragraph 4. Then for partial amount  $E_N$  the actual error will satisfy the inequality:

$$\varepsilon_{fact}(N) = |E - E_N| \leq |E - E_{N(\varepsilon_1)}| + |E_{N(\varepsilon_1)} - E_N|.$$

Changing  $N$  within the boundaries  $N(\varepsilon_1) \geq N \geq 1$ , find lowest value  $N(\varepsilon_2)$ , when running the inequality  $|E_{N(\varepsilon_1)} - E_N| \leq \varepsilon_2$ , where  $\varepsilon_2 = \varepsilon - \varepsilon_1$ .

For this choice  $\varepsilon_2$  and equity of the previous inequality, the actual error  $\varepsilon_{fact}(N(\varepsilon_2))$  do not exceed value  $\varepsilon$ .

Thus, to reduce the number of summands in the partial amount, we must:

- 1) Specify the number of  $\varepsilon_1 < \varepsilon$  and then find the value  $N(\varepsilon_1)$ , that the smallest value  $N$ , satisfy the inequality  $\Phi(N) \leq \varepsilon_1$ .
- 2) Changing a variable  $N$  from the value  $N(\varepsilon_1)$  downward, find the smallest of its value that satisfies the inequality  $|E_{N(\varepsilon_1)} - E_N| \leq \varepsilon_2$ . The resulting value is  $N(\varepsilon_2)$ .
- 3) Changing value with sample spacing  $\varepsilon_1$  and  $\varepsilon_2$  so, to  $\varepsilon_1 + \varepsilon_2 = \varepsilon$ , run the steps 1) and 2) again.
- 4) Of all the values  $N(\varepsilon_2)$ , obtained in step 3), select the smallest.

As a result of the use of this algorithm, it can be expected that the number of summable elements  $N(\varepsilon_2)$  in the partial sum will be reduced significantly as compared with the number of  $N(\varepsilon)$  while maintaining safeguards for accuracy, i.e.

$$\varepsilon_{fact}(N(\varepsilon_2)) \leq \varepsilon.$$

In tables 1 and 2 are the results of computational experiments, aimed at reducing the number of summands in partial amounts. The calculations have been carried out with the following parameters:

$$\lambda = 1 \mu m, n = 1, L = 7 \mu m, R = 5 \mu m, c = 3 \cdot 10^{14} \mu m / s, r = 1 \mu m, z = 1 \mu m, t = \frac{tc}{n} \mu m.$$

Asked value  $\varepsilon$  in increments of the maximum value of the amplitude of the wave.

Table 1. The dependence of the summands number  $N(\varepsilon)$  and  $N(\varepsilon_2)$  of coordinate  $t$  with different values  $\varepsilon$  for piecewise smooth impulse.

$\varepsilon$	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$
$N(\varepsilon)$	131	1019	9844	98079	980434
$t, \mu m$	$N(\varepsilon_2)$				
0.9	13	48	231	3116	9906
0.999	37	306	1241	6774	26632
0.99999	37	312	3072	35599	126836
1	37	312	3075	37713	377122
1.00001	37	312	3072	35599	126836
$\varepsilon$	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$
1.001	37	306	1241	6774	26633
1.1	16	68	320	3119	9906
1.7	19	34	124	1286	4086
2.5	15	32	96	928	984
4	13	25	66	612	643
5.1	16	30	75	649	1436
5.9	17	28	324	3116	3258
5.999	47	355	1262	6422	9906
5.99999	47	466	4672	35599	26632
6	47	467	4672	37713	126836
6.00001	47	465	4671	35599	377122

6.001	47	383	1461	6423	126836
6.1	17	30	360	3119	26634

Table 2. The dependence of the summands number  $N(\varepsilon)$  and  $N(\varepsilon_2)$  of coordinate  $t$  with different values  $\varepsilon$  for smooth pulse.

$\varepsilon$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$	$10^{-6}$
$N(\varepsilon)$	21	36	113	357	1128
$t, \mu m$	$N(\varepsilon_2)$				
0.9	15	18	28	62	132
0.999	15	18	28	61	136
0.99999	14	17	28	62	126
1	15	18	27	67	141
1.000001	10	21	37	91	186
1.001	10	22	42	101	211
1.1	15	17	33	61	132
1.7	10	17	37	81	181
2.5	13	15	26	67	146
4	15	22	46	101	216

From the table it can be seen that the number of summands, using uniform assessments for the respective series truncation allows you to get only the rough partial sums of lengths. These values are repeatedly exceed the values obtained from the application of the above algorithm. As can be seen from table 1, to calculate the tension of the electric field in the foreground and background areas of wave fronts requires a much larger number of terms, for example, in the range  $1.7 \mu m \leq t \leq 5.1 \mu m$  order enough 4086 parts to achieve precision  $10^{-5}$ , while in the range  $0.9 \mu m \leq t \leq 1.1 \mu m$  we want 377122 parts. This increase in the number of summands is a consequence of the weak function breaks  $\psi_1(r, t)$ , significantly slowing down the convergence of series. For the case of smooth pulse, function description  $\psi_2(r, t)$  the uneven distribution of values  $N(\varepsilon_2)$  for different  $t$  turns out to be negligible.

## 6. Conclusion

Developed and implemented programmatically algorithm provides adjustment of the partial sums length of infinite series, obtained in the course of solving boundary value problem for the wave equation. For practical application of the algorithm, it is of fundamental importance to first obtain an upper estimate for the remainder of the Fourier series that determines the solution of the boundary value problem.

The application of developed algorithm for specific series that describe the distribution of momentum in circular waveguide section allowed multiple times (from 3 up to 1500 times and more for Piecewise-smooth momentum and from 2 to 5 times for the case of smooth pulse) to reduce length of the partial sums of the series.

## References

- [1] Feng, X. A high-order compact scheme for the one-dimensional Helmholtz equation with a discontinuous coefficient. *International Journal of Computer Mathematics* 2012; 1: 1–7.
- [2] Degtyarev AA, Kozlova ES. Investigation of accuracy of numerical solution of the one-way Helmholtz equation by method of computational experiment. *Computer Optics* 2012; 36(1): 36–45.
- [3] Degtyarev AA, Praslova MO. Estimation of the error of the solution of the wave equation in the problem of modeling the distribution of the light pulse in the planary waveguide. *Proc. of ITNT-2016, Samara, SSAU 2016*; 852–859. (in Russian)
- [4] Kotlyar VV, Kozlova ES. Simulation of ultrafast 2d light pulse. *Computer Optics* 2012; 36(2): 158–164.
- [5] Kotlyar VV, Kozlova ES. Simulations of Sommerfeld and Brillouin precursors in the medium with frequency dispersion using numerical method of solving wave equations. *Computer Optics* 2013; 37(2): 146–154.
- [6] Fuchs U, Zeitner U, Tunnermann A. Ultra-short pulse propagation in complex optical system. *Optics Express* 2005; 13(10): 3852–3861.
- [7] Tikhonov AN, Samarskiy AA. *Equations of mathematical physics*. M.: Nauka 1972. (in Russian)

# Spectra and field distribution of photonic-crystal structure with inclusions of metal nanoparticles

I.A. Glukhov<sup>1</sup>, S.G. Moiseev<sup>1,2</sup>

<sup>1</sup>*Ulyanovsk State University, 42 Lev Tolstoy Str., 432017, Ulyanovsk, Russia*

<sup>2</sup>*Kotelnikov Institute of Radio Engineering and Electronics of the Russian Academy of Sciences, Ulyanovsk Branch, 48/2 Goncharov Str., 432011 Ulyanovsk, Russia*

---

## Abstract

Transmittance and reflectance spectra as well as field distribution in 1D photonic-crystal structure with embedded dielectric layer and monolayer of metal nanoparticles are characterized. The influence of plasmonic monolayer location on defect modes of photonic-crystal structure is demonstrated with respect to domains of field confinement in the cavity area.

*Keywords:* nanoplasmonics; photonic-crystal structure; defect mode; field localization

---

## 1. Introduction

In recent years 1D photonic-crystal structures (PCS) created on the basis of different materials are of a special interest to researchers. Owing to periodic modulation of refractive index, photonic spectrum of these structures has a band gap, in which incident radiation is practically totally reflected. This property is critical for practical use as it enables to control optical radiation in data-transmission systems and in laser technology. Particularly remarkable is Fabry-Perot microresonator-like structure composed of two Bragg reflectors with defect layer there between. Defect layer in such-like structure plays the role of optical microcavity (microresonator) on which electromagnetic radiation can be localized. This can add to material-radiation interaction effects.

Varying geometrical and physical properties of the structure it is possible to control spectral characteristics of PCS [1, 2] that enables to improve considerably their functionality. For example, through breakdown of the structure periodicity or using materials with controlled properties (non-linear, resonant, magnetogirotopic) photonic spectrum of PCS can be modified considerably. Metallic-dielectric nanocomposite media are advanced materials to be used as microcavity of photonic-crystal resonator. In the field of plasmonic resonance vigorous dispersion of optical properties of these materials is observed [3, 4]. This paper describes the case of ultrathin resonance structure as a monolayer of metal nanoparticles, plasmonic frequency of which coincides with defect mode frequency of PCS.

## 2. PCS material parameters and transfer matrixes

In order to calculate reflectivity and transmission of plane-layered structure with embedded monolayer of nanoparticles we employ T-matrix technique. A special case is interface, optical qualities of which are determined by Fresnel reflection and transmission coefficients [5]. Since array of nanoparticles situated in the same plane interacts with electromagnetic wave like plane interface, it can be also treated as an interface with its own reflection and transmission coefficients.

We assume that there are  $N$  interfaces in the layered medium, and they are formed by  $N-1$  interfacial boundaries and a single layer of nanoparticles. A space between interfaces is packed by media with different refraction indexes  $n_i$  ( $i = 0 \dots N$ ). Semi-infinite media are those that have  $n_0$  and  $n_N$  refraction indexes. Let a harmonic wave is incident on a layered structure in  $z$ -direction. To describe its propagation in PCS we introduce the following notation for electric field components inside structure:  $E_i(z_i^-)$  to the left of  $i$  number interface;  $E_i(z_i^+)$  to the right of  $i$  number interface;  $E_f$  for the propagating forward wave;  $E_b$  for the propagating backward wave.

According to the introduced notation, complex amplitudes of counter-propagating waves on  $m$  interface in the layer with reflection index  $n_{m-1}$  are equal to  $E_f(z_m^-)$  and  $E_b(z_m^-)$ . At the same interface but in the layer with reflection index  $n_m$  they are equal to  $E_f(z_m^+)$  and  $E_b(z_m^+)$ . Relationship of these fields on  $m$ -interface (to the left and to the right of it) can be expressed as matrix equation:

$$\begin{pmatrix} E_f(z_m^-) \\ E_b(z_m^-) \end{pmatrix} = I_{m-1,m} \begin{pmatrix} E_f(z_m^+) \\ E_b(z_m^+) \end{pmatrix}, \quad (1)$$

$$I_{m-1,m} = \frac{1}{t_{m-1,m}} \begin{pmatrix} 1 & -r_{m,m-1} \\ r_{m-1,m} & t_{m-1,m} t_{m,m-1} - r_{m-1,m} r_{m,m-1} \end{pmatrix}, \quad (2)$$

where  $r_{i,j}$ ,  $t_{i,j}$  are complex reflection and transmission coefficients of the interface deviding media with refraction indexes  $n_i$  and  $n_j$  when the lightwave is incident from the medium with refraction index  $n_i$ . In case of plane interface  $r_{i,j}$  and  $t_{i,j}$  are Fresnel coefficients [5]. Relationship of the fields on two interfaces of  $m$  and  $m+1$  numbers confining homogeneous layer of  $m$  number is via transfer matrix  $\hat{F}_m$  :

$$\begin{pmatrix} E_f(z_m^-) \\ E_b(z_m^-) \end{pmatrix} = \hat{F}_m \begin{pmatrix} E_f(z_m^+) \\ E_b(z_m^+) \end{pmatrix}, \quad (3)$$

$$\hat{F}_m = \begin{pmatrix} \exp(-i\delta_m) & 0 \\ 0 & \exp(i\delta_m) \end{pmatrix}, \quad (4)$$

where  $\delta_m = kn_m L_m$  is phase thickness of the layer;  $k = \omega/c$  is the wave number.

Applying expressions (1) – (4) to the entire PCS we obtain relation for the amplitudes to the left of the first interface and to the right of the last (with number  $N$ ) interface:

$$\begin{pmatrix} E_f(z_1^-) \\ E_b(z_1^-) \end{pmatrix} = \hat{G} \begin{pmatrix} E_f(z_N^+) \\ E_b(z_N^+) \end{pmatrix}, \quad (5)$$

$$\hat{G} = I_{0,1} F_1 I_{1,2} F_2 \dots F_{N-1} I_{N-1,N}. \quad (6)$$

Note that in semi-infinite medium with refraction index  $n_N$  there exists only transmitted wave, therefore we shall assume  $E_b(z_N^+) = 0$  in (5).

Reflectance and transmittance of PCS are calculated from the formulas

$$T = \left| \frac{1}{\hat{G}_{11}} \right|^2, \quad R = \left| \frac{\hat{G}_{21}}{\hat{G}_{11}} \right|^2. \quad (7)$$

To calculate field distribution in PCS we can use expressions similar to (5), in the left part of which column elements are substituted with amplitudes of local fields in the corresponding points.

### 3. Reflection and transmission spectra of nanoparticle monolayer

For the analysis of the properties of PCS with embedded nanocomposite monolayer film of nanoparticles placed into dielectric matrix, it is necessary to know amplitude coefficients of monolayer reflection and refraction. Analytical calculation of these coefficients is not a trivial task; therefore we use numerical technique – FEM implemented in COMSOL Multiphysics software. For simplicity we consider the case of the ordered monolayer film in which nanoparticles are located at the sites of square lattice lying in the palne ( $xy$ ).

Taking into account structure symmetry, we took fourth of the structure's unit cell (Fig. 1). Such domain contains fourth of nanoparticle and is of the size equal to  $\frac{1}{2}$  period of the structure along the direction of  $x$  and  $y$  coordinate axes. In order to obtain the entire monolayer it is necessary to apply reflection operations to the domain shown in Fig. 1 (to complete the unit to the full with a spherical shape particle), and then to apply transmission operations along  $x$  and  $y$  coordinate axes. The structure thus obtained will be 2D array of nanoparticles in ( $xy$ ) plane. Boundary conditions are selected in such a way that the model fits the case of normal (to the plane of monolayer film) incidence of the light wave. Incident polarization is oriented parallel to one of the crystal axes of monolayer film.

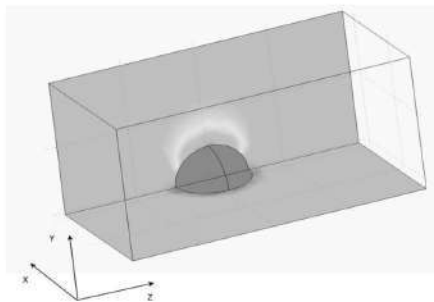


Fig. 1. Modeling domain in Comsol Multiphysics software.

As a medium in which nanoparticle monolayer is weighted, we use material with dielectric constant  $\varepsilon_m = 2.25$ . To calculate dielectric constant of metal nanoparticles we use relation of Drude theory [5]:

$$\varepsilon_p(\omega) = \varepsilon_0 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma}, \quad (8)$$

where  $\omega_p$  is plasmonic frequency,  $\varepsilon_0$  is lattice contribution,  $\gamma$  is relaxation parameter. For definiteness, as nanoparticle material we use silver, for which  $\omega_p = 1.36 \cdot 10^{16} \text{ c}^{-1}$ ,  $\varepsilon_0 = 5$ ,  $\gamma = 3 \cdot 10^{13} \text{ c}^{-1}$ .

Modeling outcomes for optical properties of silver nanoparticle monolayer are shown in Fig. 2. It is obvious that in the domain of surface plasmonic resonance of nanoparticles (resonant wavelength falls within 435 nm) monolayer film reflection and transmission spectra are subject to strong changes. Amplitudes of the observed values in resonance region depend on the surface concentration of nanoparticles: with decrease in the average distance among nanoparticles resonance becomes more pronounced; the width of resonance increases and frequency shift of resonance towards short wavelength region is observed. Thus, spectral characteristics of the monolayer of metal nanoparticles depend on internal geometrical parameters which make it possible to control to a certain extent its influence on spectrum of PCS.

#### 4. Analysis of the properties of photonic-crystal structure with monolayer of nanoparticles

Let us consider PCS, in which between two dielectric reflectors there is a defect layer consisting of dielectric matrix and a monolayer film of nanoparticles. Transfer matrix of such structure can be expressed as:

$$N = I_{0,1} F_1 \quad I_{d-1,d} F_{d1} F_s F_{d2} I_{d,d+1} \quad F_{N-1} I_{N-1,N} = M^a D M^b, \quad (9)$$

where  $F_s$  is transfer matrix of nanoparticle monolayer;  $F_{d1}$  and  $F_{d2}$  are transfer matrixes of the layers that edge monolayer of nanoparticles;  $M^a$  and  $M^b$  are transfer matrixes that describe dielectric reflectors containing  $a$  and  $b$  binary layers respectively. Binary layers of dielectric reflectors consist of two layers of isotropic dielectric material of real transmissivities  $\varepsilon_j$  and thickness  $L_j$  ( $j = 1, 2$ ). For modeling optical properties of the structure we assumed  $\varepsilon_1 = 6.25$  and  $\varepsilon_2 = 2.25$ . Thickness of layers of the structure meets the requirements  $L_{1,2} = \lambda_0 / 4 \sqrt{\varepsilon_{1,2}}$ ; thickness of defect layer is equal to  $L_d = \lambda_0 / \sqrt{\varepsilon_d}$ , where  $\lambda_0$  – wavelength in vacuum calculated for the central frequency of the photonic band gap. Presence of defect layer in PCS leads to the occurrence in the photonic band gap of narrow spectral transmission band with the peak value of transmission index which is close to 1.

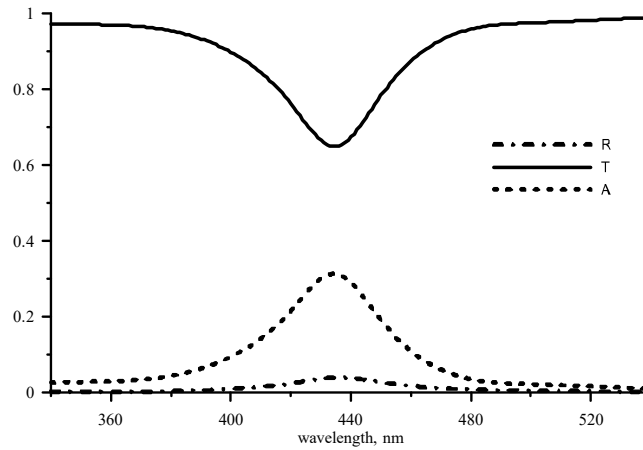


Fig. 2. Reflection, transmission and absorption spectra for monolayer of silver nanoparticles. Period of structure is 10 nm.

Figures 3 and 4 show field distribution (square amplitude of dielectric field intensity) and transmission and reflection spectra of PCS of  $M^5 D M^5$ -type depending on location of the monolayer of plasmonic nanoparticles. Location of the monolayer is given by relations  $L_{d1} = L_d / 2 + \Delta$ ,  $L_{d2} = L_d / 2 - \Delta$ , where  $L_d$  is thickness of the central layer, which is divided by a monolayer of nanoparticles into two domains of the thickness  $L_{d1}$  and  $L_{d2}$  respectively. The domain of the thickness  $L_{d1}$  is located from incidence of external electromagnetic wave. Thus,  $\Delta$  is a value of monolayer film displacement from the center of PCS. Parameters of dielectric layers of PCS correlate such that in the center of PCS, i. e. for  $\Delta = 0$ , field amplitude is close to zero (standing-wave node), and in displacement by  $\Delta = \pm 71 \text{ nm}$  the amplitude reaches a maximum value (antinode of standing wave). The given relations indicate that if monolayer film of nanoparticles is located in the center ( $\Delta = 0$ ), field distribution and transmission and reflection spectra are practically comparable with the case of monolayer-free PCS. Such peculiarity is explained practically by total absence of electrodynamic interaction between monolayer film and lightwave in standing-wave field ( $\Delta = 71 \text{ nm}$ ), when intensity

of electromagnetic scattering by monolayer is the highest, defect mode transformation is observed. Transformation is displayed as decrease in its amplitude and splitting of spectrum curves.

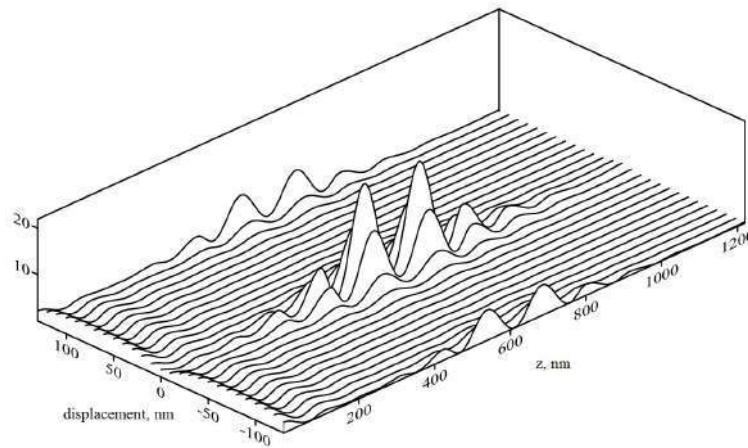


Fig. 3. The distribution of electromagnetic field amplitude (in relative units) throughout PCS as a function of the value of displacement of monolayer of nanoparticles from the center of the structure.

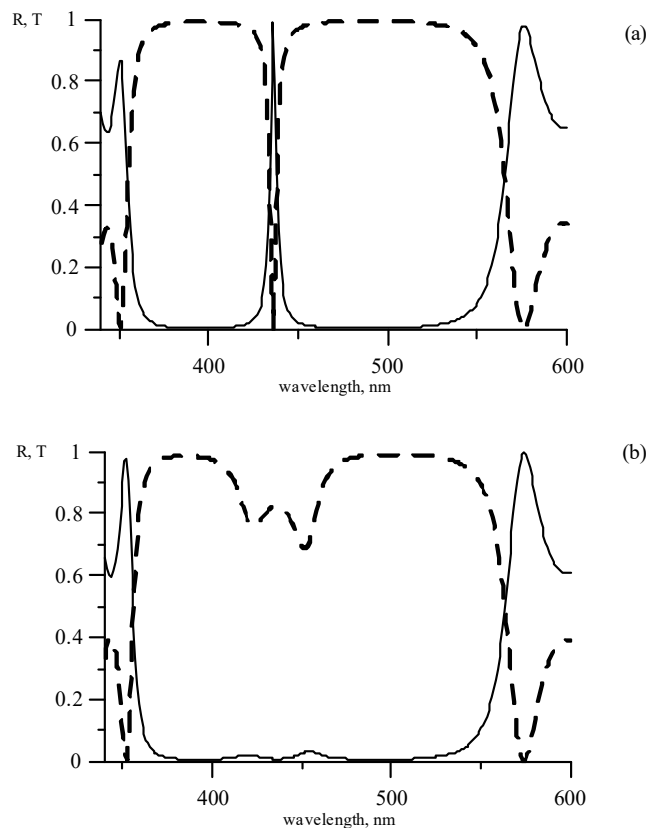


Fig. 4. Transmission (dashed line) and reflection (continuous line) spectra of PCS when monolayer displacement is (a)  $\Delta = 71$  nm, (b)  $\Delta = 0$ .

## 5. Conclusions

We have demonstrated that control of transmission and reflection coefficients corresponding to defect mode in photonic band gap of PCS is capable owing to the use of nanoparticle monolayer with plasmonic resonance. It is shown that defect mode amplitude is heavily dependent on the location of monolayer. Dependence of spectral characteristics of the layered structure on the location of plasmonic monolayer is attributed to the inhomogeneity of electromagnetic field distribution in the optical microcavity placed between distributed Bragg reflectors.

## Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation (State Contracts Nos. 3.5698.2017/P220 and 3.8388.2017/ITR) and the Russian Foundation for Basic Research (Projects Nos. 15-07-08111 and 17-02-01382).



## References

- [1] Vorobev LE, Ivchenko EL, Firsov DA, Shalygin VA. Optical properties of nanostructures. Saint Petersburg: “Nauka” Publisher 2001.
- [2] Gaponenko SV, Rosanov NN, Ivchenko EL. Optics of nanostructures. Saint Petersburg: “Nedra” 2005.
- [3] Moiseev SG, Ostatochnikov VA. Defect modes of one-dimensional photonic-crystal structure with a resonance nanocomposite layer. *Quantum Electron* 2016; 46(8): 743–748.
- [4] Vetrov SY, Avdeeva AY, Timofeev IV. Spectral properties of a one-dimensional photonic crystal with a resonant defect nanocomposite layer. *JETP* 2011; 113(5): 755–761.
- [5] Born M, Wolf E. Principles of Optics. Cambridge: Cambridge University Press 1999.

# Microexplosions polystyrene microparticles on substrate covered by aluminum

V.S. Vasilev<sup>1,2</sup>, R.V. Skidanov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Annotation

Experimental results of microexplosion polystyrene microparticles diameter equals 5 micron located on substrate covered by aluminum layer thickness in 100 nanometers were showed. Produced a comparison results of experiment on quartz substrate and on aluminum substrate. As a source of radiation was chosen a laser with wavelength equals 355 nanometers.

*Keywords:* microexplosions; ultraviolet beam, polystyrene microbits; velocity of microparticles; quartz substrate; aluminum substrate

---

## 1. Introduction

Currently, all manipulation works are now heading for reducing the size of the moving objects. There are plenty of microparticles manipulation methods using optical traps of different types. However, it is desirable to have a method to move a relatively large micro-objects. Especially such objects could be found in biological research (sporules, microslides of tissue, large cells). And it is necessary to prevent an effect even of a minimum quantity of light radiation on the biological micro-object. It is possible to carry out mechanical micro-manipulation by means of mechanical micro tweezers. This method is invasive for microparticles manipulating.

Typical sizes of the roaming micro-objects using common optical trap are from fractions of a micrometer to about ten micrometers. It is necessary to significantly increase the power of the light beam with the increase of the micro-objects size. Because anyway some fraction of the light beam energy is absorbed by the object, there is a certain limit for the microparticles size which can be moved by the forces of optical trapping. The precise value of this size depends on many parameters: the micro-objects absorption coefficient, fluid properties, micro-object surface shape, etc. A rough estimate of this size for transparent spherical micro-objects gives a value of about 30  $\mu\text{m}$ . It should be noted that the micro-object with a size close to the limit is experienced strong thermal impact. Things get worse with the movement of opaque micro-objects in the light traps. The limiting size is reduced by one and a half to two times. At the same time, the micro-objects with sizes up to 100 microns are still quite small for a mechanical movement. There is another complicate combined method of micro-objects grasping with the usage of light and ultrasound [5]. But here is another [6] described simpler method of moving micro-objects by means of the microexplosions of polystyrene microparticles in a beam of an ultraviolet laser with a wavelength of 355 nm.

In this paper we will consider a method of moving micro-objects with explosion of polystyrene microparticles on a substrate coated with a layer of aluminium. As compared with microexplosions on a quartz substrate, this method significantly reduces expended energy of microexplosions due to the interference of the beam, which is incident on the substrate and a beam reflected from the surface of the substrate. Object of study in this research is calculation average velocity polystyrene microparticles with diameter equals 5 micron after microexplosion occurring under the pressure of laser emission with wavelength 355 nanometers.

The subject of the study is the behavior of polystyrene micro-particles in the explosion.

The aim of this work is the experimental test of interference effect between incident and reflected beams in the explosion on a substrate covered with aluminum. Another aim of this work is the calculation of scattering speed of microparticles after the explosion and finding the parameters for explosion occurrence.

In accordance with the intended aims following tasks were summarized:

- in situ observation of polystyrene microparticles microexplosion on a substrate covered with aluminum under the action of ultraviolet laser with a wavelength of 355 nm;
- the calculation of the average dispersion velocity of polystyrene microparticles after the explosion;
- finding the system parameters for the observed microexplosions of polystyrene microspheres.

### Scientific and practical novelty and significance of the results:

- full-scale experiment with the microexplosion of polystyrene microparticles on a substrate covered with aluminum under the action of an ultraviolet laser with a wavelength of 355 nm was successfully completed.
- the average expansion rate of polystyrene microparticles after microexplosion was calculated.
- the system parameters for the observed polystyrene microspheres microexplosions were found.

## 2. Theoretical description observed effect

In this experiment, there is an effect of interference of incident and reflected waves. As a result of usage a single source of light radiation the incident and reflected waves are coherent. Thus, the total intensity of incident and reflected waves [7] can be represented as follows:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta \quad (1)$$

where  $I_1$ ,  $I_2$  - the intensity of the incident and reflected beams, respectively,  $\delta$  - the phase difference between these beams. Considering that quartz glass has a transmittance of 99% and aluminum reflects about 93-94%, it is possible to write the following:

$$I_1 > I_2 \quad (2)$$

Then, inserting into the formula above, we get that  $I_{\max} = 4I_1$  and  $I_{\min} = 0$ . This formula will be tested experimentally — if the microexplosions will occur on the aluminum substrate at the power level at which microexplosions of polystyrene microparticles on a quartz substrate were not observed, so it is possible to draw a conclusion about the strengthening of the two beams through their interference.

### 3. Experiment

Consider the installation, which was used in the observation process of the full-scale experiment with microexplosion of polystyrene microparticles on a substrate coated with aluminium in thickness of 100 nm. The application of the aluminium substrate was made with the use of plant «Carolina D 12 A» designed for magnetron sputtering on ceramic, silicon and other substrates with sizes up to 100 mm. The following keys were added to the optical diagram in figure 1: 1 – continuous UV laser DTL – 375 with wavelength 355 nm and maximal average power equals to 40 mV [4]; 2,3 – rotary mirrors; 4 – semitransparent cubic; 5 – focused microobjective (20x); 6 – substrate covered by aluminum with polysterene microparticles; 7 – CCD – camera FastVideo 500 E with resolution 640x480.

Now let's move to the description of the experiment. The light by means of rotary mirrors and microscope objective is focused into the required area of the substrate with microparticles. In view of the high reflection coefficient of aluminium (about 93-94%) the process of microexplosions on the aluminium substrate needed to be monitored in the reflected light. For separation of the incident and the reflected beams it is used a cube with translucent mirrors. Reflected light falls on the camera and the resulting image is processed on the computer.

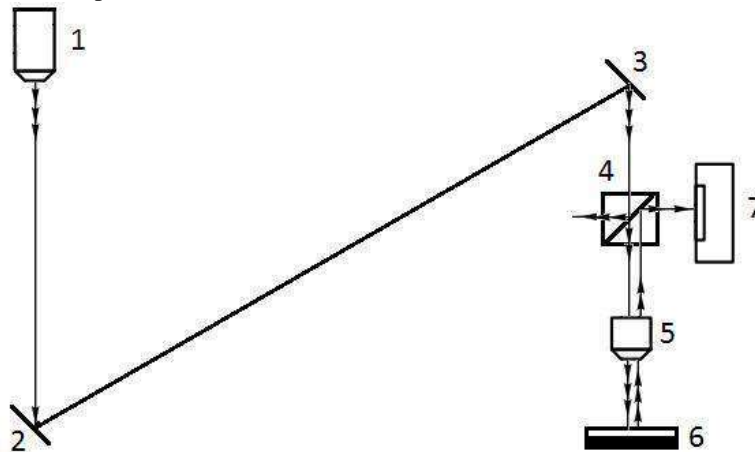


Fig.1. Optical setup for microexplosions polysterene microparticles on substrate covered by aluminum.

Let's turn to the experimental results. The original experiment took place at the maximum value of laser radiation power obtained with the pulse frequency of 3000 Hz. But with this value there was a destruction of the substrate surface covered with aluminum. Consequently, the beam power was reduced to a value of 6.17 kW, which stops the observed damage to the surface of the substrate. Thus, it minimized the influence of melting and destruction of the substrate surface on the movement of microparticles of polystyrene.

An experiment was carried out with deposition of polystyrene microparticles floating in the water on the substrate covered with aluminum. As the result, it was obtained a footage of polystyrene microparticles movement after the explosion, which is achieved by superposition of the incident and reflected beams of laser radiation. Now we will explain how we made these conclusions.

Firstly, in order to eliminate the influence of melting and destruction of the substrate surface covered with aluminium on the movement of polystyrene particles, the particles were inputted to the surface in the water. Thus, the water absorbs the portion of energy that came from breaking the surface of the substrate.

Given that the light almost does not pass through the surface covered with aluminum, then all the power goes into heating the surface (this effect was eliminated due to the choice of required capacity and the introduction of particles in solution with water), and the remaining part is reflected (about 93%). We find that in the explosion of polystyrene microparticles affects only the energy of the incident beam and the energy of the reflected beam.

In order to see whether the incident and reflected beams interact with each other or the main contribution is made only by the incident beam, another experiment was carried out which eliminates these issues. The substrate covered with aluminium was replaced by a quartz substrate with the transmittance of about 99%. Thus, if the former power of the laser radiation the microexplosion of polystyrene microparticles will not occur it turns out that this reflected beam interacts with the incident beam in a certain way.

After conducting full-scale experiments on a substrate covered with aluminum, there was an explosion of polystyrene microparticles, in which a displacement of nearby microparticles of polystyrene occurred. In case of changing of substrate coated with aluminium on a quartz substrate and leaving unchanged all the parameters of the scheme the explosion of polystyrene particles was not observed. As the result, we can conclude that explosion of polystyrene microparticles occurs on a substrate covered with aluminum due to the interaction of the incident and reflected beams.

It was also calculated the average rate of expansion of the polystyrene particles located on a substrate coated with aluminium after the explosion of a nearby polystyrene particles. This value is 0.77 mm/s.

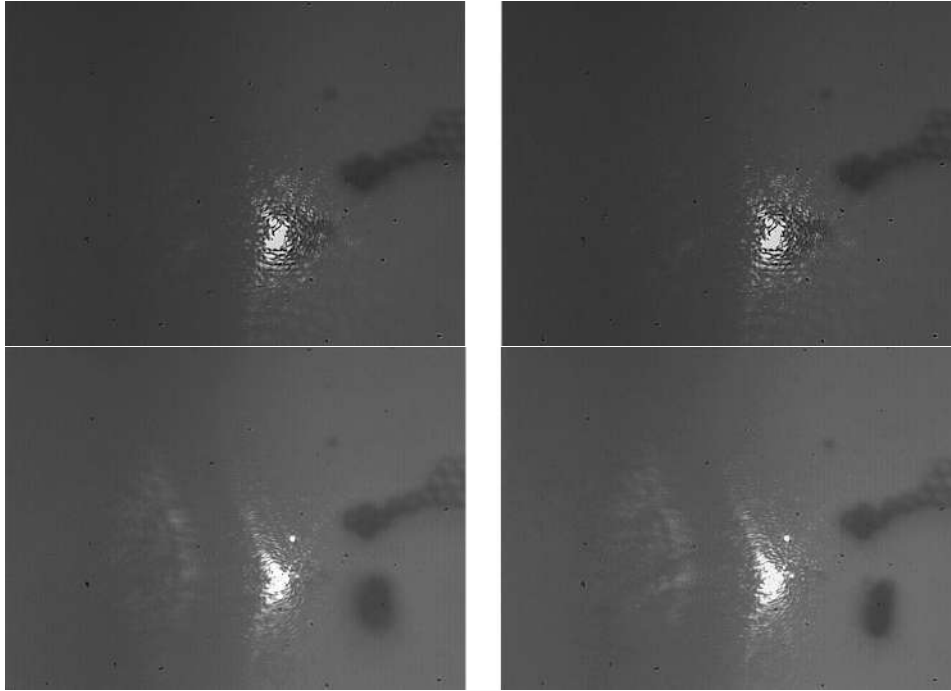


Fig.2. Results of the experiment on explosions polysterene microparticles using substrate covered by aluminum. Time interval between frames is equal to 10 ms.

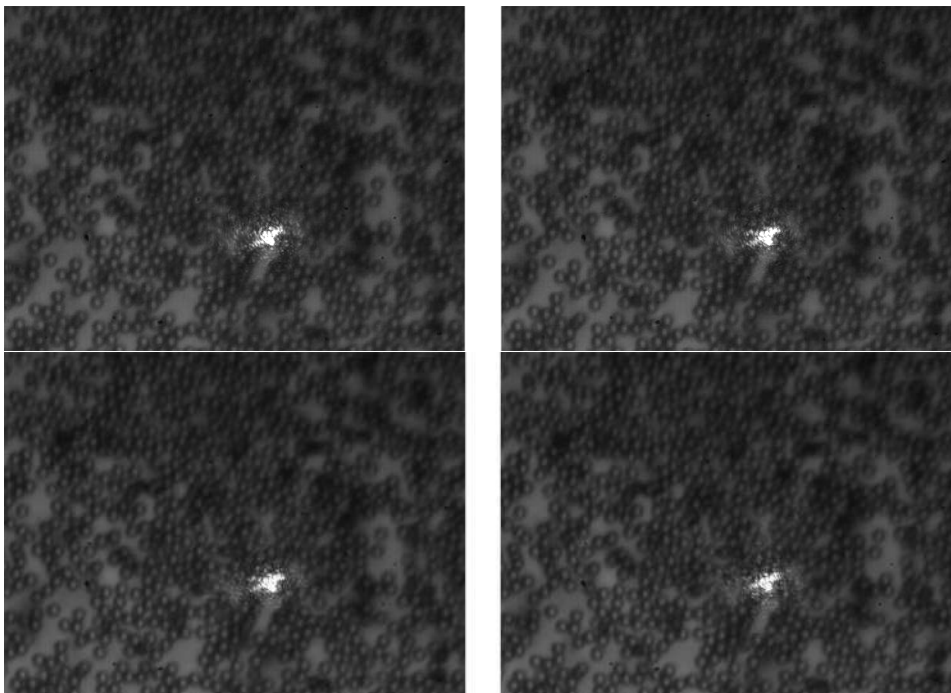


Fig.3. Results of the experiment on explosions polysterene microparticles using quartz substrate. Time interval between frames is equal to 10 ms.

#### 4. Conclusion

In the process of performing this work the following results were obtained:

- 1) An experiment was conducted, in which we discovered the existence of the interference between incident and reflected waves using aluminum substrate.
- 2) The expansion velocity of the particles using a 20x focusing microobjective was experimentally calculated. The speed value is 0.77 mm/s;
- 3) The parameters of explosion occurrence were found. The parameter is the average radiation power. The explosion occurs when values are higher than 6.17 kW.

## Acknowledgements

The work was partially funded by the RF Ministry of Science and Education as part of state-assigned task No. 3.3025.2017/PCh and by the Russian Foundation for Basic Research (RFBR) (grants Nos. 16-29-11744 ofi\_m and 16-29-09528 ofi\_m).

## References

- [1] Zemranek P, Jonras A, Sramek L, Liska M. Optical trapping of nanoparticles and microparticles by a Gaussian standing wave. *Optics Letters* 1999; 24(21): 1448–1450.
- [2] De AK, Roy D, Dutta A, Goswami D. Stable optical trapping of latex nanoparticles with ultrashort pulsed illumination. *Applied Optics* 2009; 48(31): 33–37.
- [3] Bosanac L, Aabo T, Bendix PM, Oddershede LB. Efficient Optical Trapping and Visualization of Silver Nanoparticles. *Nano Letters* 2008; 8(5): 1486–1491.
- [4] Khonina SN, Kotlyar VV, Skidanov RV, Soifer VA. Diffraction optical elements for optical manipulation of microparticles. Official materials scientific and practical conference “Holography in Russia and abroad. Science and practice” 2004; 2(4): 62.
- [5] Thalhammer G, Steiger R, Meinschad M, Hill M, Bernet S, Ritsch-Marte M. Combined acoustic and optical trapping. *Biomedical Optics Express* 2011; 2(10): 2859–2870.
- [6] Skidanov RV, Morozov AA, Porfirev AP. Composite light beam and microexplosion for optical micromanipulation. *Computers Optics* 2010;34(3): 371–375.
- [7] Sivuhin DV. Course of general physics V. 4. Optics. M.: Science, 1980.
- [8] Bardin AN. Assembly and alignment optical devices. M.: Higher school, 1968.

# Generation of regular optical pulses in VCSELs below the static threshold

A.A. Krents<sup>1,2</sup>, N.E. Molevich<sup>1,2</sup>, D.A. Anchikov<sup>1</sup>, S.V. Krestin<sup>2</sup>

<sup>1</sup>Samara National Research University, Moskovskoye Shosse 34, 443086, Samara, Russia

<sup>2</sup>Lebedev Physical Institute, Novo-Sadovaya Str. 221, 443011, Samara, Russia

---

## Abstract

We study numerically the dynamics of a vertical-cavity surface emitting laser (VCSEL) with external optical injection and asymmetrical triangular current modulation. Even if the average current is below the threshold, the VCSEL without optical injection emits irregular optical pulses. External optical injection stabilizes the laser output, reduces the standard deviation of the generated pulses and increases their averaged amplitude. The results of this study make it possible to reduce the threshold current.

*Keywords:* vertical-cavity surface-emitting lasers (VCSELs); optical injection; polarization; semiconductor lasers

---

## 1. Introduction

Vertical cavity surface emitting lasers (VCSELs) are among the most attractive light sources in modern optical devices, especially for both digital and analog photonic communication systems. They have the low threshold current, high modulation bandwidth, and emit a single-longitudinal mode and circular output beams that result in high coupling efficiencies into optical fibers. The output radiation of VCSELs is polarized along one of the two linearly polarized modes, aligned to the crystallographic directions. Under current modulation, nonlinear effects such as period doubling, chaos and multistability can arise [1, 2]. In this paper, using the standard spin-flip model extended to optical injection, we demonstrate that VCSEL, with asymmetric triangular current modulation and a small coherent optical injection, can generate regular optical pulses in both orthogonal polarizations.

Using asymmetric modulation makes it possible to reduce the threshold current and thermal heating of laser active medium.

## 2. Model

Polarization properties of VCSEL are described by the spin-flip model extended to optical injection [3]:

$$\begin{aligned} E_x &= k(1+i\alpha)\left[(N-1)E_x + inE_y\right] - i(\gamma_p + \Delta\omega)E_x - \gamma_a E_x + kE_{inj} \cos(\psi) + \sqrt{\beta_{sp}} \xi_x, \\ E_y &= k(1+i\alpha)\left[(N-1)E_y - inE_x\right] + i(\gamma_p - \Delta\omega)E_y + \gamma_a E_y + kE_{inj} \sin(\psi) + \sqrt{\beta_{sp}} \xi_y, \\ N &= \gamma_N \left[ \mu(t) - N \left( 1 + |E_x|^2 + |E_y|^2 \right) - in(E_y E_x^* - E_x E_y^*) \right], \\ n &= -\gamma_s n - \gamma_N \left[ n \left( |E_x|^2 + |E_y|^2 \right) + iN(E_y E_x^* - E_x E_y^*) \right], \end{aligned} \quad (1)$$

where  $k$  is the field decay rate,  $\gamma_N$  is the decay rate of the total carrier population,  $\gamma_s$  is the spin-flip rate which accounts for the mixing of carrier populations with different spins,  $\alpha$  is the linewidth enhancement factor,  $\gamma_a$  and  $\gamma_p$  are linear anisotropies representing dichroism and birefringence,  $\Delta\omega$  is the detuning parameter,  $\Psi$  is the angle between the x axis and the direction of the linearly polarized optical injection,  $\beta_{sp}$  is the noise strength,  $\xi_{x,y}$  are uncorrelated Gaussian white noises, and  $\mu(t)$  is the normalized injection current parameter (the static threshold is at  $\mu_{th,s} = 1$ ).

The current is modulated with an asymmetric triangular periodic signal of amplitude  $\Delta\mu$ , rising from  $\mu_0$  during the time interval  $T_1$  and falling back to  $\mu_0$  during the time interval  $T_2$ . One modulation cycle is:

$$\begin{aligned} \mu(t) &= \mu_0 + \Delta\mu(t/T_1) \text{ for } 0 \leq t \leq T_1, \\ \mu(t) &= \mu_0 + \Delta\mu \left[ 1 - (t - T_1)/T_2 \right] \text{ for } T_1 \leq t \leq T_1 + T_2. \end{aligned}$$

The average current,  $\mu_{ave} = \mu_0 + \Delta\mu/2$ , is independent of the modulation period,  $T = T_1 + T_2$ . The asymmetry of the modulation is characterized by the parameter  $\alpha_a = T_1/T$  with  $0 \leq \alpha_a \leq 1$ .

## 3. Results and Discussion

The equations were simulated with typical VCSEL parameters [4]:  $k = 300 \text{ ns}^{-1}$ ,  $\alpha = 3$ ,  $\gamma_N = 1 \text{ ns}^{-1}$ ,  $\gamma_a = 0.5 \text{ ns}^{-1}$ ,  $\gamma_p = 50 \text{ rad/s}$ ,  $\gamma_s = 50 \text{ ns}^{-1}$ , and  $\beta_{sp} = 10^{-6} \text{ ns}^{-1}$ . Asymmetrical triangular modulation of current leads to the generation of irregular optical pulses even if, on average, the current is below the threshold [3]. There is an optimal modulation asymmetry, typically  $\alpha_a \cong 0.8$ , for which the averaged intensity and the averaged pulse amplitude reach their maximum value, and for this asymmetry, the dispersion of the pulse amplitude reaches its minimum value. Figure 1(a)–(f) displays time traces of the

$I_x = |E_x|^2$  and  $I_y = |E_y|^2$  for three different optical injection values and fixed other parameters: average current value  $\mu_{ave} = 0.87$ , asymmetry  $\alpha_a = 0.8$ , frequency detuning between injection and laser mode  $\Delta\omega = 0$ ,  $\Psi = \pi/4$  (injection in both polarizations is equal).

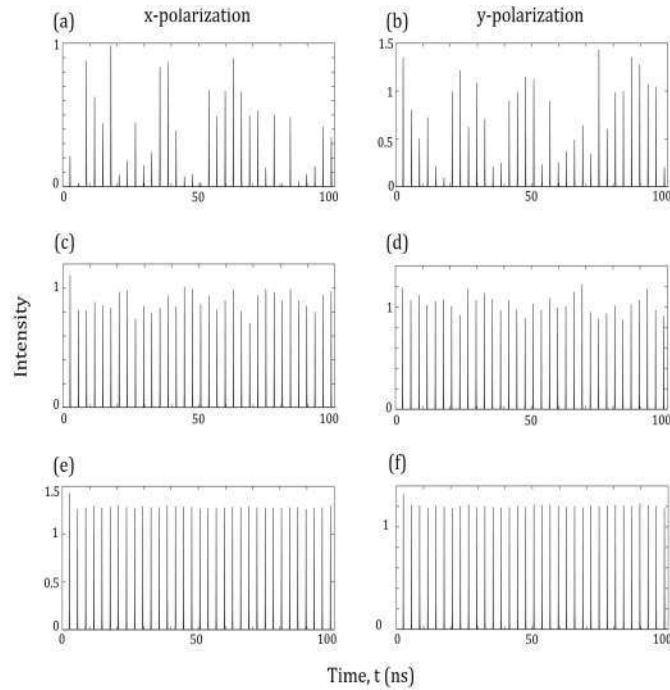


Fig.1. Time traces of intensities of the orthogonal linear polarization: (a), (b)  $E_{inj} = 0$ , (c), (d)  $E_{inj} = 10^{-5}$ , (e), (f)  $E_{inj} = 10^{-4}$ .

Figure 2(a) displays standard deviation of the intensity of the pulse, depending on the value of the optical injection. It can be observed that injection of the optical signal leads to more regular pulses. Also, injection leads to increasing of the mean value of generated pulse amplitude for both polarizations (Figure 2(b)). Thus, optical injection stabilizes the laser output and increases the laser efficiency.

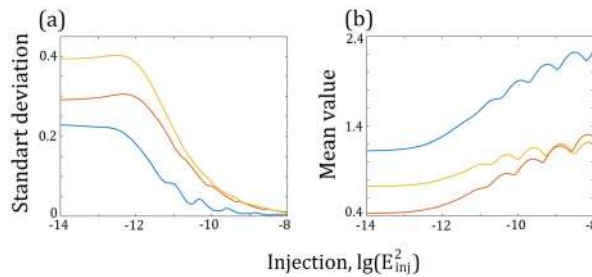


Fig.2. Standard deviation of the intensity of the pulses (a). Mean of the intensity of the pulses (b). Red is the x-polarization, yellow is the y-polarization and blue is the total intensity.

Figure 3(a) displays the standard deviation of the total intensity of the pulses, depending on the value of the injection angle. Figure 3(b) displays the mean of the total intensity of the pulses. The mean of the total intensity has the maximum value for the angle  $\Psi = \pi/2$  (parallel optical injection). Standard deviation has minimum for  $\Psi = \pi/2$  and  $\Psi = 0$ . x-polarization vanishes for parallel optical injection and y-polarization vanishes for orthogonal optical injection ( $\Psi = 0$ ). Thus, the optical injection in the y-mode is the most benefit.

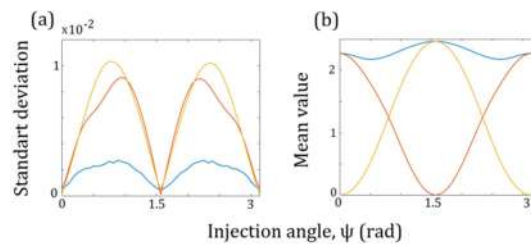


Fig. 3. Standard deviation of the intensity of the pulses (a). Mean of the intensity of the pulses (b).  $E_{inj} = 10^{-4}$ . Red is the x-polarization, yellow is the y-polarization and blue is the total intensity.

The frequency detuning between injected and generated radiation  $\Delta\omega$  is the important parameter of the model. Figures 4(a), (b) display the standard deviation of the total intensity of the pulses and the mean of the total intensity of the pulses for  $\Psi = \pi/4$ . The standard deviation of the total intensity of the pulses has minimum for  $\Delta\omega \approx \gamma_p = 50$  (in this case injection is coherent to the y-mode). The mean of the total intensity of the pulses has maximum for the same detuning value. For  $\Psi = 0$ , the standard deviation of the total intensity of the pulses has minimum for  $\Delta\omega = 0$  and the mean of the total intensity of the pulses has maximum for the same detuning value. For  $\Psi = \pi/2$ , the standard deviation of the total intensity of the pulses has minimum for  $\Delta\omega \approx \gamma_p = 50$  and the mean of the total intensity of the pulses has maximum for the same detuning value.

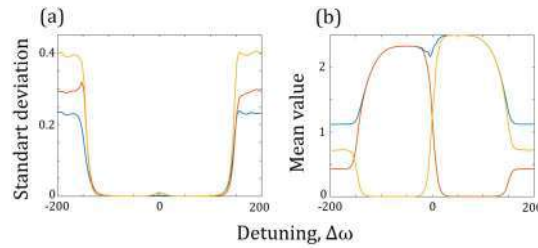


Fig.4. Standard deviation of the intensity of the pulses (a). Mean of the intensity of the pulses (b).  $E_{inj} = 10^{-4}$ ,  $\Psi = \pi/4$ . Red is the x-polarization, yellow is the y-polarization and blue is the total intensity.

Thus, both amplitude and polarization of pulses generated by asymmetrically modulated VCSELs can be stabilized by weak external optical injection.

#### 4. Conclusion

In summary, asymmetrically periodically modulated VCSELs with a pumping current below the threshold and an external optical injection, have been numerically investigated. We have shown that injection of weak external optical signal stabilizes the VCSEL generation. We showed that generation of quasiregular optical pulses is already possible for the injection value  $E_{inj} = 10^{-4}$ , which is only about  $5 \cdot 10^{-9}$  of the output intensity of generation. Optical injection increases the mean amplitude of generated pulses and decreases the standard deviation of the intensity of the pulses. Control of the output radiation polarization is also possible. We showed the possibility of smooth adjustment of the polarization of the generated pulses by varying the optical injection angle  $\Psi$ . We found that the standard deviation of the intensity of the pulses has minimum for injection in y-mode (parallel optical injection). Method proposed in this paper provides a generation of regular optical pulses in VCSELs below the static threshold.

#### Acknowledgements

This research was supported by Russian Foundation for Basic Research (16-32-60151 mol\_a\_dk); State assignment to educational and research institutions under project 3.1158.2017.

#### References

- [1] Agrawal GP. Effect of gain nonlinearities on period doubling and chaos in directly modulated semiconductor lasers. Appl. Phys. Lett. 1986; 49: 1012–1015
- [2] Zamora-Munt J, Masoller C. Generation of optical pulses in VCSELs below the static threshold using asymmetric current modulation. Opt. Express 2008; 16(22): 17848–17853.
- [3] San Miguel M, Feng Q, Moloney JV. Light-polarization dynamics in surface-emitting semiconductor. Phys. Rev. A 1995; 52: 1728–1739.
- [4] Martin-Regalado J, Prati F, San Miguel M, Abraham NB. Polarization properties of vertical-cavity surface-emitting lasers. IEEE J. Quantum Electron 1997; 33: 765–783.



# Experimental observing of transformation Bessel beam spreading along axis of crystal during wavelength changes

V.S. Vasilev<sup>1</sup>, V.V. Podlipnov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

## Abstract

In paper describe experimental observing transform bessel beam, formed by diffraction axicon in moment propagation through anisotropic birefringence crystal. This observation covers large range wavelength changes (from 520 nm to 534 nm). Theoretical explain effect is given.

*Keywords:* laser with changing wavelength; diffraction axicon; birefringent crystal; bessel beams

## 1. Introduction

It is well-known the usage of anisotropic elements to convert beams with the homogeneous polarization into cylindrical vector beams [1-6]. At the same time, it is necessary to implement the separation of the longitudinal modes along the optical axis of the system, which is parallel to the axis of the crystal. To improve convergence of the beams in the crystal, it is possible to use telescopic system or to form beams with high numerical aperture. Polarization and mode conversion during propagation along the axis of the crystal were considered for both Bessel and Gaussian beams [7-16].

It has been shown in the studies [17, 18] that during the propagation along the crystal axis neuraxial Bessel beams have other properties than Gaussian beams, namely, experiencing a uniform periodic change of intensity. In this case, the Bessel beam of zero order and second-order are periodically converted from one to another [7-9, 17, 18]. The oscillation period is directly proportional to the wavelength of the laser radiation and inversely proportional to the square of the spatial frequency of the laser beam and the difference of the dielectric capacitivity, which is corresponding to the ordinary and extraordinary rays. This dependence allows control occurring transformation in the crystal due to changes of the characteristics in either Bessel beam or crystal. In particular, the spatial frequency of the beam depends on the numerical aperture of the axicon [19-22] which shapes the beam, also it is possible to adjust characteristics of the beam by changing the beam divergence [23]. To change the parameters of the crystal, it can be heated [24] or effected by electro-optic [25]. However, the most convenient way of adjustment is to change the wavelength of the laser radiation which has a direct linear relationship from the period of transformation [26].

It was experimentally demonstrated the ability [26] to manage the transformation of the Bessel beam at the output of the CaCO<sub>3</sub> crystal by changing the wavelength of the radiation illuminating the diffractive axicon. It was achieved almost complete transformation of the Bessel beam of zero order beam to the second order using the axicon period of 2 μm and the wavelength at Δλ = 1.5 of the initial value of λ = 637.5. The variation of the wavelength within a small range was achieved by changing the temperature of the laser. In contrast to this method, the usage of a laser with variable wavelength provides a wide range of Δλ, and therefore the possibility of achieving complete conversion using the axicon with a large period, i.e., a smaller numerical aperture. Note that the usage of axicons with high numerical aperture is limited not only with technological possibilities of production [27] and reduction of non-diffraction distribution cut [20], but with the limiting numerical aperture [28], in which propagating waves occur in the considered optical medium.

This paper shows the results of experimental observation of the mode conversion of Bessel beam formed by the axicon amplitude with a period of 3 μm with the output of a deuterated potassium dihydrogen phosphate crystal when the wavelength of the laser EKSPLA NT 200 radiation is changed.

## 2. Theoretical analysis

Consider an anisotropic crystal whose axis is oriented along the optical axis.

The intensity distribution I(x,y,z) in the propagation of Bessel beam along the axis of the crystal is as follows [9, 11, 17, 18]:

$$I(x, y, z) \approx \frac{1}{2} \left[ |C(z)|^2 J_0^2(k \sqrt{x^2 + y^2}) + |S(z)|^2 J_2^2(k \sqrt{x^2 + y^2}) \right] \quad (1)$$

where  $J_0(\cdot)$  и  $J_2(\cdot)$  - Bessel functions of zero and second order, respectively,

$$\begin{aligned} C(z) &= \exp(ikz\gamma_o) + \exp(ikz\gamma_e), \\ S(z) &= \exp(ikz\gamma_o) - \exp(ikz\gamma_e), \end{aligned} \quad (2)$$

where α – numerical aperture of the beam, z – is the distance traveled; γ<sub>o</sub>, γ<sub>e</sub> - are the values which are determining the direction of propagation of the ordinary and extraordinary rays:

$$\sqrt{\quad}$$

$$\begin{aligned}\gamma_o &= n_o^2 - \alpha^2, \\ \gamma_e &= \sqrt{n_o^2 - \alpha^2 n_{oe}^2 / n_e^2},\end{aligned}\quad (3)$$

where  $n_o$ ,  $n_e$  – ordinary and extraordinary refractive indices of the crystal.

A complete transformation of the Bessel beam of zero order to the beam of second order will periodically occur at distances that are multiples of the value:

$$z_p = \frac{\lambda}{\gamma_o - \gamma_e} \approx \frac{2\lambda n_o^2}{\alpha^2 (n_o^2 - n_e^2)} \quad (4)$$

Full transformation period depends on the refractive indices of the crystal and the numerical aperture of the axicon, as well as on the wave length of radiation. Moreover, the wavelength dependence is direct and linear, i.e. the most convenient to dynamically change the value of period so that the output of the crystal is formed the desired pattern.

### 3. Experimental results

#### 3.1. Method of experiment

In this paper experiments were conducted using the optical arrangement shown in the fig.1, where 1 – laser with changing wavelength EKSPLA NT 200, 2 – diaphragm, 3 – collimator, 4 – diaphragm, 5 - DOE, 6 – anisotropic crystal, 7 – 20x microobjective, 8 – digital USB camera TOUPCAM UCMOS05100KPA.

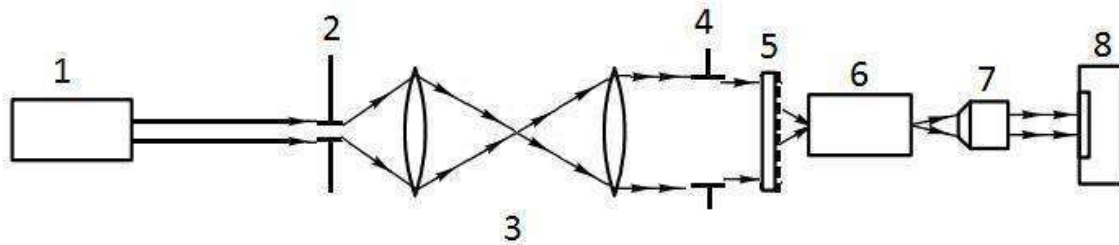


Fig. 1. Optical setup of the experiment.

A laser with variable wavelength was used as a radiation source EKSPLA NT 200. In the considered range of wavelength variation (520 – 534 nm), the laser beam has a horizontal X-polarization. The energy of the laser radiation obtained in the range of the visible spectrum wavelength is variable from 610 microjoule (450 nm) to 45 microjoule (700 nm) . The extension of the beam is done by the collimator. Owing to the fact that the beam emerging from the laser has a Gaussian intensity distribution, it has become necessary to select a part of the beam with a small change of intensity. This problem can be solved by introduction of a diaphragm 2. Septum 4 allows to limit the numerical aperture and to enable formation of the propagating waves. The intensity distribution of the output beam was recorded with a digital USB camera with a resolution of 5 mega pixels and ADC digit capacity of 12 bits.

The Bessel beam of zero order is formed by using a diffraction amplitude of the axicon with period which operates with nearly the same effectiveness in the considered wavelength range. The Bessel beam was directed along the axis of a crystal with cross-sectional dimension and length 20 mm. As a result of Bessel beams transformation there were formed interference pattern intensity distribution for different wavelengths and it was recorded with the microscope objective and digital cameras (table 1). To highlight different X and Y components of the transformed beams a rotating analyzer was installed in front of the digital camera.

#### 3.2. Results and discussion

As you can see in the images, when the wavelength changes by  $\Delta\lambda=14$  nm there is a complete transformation of the Bessel beam of the first order to the second, which is caused by the reaction of doubly refracting crystal. The observed phenomenon is explained by the formula (4), where  $\lambda$  is in the numerator. At the same time, change of wavelength is similar to the changes of the propagation length of the beam, as if it had been changed the dimensions of the crystal. To verify the observed phenomenon in the described conversion model it was carried out an additional numerical calculation for the wavelength of 520 nm and 532 nm. Intensity distributions of the Bessel beams images which were converted by electro-optic crystal for given experimental conditions obtained by numerical calculation are presented in table 2.

Based on the simulation results, we can conclude that the observed experimental results are very similarity with the mathematical description for the Bessel beams conversion in the considered wavelength range.

Table 1. Distribution of intensity bessel beams transformed in birefringent crystal.

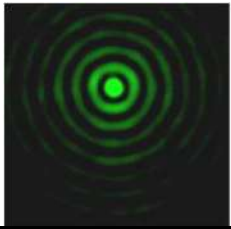
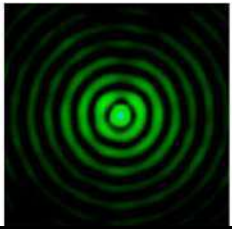
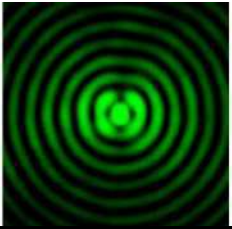
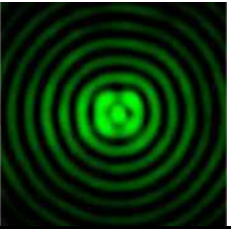
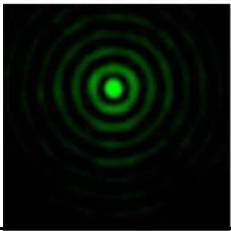
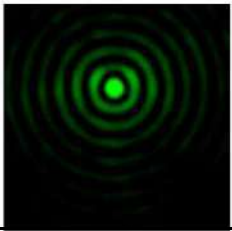
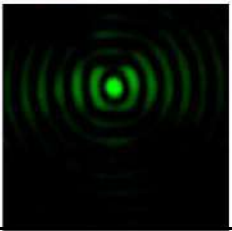
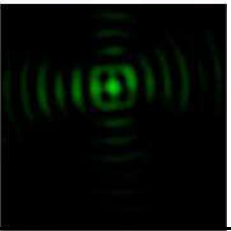
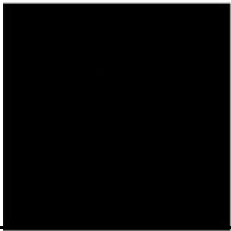

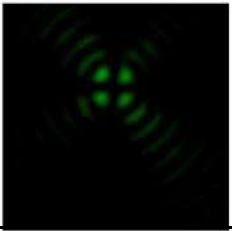
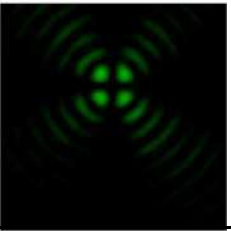
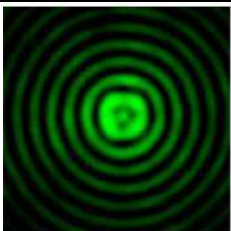
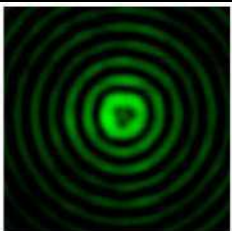
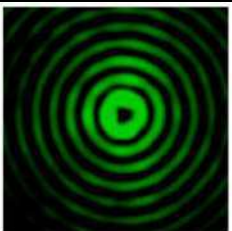
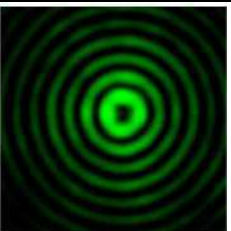


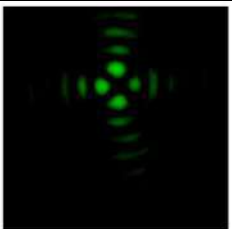
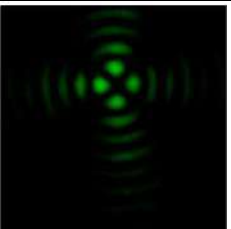
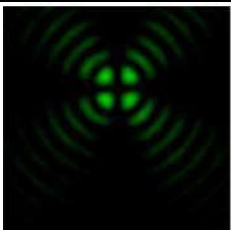
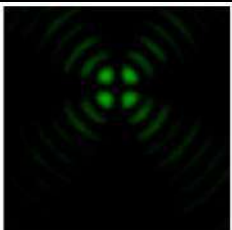
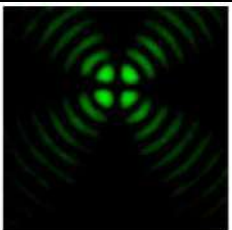
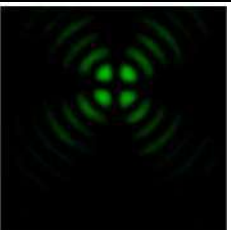
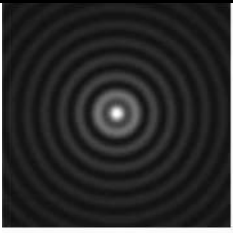
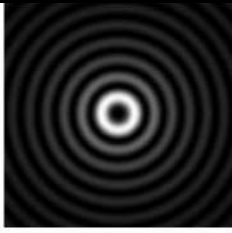
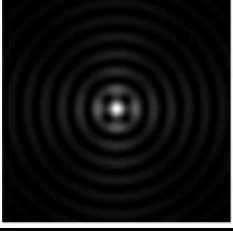
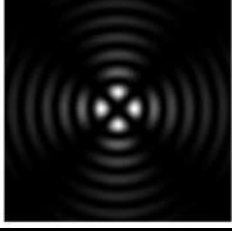
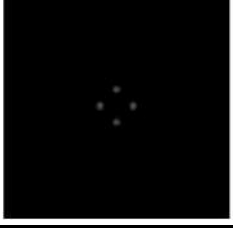
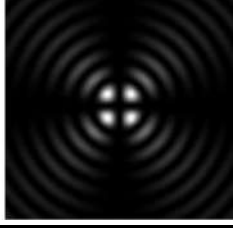
Wavelength	520	522	524	526
				
X component				
Y component				
Wavelength	528	530	532	534
				
X component				
Y component				

Table 2. Modeling distribution of intensity bessel beams transformed in birefringent crystal.

Wavelength	520	532
		
X component		
Y component		

#### 4. Conclusion

It was experimentally demonstrated the conversion of Bessel beams of zero order, generated by the axicon with period in birefringence crystal, depending on the change of the wavelength of the laser radiation in the range of  $\lambda = 520\text{-}534$  nm to the Bessel beams of the second order, which has an annular intensity distribution. Further increase of the wavelength has showed a recurrent re-transformation into a Bessel beam of zero order. Comparative analysis of experimental images of full intensity and their components with images obtained by the numerical simulation has showed their similarity.

#### Acknowledgement

This work was supported by the Ministry of Education, by the Russian Foundation for Basic Research (grant 16-29-11698 ofi\_m and 16-07-00494 a) and by the grant from the President of the Russian Federation to support young Russian scientists–doctors of science, (project no. MD- 5205.2016.9).

#### References

- [1] Machavariani G, Lumer Y, Moshe I, Meir A, Jackel S, Davidson N. Birefringence-induced bifocusing for selection of radially or azimuthally polarized laser modes. *Applied Optics* 2007; 46(16): 3304.
- [2] Yonezawa K, Kozawa Y, Sato S. Compact laser with radial polarization using birefringent laser medium. *Journal of Applied Physics* 2007; 1(1): 5160.
- [3] Zhan Q. Cylindrical vector beams: from mathematical concepts to applications. *Advances in Optics and Photonics* 2009; 1(1): 57.
- [4] Fadeyeva T, Shvedov V, Shostka N, Alexeyev C, Volyar A. Natural shaping of the cylindrically polarized beams. *Optics Letters* 2010; 235(22): 3787.
- [5] Khonina SN, Karpeev SV, Alferov SV. Theoretical and an experimental research of polarizing transformations in uniaxial crystals for generation cylindrical vector beams of high orders. *Computer Optics* 2014; 38(2): 171–180.
- [6] Khonina SN, Karpeev SV, Alferov SV, Soifer VA. Generation of cylindrical vector beams of high orders using uniaxial crystals. *Journal of Optics* 2015; 17(1): 11.
- [7] Khilo NA, Ryzhevich AA, Petrova ES. Transformation of the order of Bessel beams in uniaxial crystals. *Quantum Electronics* 2001; 31(1): 85-89.
- [8] Khilo NA. Diffraction and order conversion of Bessel beams in uniaxial crystals. *Optics Communications* 2012; 285(1): 503–509.
- [9] Khonina SN, Morozov AA, Karpeev SV. Effective transformation of a zero-order Bessel beam into a second-order vortex beam using a uniaxial crystal. *Laser Phys.* 2014; 24(1): 5.
- [10] Khonina SN, Parandin VD, Ustinov AV, Krasnov AP. Astigmatic transformation of Bessel beams in a uniaxial crystal. *Optica Applicata* 2016; Vol. 46(1): 5–18.
- [11] Khonina SN, Karpeev SV, Morozov AA, Parandin VD. Implementation of ordinary and extraordinary beams interference by application of diffractive optical elements. *Journal of Modern Optics* 2016; 63(13): 1239–1247.
- [12] Ciattoni A, Cincotti G, Palma C. Circularly polarized beams and vortex generation in uniaxial media. *J. Opt. Soc. Am. A* 2003; 20(1): 163–171.
- [13] Marrucci L, Manzo C, Paparo D. Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media. *Phys. Rev. Lett.* 2006; 96(1): 130–135.
- [14] Loussert C, Brasselet E. Efficient scalar and vectorial singular beam shaping using homogeneous anisotropic media *Optics Letters* 2010; 35(1): 7–9.

- [15] Fadeyeva TA, Shvedov VG, Izdebskaya YV, Volyar AV, Brasselet E, Neshev DN, Desyatnikov AS, Krolikowski W, Kivshar YS. Spatially engineered polarization states and optical vortices in uniaxial crystals. *Optics Express* 2010; 18(10): 63.
- [16] Picon A, Benseny A, Mompart J, Calvo GF. Spin and orbital angular momentum propagation in anisotropic media: theory. *J. Opt.* 2011; 13(1): 7.
- [17] Khonina SN, Volotovskiy SG, Kharitonov SI. Features of nonparaxial propagation of Gaussian and Bessel beams along the axis of the crystal. *Computer Optics* 2013; 37(3): 297–306.
- [18] Khonina SN, Kharitonov SI. Comparative investigation of nonparaxial mode propagation along the axis of uniaxial crystal. *Journal of Modern Optics* 2015; 62(2): 125–134.
- [19] McLeod, JH. The axicon: a new type of optical element. *Journal of the Optical Society of America* 1954; 44: 592–597.
- [20] Turunen J, Vasara A, Friberg AT. Holographic generation of diffraction-free beams. *J. Appl. Opt.* 1988; 27(19): 3959–3962.
- [21] Khonina SN, Kotlyar VV. Bessel-mode formers. *Proceedings of SPIE* 1994; 23(63): 184–190.
- [22] Chattrapiban N, Rogers E, Cofield D, Hill W, Roy R. Generation of nondiffracting Bessel beams by use of a spatial light modulator. *Opt. Lett.* 2003; 28(22): 2183–2185.
- [23] Parinin VD, Karpeev SV, Khonina SN. Control of the formation of vortex Bessel beams in uniaxial crystals by varying the beam divergence. *Quantum Electronics* 2016; 46(2): 163–168.
- [24] Parinin VD, Khonina SN, Karpeev SV. Control of the optical properties of a CaCO<sub>3</sub> crystal in problems of generating Bessel vortex beams by heating. *Optoelectronics, Instrumentation and Data Processing* 2016; 52(2): 174–179.
- [25] Khonina SN, Parinin VD. Electro-optical correction of Bessel beam conversion along axis of a barium niobate-strontium crystal. *Computer Optics* 2016; 40(4): 475–481. DOI: 10.18287/2412-6179-2016-40-4-475-481.
- [26] Parinin VD, Karpeev SV, Khonina SN. Transformation of Bessel beams in c-cuts of uniaxial crystals by varying the emission source wavelength. *Journal of Russian Laser Research* 2016; 37(3): 207–210.
- [27] Cherkashin VV, Kharissov AA, Korol'kov VP, Koronkevich VP, Poleshchuk AG. Accuracy potential of circular laser writing of DOEs. *Proceedings of SPIE* 1997; 3348: 58–68.
- [28] Ustinov AV, Khonina SN. Analysis of laser beam diffraction by axicon with the numerical aperture above limiting. *Computer Optics* 2014; 38(2): 213–222.

# Amplitude and polarization transformations of the Bessel beam as it passes through an anisotropic crystal perpendicular to the axis of the crystal

A.V. Glazkova<sup>1</sup>, M.V. Zablovskaya<sup>1</sup>, V.V. Podlipnov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

## Abstract

A comparative numerical calculation of the propagation of a zero-order Bessel laser beam in a uniaxial crystal perpendicular to its axis is performed using the Rayleigh-Sommerfeld integral operator, generalized for an anisotropic environment. Numerical simulation is performed with a different type of beam polarization and different characteristics of the Bessel beam. Patterns of the beam intensity during the passing of different distances in the crystal are obtained, showing the degree of astigmatic transformation, which makes it possible to determine the conditions under which the greatest astigmatic distortion of the beams occurs. The above analysis can be useful in practice for determining the anisotropy characteristics of a crystal.

**Keywords:** diffraction axicon; birefringent crystal; polarization transformations; amplitude transformations, Bessel beams; astigmatism

## 1. Introduction

Optical devices are becoming more and more interesting and practical. They allow to transform certain properties of electromagnetic radiation into others. Most often, modal transformations (from the fundamental mode to higher order distributions) and polarization (from homogeneous linear polarization to more complex ones) are required. One of the tools of such transformations are anisotropic crystals. The propagation of laser modes with a high numerical aperture in an environment with strong anisotropy leads to complex polarization-mode transformations [1-6].

In particular, when propagating along the crystal axis, the spin angular momentum is transformed, which has a circularly polarized beam at the orbital angular momentum [7-13]. It was shown in [6, 7, 11-13] that when propagating along the crystal axis, nonparaxial Bessel beams undergo a periodic change in intensity, corresponding to a transformation into a higher-order beam. In publications [14-20], polarization transformations of beams focused along the crystal axis were considered.

The propagation of various types of laser beams perpendicular to the axis of the crystal was investigated in [21-26]. The most interesting transformations were observed for Bessel beams [16, 21, 24, 27], since in this case there is a visually pronounced astigmatic distortion of the ring structure of the beam. A similar distortion can be observed with oblique incidence of a plane wave on a diffraction axicon [28-30], and also with a cylindrical lens [31]. This analogy was noted in [24], and the analytical basis for such an effect was given in [27].

In this paper, the effect of the astigmatic transformation of Bessel beams propagating perpendicular to the crystal axis is studied in detail on the basis of numerical simulation. The calculation was carried out using the Rayleigh-Sommerfeld integral operator, generalized for an anisotropic environment [32, 33]. Numerical simulation is performed for different types of beam polarization and different characteristics of the Bessel beam. The formation of Bessel beams [34-37] was carried out with the diffraction axicon with different period of the radial lattice. The effect of the relative position of the polarization plane of the radiation and the c-axis of the crystal on the intensity distributions formed in different vector components of ordinary and extraordinary beams is investigated. Patterns of the beam intensity are obtained during the passing of different distances in the crystal, showing the degree of astigmatic transformation, which makes it possible to determine the conditions under which the greatest astigmatic distortion of the beams occurs. The above analysis can be useful in practice for determining the anisotropy characteristics of a crystal.

## 2. Theoretical analysis

Consider an anisotropic crystal whose axis is oriented perpendicular to the propagation axis and coincides with the Oy axis. In this case, the field propagation in a crystal with dielectric permittivities, (ordinary and extraordinary) can be described by an expression similar to the Rayleigh-Sommerfeld integral [32, 33]:

$$\mathbf{E}(u, v, z) = \frac{2\pi z}{\lambda^2} \sum_{j=1}^2 \iint \mathbf{e}_j(\alpha_{jc}, \beta_{jc}) \left[ \mathbf{w}_j(\alpha_{jc}, \beta_{jc})^T \mathbf{E}_\perp(x, y, 0) \right] \frac{\sqrt{d_j s_j t_j}}{R_j^2} \exp \left\{ ik \sqrt{\frac{d_j}{s_j t_j}} R_j \right\} dx dy, \quad (1)$$

where the indices correspond to the ordinary ( $j = 1$ ) and extraordinary ( $j = 2$ ) waves,  $d_1 = d_2 = \varepsilon_o$ ,  $s_1 = t_1 = 1$ ,  $s_2 = t_2 = \varepsilon_o / \varepsilon_e$ .

For transverse (x- and y- components):

$$\begin{aligned}
 e_{1x}(\alpha, \beta) &= 1, \\
 e_{1y}(\alpha, \beta) &= 0, \\
 e_{2x}(\alpha, \beta) &= \alpha\beta, \\
 e_{2y}(\alpha, \beta) &= \beta^2 - \varepsilon_o.
 \end{aligned} \tag{2}$$

$$\begin{aligned}
 \mathbf{w}_1(\alpha, \beta)^T &= \left( 1, \frac{\alpha\beta}{(\varepsilon_o - \beta^2)} \right), \\
 \mathbf{w}_2(\alpha, \beta)^T &= \left( 0, -\frac{1}{(\varepsilon_o - \beta^2)} \right).
 \end{aligned} \tag{3}$$

$$\begin{cases} \alpha_{1c} = \sqrt{\varepsilon_o} \frac{(u-x)}{R_1}, \\ \beta_{1c} = \sqrt{\varepsilon_o} \frac{(v-y)}{R_1}, \end{cases} \quad \begin{cases} \alpha_{1c} = \sqrt{\varepsilon_o} \frac{(u-x)}{R_1}, \\ \beta_{1c} = \sqrt{\varepsilon_o} \frac{(v-y)}{R_1}, \end{cases} \tag{4}$$

where

$$\begin{aligned}
 R_1 &= \sqrt{(u-x)^2 + (v-y)^2 + z^2}, \\
 R_2 &= \sqrt{\frac{\varepsilon_e}{\varepsilon_o} \sqrt{(u-x)^2 + \frac{\varepsilon_o}{\varepsilon_e} (v-y)^2 + z^2}}.
 \end{aligned} \tag{5}$$

Similar results can be obtained if the crystal axis is directed along the Ox axis.

### 3. Results of numerical simulation

During the experiment, the axicon was used. The scheme of the axicon's work is shown in Fig. 1

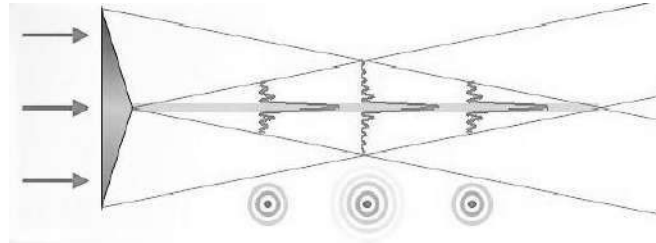


Fig. 1. The scheme of the axicon's work.

In order to carry out the simulation as an anisotropic medium, a lithium niobate crystal of the X-cut was chosen in this study, the dielectric constant of which is  $\varepsilon_0 = 5.2273505956$ ,  $\varepsilon_e = 4.8517551289$ . The refractive indices of this crystal are:  $n_0 = 2.28634$ ,  $n_e = 2.20267$ . For the formation of zero-order Bessel beams, diffraction axicons with periods  $d_1 = 1.2 \mu\text{m}$ ,  $d_2 = 2 \mu\text{m}$ ,  $d_3 = 4 \mu\text{m}$  were used and illuminated with light polarized linearly along the OY axis with a wavelength of  $\lambda = 632.8 \text{ nm}$ . We also compared the results of the transformation for different crystal thicknesses, which were chosen  $h_1 = 1047 \mu\text{m}$  and  $h_2 = 843 \mu\text{m}$ . To analyze the transformation of Bessel beams with axicons, the results of the simulation were presented in the form of patterns of light distribution of propagating beams separately for polarized light along OX, separately for OY and their superposition. The results of the simulation are presented in Table 1.

It can be noted that the picture of the Y component almost does not differ from the superposition picture of the X and Y components, which means that the X component has a negligible intensity, and the linearly polarized light at the exit from the lithium niobate crystal has not changed its polarization.

As can be seen from the modeled intensity distribution maps of lithium niobate transformed into an anisotropic lithium crystal by Bessel beams, the beams formed by axicons with the minimal period are subjected to the strongest astigmatic distortions. With an increase in crystal thickness, the degree of astigmatism increases in proportion to the propagation length.

When analyzing patterns of light intensity distribution at the output of an anisotropic crystal for linearly polarized light along the Y axis, with circular polarization, polarization rotated through an angle of  $45^\circ$  about the X axis, the above-described character of the Bessel beam transformation is preserved.

Table 1. Patterns of propagation of Bessel beams formed with axicons under illumination by light polarized along the OY axis through an anisotropic X-cut crystal.

	$d_1=1,2 \mu\text{m}$		$d_2=2 \mu\text{m}$		$d_3=4 \mu\text{m}$	
Component	$h_1=1,047 \text{ mm}$	$h_2=0,843 \text{ mm}$	$h_1=1,047 \text{ mm}$	$h_2=0,843 \text{ mm}$	$h_1=1,047 \text{ mm}$	$h_2=0,843 \text{ mm}$
General						
x						
y						

#### 4. Conclusion

In the work, to analyze the dependence of the propagation of the zero-order Bessel beam on the polarization angle, on the period and the radius of the axicon, we used the calculation with the Rayleigh-Sommerfeld integral operator generalized for an anisotropic medium. The Bessel beams formed by an axicon with the smallest period and passing through an anisotropic Crystal at the greatest distance. The described regularities can be used in practice to determine the degree of anisotropy or the exact thickness of the crystal cuts.

#### Acknowledgments

This work was supported by the Ministry of Education of the Russian Federation., by the Russian Foundation for Basic Research (RFBR grants 16-07-00825, 16-29-11698 ofi\_m and 16-07-00494 a) and by the grant from the President of the Russian Federation (project no. MD- 5205.2016.9).

#### References

- [1] Zhou Y, Wang X, Dai Ch. Nonparaxial analysis in the propagation of a cylindrical vector Laguerre-Gaussian beam in a uniaxial crystal orthogonal to the optical axis. *Opt. Commun.* 2013; 305: 113–125.
- [2] Khonina SN, Volotovskiy SG, Kharitonov SI, Features of nonparaxial propagation of Gaussian and Bessel beams along the axis of the crystal. *Computer Optics* 2013; 37(3): 297–306.
- [3] Loussert C, Brasselet E. Efficient scalar and vectorial singular beam shaping using homogeneous anisotropic media. *Opt. Lett.* 2010; 35(1): 7–9.
- [4] Khonina SN, Zoteeva OV, Kharitonov SI. Nonparaxial propagation of gaussian beams on the angle to the axis of the anisotropic crystal. *Computer Optics* 2012; 36(3): 346–356.
- [5] Khonina SN, Zoteeva OV, Kharitonov SI, Sharp focusing of laser beams in anisotropic uniaxial crystals. *Journal of Optical Technology* 2015; 82(4): 212–219.
- [6] Khonina SN, Kharitonov SI. Comparative investigation of nonparaxial mode propagation along the axis of uniaxial crystal. *J. Mod. Opt.* 2015; 62(2): 125–134.
- [7] Khilo NA, Petrova ES, Ryzhevich AA. Transformation of the order of Bessel beams in uniaxial crystals. *Quantum Electronics* 2001; 31(1): 85–89.
- [8] Ciattoni A, Cincotti G, Palma C. Circularly polarized beams and vortex generation in uniaxial media. *J. Opt. Soc. Am. A* 2003; 20(1): 163–171.
- [9] Marrucci L, Manzo C, Paparo D. Optical spin-to-orbital angular momentum conversion in inhomogeneous anisotropic media. *Phys. Rev. Lett* 2006; 96: 163905–163908.
- [10] Picon A, Benseny A, Mompart J, Calvo GF. Spin and orbital angular momentum propagation in anisotropic media: theory. *J. Opt* 2011; 13: 064019 (7 pp).
- [11] Khilo NA. Diffraction and order conversion of Bessel beams in uniaxial crystals. *Opt. Commun.* 2012; 285(5): 503–509.
- [12] Khonina SN, Morozov AA, Karpeev SV. Effective transformation of a zero-order Bessel beam into a second-order vortex beam using a uniaxial crystal. *Laser Phys.* 2014; 24(5): 056101 (5 pp).
- [13] Khonina SN, Karpeev SV, Morozov AA, Parinin VD. Implementation of ordinary and extraordinary beams interference by application of diffractive optical elements. *Journal of Modern Optics* 2016; 63(13): 1239–1247.



- [14] Machavariani G, Lumer Y, Moshe I, Meir A, Jackel S, Davidson N. Birefringence-induced bifocusing for selection of radially or azimuthally polarized laser modes. *Applied Optics* 2007; 46: 3304–3310.
- [15] Yonezawa K, Kozawa Y, Sato S. Compact laser with radial polarization using birefringent laser medium. *Japanese Journal of Applied Physics* 2007; 46: 5160–5163.
- [16] Hacyan S, Jáuregui R. Evolution of optical phase and polarization vortices in birefringent media. *J. Opt. A: Pure Appl. Opt.* 2009; 11(8): 085204.
- [17] Fadeyeva T, Shvedov V, Shostka N, Alexeyev C, Volyar A. Natural shaping of the cylindrically polarized beams. *Optics Letters* 2010; 35(22): 3787–3789.
- [18] Fadeyeva TA, Shvedov VG, Izdebskaya YV, Volyar AV, Brasselet E, Neshev DN, Desyatnikov AS, Krolikowski W, Kivshar YS. Spatially engineered polarization states and optical vortices in uniaxial crystals. *Optics Express* 2010; 18(10): 10848–10863.
- [19] Khonina SN, Karpeev SV, Alferov SV. Theoretical and an experimental research of polarizing transformations in uniaxial crystals for generation cylindrical vector beams of high orders. *Computer Optics* 2014; 38(2): 171–180.
- [20] Khonina SN, Karpeev SV, Alferov SV, Soifer VA. Generation of cylindrical vector beams of high orders using uniaxial crystals. *Journal of Optics* 2015; 17: 065001 (11 pp).
- [21] Ciattoni A, Palma C. Nondiffracting beams in uniaxial media propagating orthogonally to the optical axis. *Opt. Commun.* 2003; 224(4): 175–183.
- [22] Liu D, Zhou Z. Various dark hollow beams propagating in uniaxial crystals orthogonal to the optical axis. *J. Opt. A: Pure Appl. Opt.* 2008; 10(9): 095005 (9 pp).
- [23] Tang B. Hermite-cosine-Gaussian beams propagating in uniaxial crystals orthogonal to the optical axis. *J. Opt. Soc. Am. A* 2009; 26(12): 2480–2487.
- [24] Zusin DH, Maksimenka R, Filippov VV. Bessel beam transformation by anisotropic crystals. *J. Opt. Soc. Am. A* 2010; 27(8): 1828–1833.
- [25] Zhao C, Cai Y. Paraxial propagation of Lorentz and Lorentz-Gauss beams in uniaxial crystals orthogonal to the optical axis. *J. Mod. Opt* 2010; 57(5): 375–384.
- [26] Zhou G, Chen R, Chu X. Propagation of Airy beams in uniaxial crystals orthogonal to the optical axis. *Opt. Express* 2012; 20(3): 2196–2205.
- [27] Khonina SN, Paragin VD, Ustinov AV, Krasnov AP. Astigmatic transformation of Bessel beams in a uniaxial crystal. *Optica Applicata* 2016; 46(1): 5–18.
- [28] Bin Z, Zhu L. Diffraction property of an axicon in oblique illumination. *Appl. Opt.* 1998; 37(13): 2563–2568.
- [29] Khonina SN, Kotlyar VV, Soifer VA. Astigmatic Bessel laser beams. *J. Mod. Opt.* 2004; 51(5): 677–686.
- [30] Bendersky A, Quintian FP, Rebollo MA. Modification of the structure of Bessel beams under oblique incidence. *J. Mod. Opt.* 2008; 55(15): 2449–2456.
- [31] Anguiano-Morales M. Transformation of Bessel beams by means of a cylindrical lens. *Appl. Opt.* 2009; 48(25): 4826–4831.
- [32] Khonina SN, Kharitonov SI. Analogue of Rayleigh-Sommerfeld integral for anisotropic and gyrotropic media. *Computer Optics* 2012; 36(2): 172–182.
- [33] Khonina SN, Kharitonov SI. An analog of the Rayleigh-Sommerfeld integral for anisotropic and gyrotropic media. *Journal of Modern Optics* 2012; 60(10): 814–822.
- [34] Turunen J, Vasara A, Friberg AT. Holographic generation of diffraction-free beams. *J. Appl. Opt* 1988; 27(19): 3959–3962.
- [35] Khonina SN, Kotlyar VV. Bessel-mode formers. *Proceedings of SPIE* 1994; 2363: 184–190.
- [36] Kotlyar VV, Khonina SN, Soifer VA. Algorithm for the generation of non-diffracting Bessel modes. *Journal of Modern Optics* 1995; 42(6): 1231–1239.
- [37] Chattrapiban N, Rogers EA, Cofield D, Hill WT, Roy R. Generation of nondiffracting Bessel beams by use of a spatial light modulator. *Opt. Lett* 2003; 28(22): 2183–2185.

# Raman spectra analysis of human blood protein fractions using the projection on latent structures method

A.A. Lykina<sup>1</sup>, D.N. Artemyev<sup>1</sup>, I.A. Bratchenko<sup>1</sup>, Yu.A. Khristoforova<sup>1</sup>, O.O. Myakinin<sup>1</sup>,  
T.P. Kuzmina<sup>2</sup>, I.L. Davydkin<sup>2</sup>, V.P. Zakharov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Samara State Medical University, 89 St. Chapayevskaya, 443099, Samara, Russia

---

## Abstract

This work is devoted to the study of human blood protein fractions by Raman spectroscopy. Whole blood and blood plasma were used as the tested samples. For the pure Raman spectra analysis the autofluorescence background was subtracted by using of two mathematical approaches: polynomial approximation and baseline correction with asymmetric least squares. The study allowed for revealing the differences between the spectral features of blood plasma and whole blood plasma, which are changes in the relative Raman intensities of plasma and whole blood and appearance Raman bands  $670\text{ cm}^{-1}$ ,  $750\text{ cm}^{-1}$ ,  $1120\text{ cm}^{-1}$  and  $1550\text{ cm}^{-1}$  correspond to hemoglobin bonds in whole blood. The spectral features were used for total protein concentration measurement of plasma and whole blood. PLS regression method was utilized for spectral data analysis with different protein concentrations. The VIP-scores make it possible to determine the most informative spectral bands:  $1002\text{ cm}^{-1}$ ,  $1227\text{ cm}^{-1}$ ,  $1400\text{ cm}^{-1}$ ,  $1630\text{ cm}^{-1}$  for proteins analysis.

*Keywords:* whole blood; blood plasma; Raman spectroscopy; Projection on Latent Structures

---

## 1. Introduction

Proteins are high molecular weight polypeptides consisting of amino acids and are part of all human body fluids. Changes in the blood protein amount, as well as the certain fractions quantity, allows for drawing a conclusion about the human body state and pathology presence, also it helps to define the treatment efficiency [1].

Various spectral methods allow for obtaining an individual spectral “fingerprint” of the tested sample chemical compounds and these techniques are used for qualitative evaluation of blood protein. Raman spectroscopy (RS) is one of the most sensitive optical methods [2-6], as this approach has been used for various blood proteins analysis [3, 4]. The aim of this study was to compare Raman spectra of plasma and whole blood (mixture of plasma and formed elements, such as erythrocytes) [6]. The application of two mathematical approaches (polynomial approximation and baseline with asymmetric least squares) was demonstrated for Raman signal separation from the autofluorescence background. Regression model of the Projection on Latent Structures (PLS) method was constructed for the determination of total protein concentration by the analysis of plasma and whole blood Raman spectra. Variable importance in projection-scores allow for determining the most informative spectral bands.

## 2. Material and methods

### 2.1. Experimental setup

Raman spectra were collected by setup including thermally stabilized diode laser LML-785.0RB-04 (785 nm, 200 mW), commercial Raman probes (Inphotonics RPB785), and spectrograph Shamrock SR-500i-D1-R with deeply cooled digital camera Andor iDus DU416A-LDC-DD (air-cooled up to  $-70\text{ }^{\circ}\text{C}$ ). Detailed information about the experimental setup presented in paper [7].

All spectra were recorded in 780-950 nm spectral range, the exposure time was 60 seconds. The recording of three spectra for each studied sample was performed sequentially. The total time of Raman spectra registration was 3 minutes.

### 2.2. Tested samples preparation and spectra registration

The standardized collection of whole blood samples from patients with pathological blood disease was performed. The whole blood samples were obtained from the biochemical laboratory of the Samara State Medical University. Blood plasma was produced by sedimentation of whole blood in a test-tube at  $+2\text{ }^{\circ}\text{C}$  up to the complete drop-out of the formed elements to the bottom of the tube. After the blood plasma sample has been studied, it was mixed with the formed elements for the subsequent analysis of the whole blood. Altogether we performed our study for 45 samples. The tested samples were placed in the aluminum cuvette with a volume of 0.9 ml. Choice of the cuvette geometry was made based on our previous study results

[7]. The cuvette had a cylindrical shape with a flat bottom, the cuvette depth was 45 mm, the hole diameter was  $\varnothing$  5 mm. The chosen cuvette geometry provides the increase of “light” volume due to the laser radiation reflection from the side walls surface.

### 2.3. Data processing methods

PLS method was used for the experimental data analysis [8], as this method interprets the results based on a smaller number of bilinear components.

The registered signal includes the autofluorescence and Raman components, so a raw spectrum preprocessing was performed for the autofluorescence background removal. Two methods of the Raman spectrum extraction were utilized: polynomial approximation method and baseline correction with asymmetric least squares based on procedures implemented in the TPTcloud (<https://tptcloud.com/>) cloud service. Baseline correction for background component removal was performed using the method of asymmetric least squares (baseline als) [9]. Another approach utilized in this study for Raman spectra separation from the autofluorescence signal is the polynomial approximation [10].

Spectral informative bands during the regression model construction were determined by the analysis of the variable importance in projection (VIP) [11]. The higher the VIP-score of an individual variable corresponds to the more significant values in the constructed model. Variables with a low VIP-score are less important, and may be regarded as candidates for exclusion from the model.

## 3. Results and discussion

### 3.1. Raw spectra processing

The raw spectra of plasma and whole blood were registered for the study of protein fractions. The common bands of protein fractions were obtained on the basis of two mathematical approaches: polynomial approximation and method of asymmetric least squares. Analysis of the plasma and whole blood spectra allows for detection of the differences in the proteins component composition. Fig. 1 demonstrates Raman bands of blood plasma 820  $\text{cm}^{-1}$  (vibrations of tyrosine), 950  $\text{cm}^{-1}$  (deformation vibrations of CH group), 1002  $\text{cm}^{-1}$  and 1080  $\text{cm}^{-1}$  (vibrations of phenylalanine), 1160  $\text{cm}^{-1}$  (deformation vibrations of CC group), 1250  $\text{cm}^{-1}$  ( $\alpha$ -helix in Amide III), 1330  $\text{cm}^{-1}$  (vibrations of tryptophan), 1450  $\text{cm}^{-1}$  (deformation vibrations of  $-\text{CH}_2$  group), 1650  $\text{cm}^{-1}$  ( $\beta$ -helix in Amide I) [12]. Fig. 2 shows Raman spectra of whole blood. The bands are similar to Raman bands of plasma excluding 570  $\text{cm}^{-1}$  (deformation vibrations of  $\text{FeO}_2$  group), 670  $\text{cm}^{-1}$  and 750  $\text{cm}^{-1}$  (vibrations of pyrrole), 1120  $\text{cm}^{-1}$  (deformation vibrations of C-N group), 1227  $\text{cm}^{-1}$  (deformation vibrations of CH group), 1550  $\text{cm}^{-1}$  (vibrations of phenylalanine) [13]. Each processed spectrum was a subject to the multivariate analysis for the construction of regression model. The spectra of plasma and whole blood processed by two approaches (baseline als and polynomial approximation) and normalized using the standard deviation are shown in Fig. 1 (b) and Fig. 2 (b).

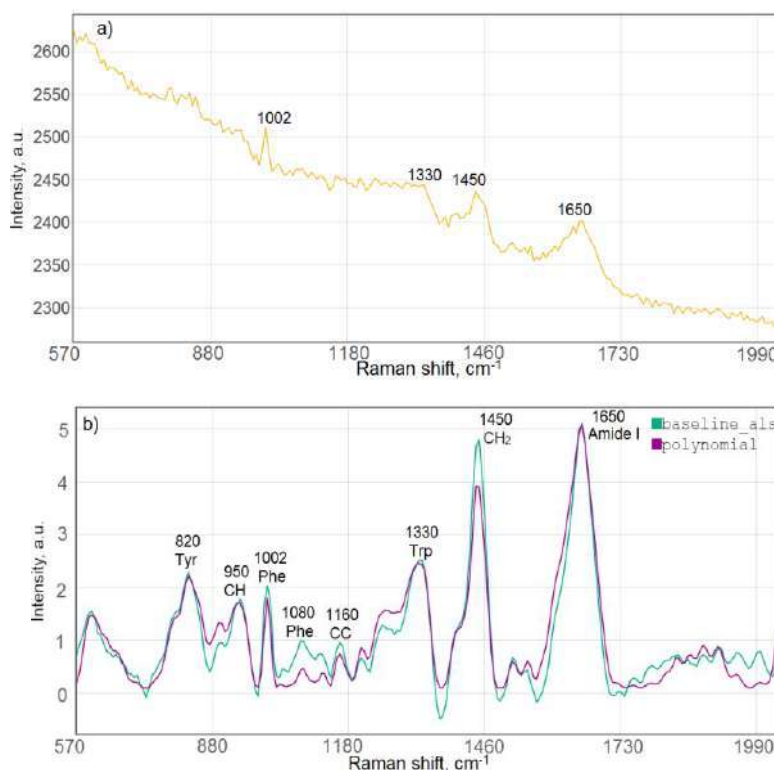


Fig. 1. Raman spectra of blood plasma processed by methods of baseline als, polynomial approximation (Phe- phenylalanine, Trp- tryptophan, Tyr-tyrosin) a) raw spectrum b) pure Raman spectrum.

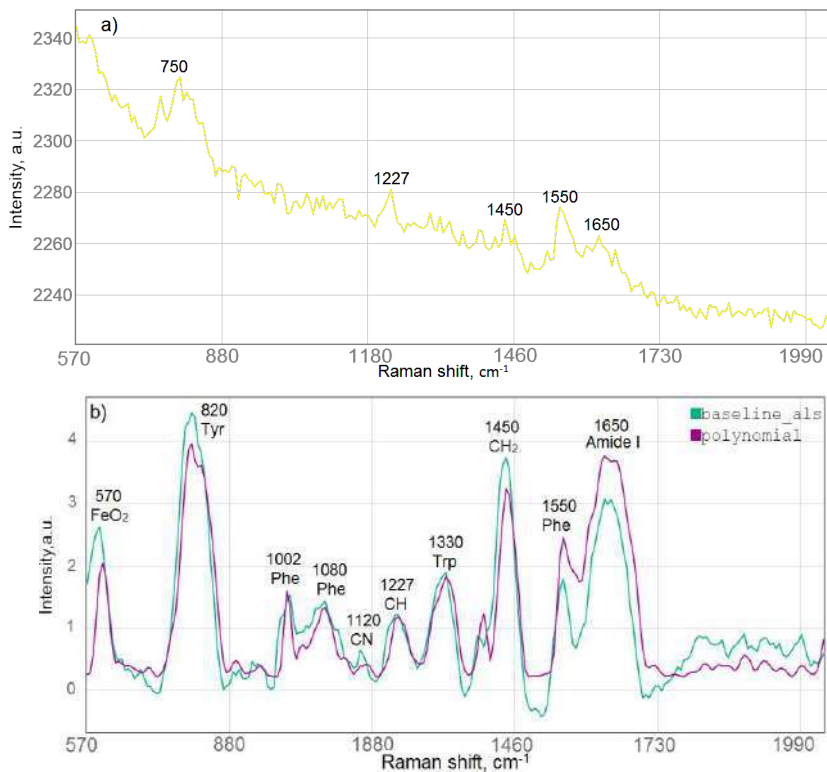


Fig. 2. Raman spectra of whole blood processed by methods of baseline als, polynomial approximation (Phe- phenylalanine, Trp- tryptophan, Tyr-tyrosin)  
a) raw spectrum b) pure Raman spectrum.

As shown in Fig.1 utilization of baseline als and polynomial approximation provides the possibility to observe Raman bands corresponding to the contribution of certain molecular vibrations. Analysis of Fig. 1 (b) demonstrate that the shape and intensity of Raman peaks in the blood plasma spectrum processed by baseline als and polynomial approximation are coincide in spectral ranges: 980- 1100  $\text{cm}^{-1}$ , 1150-1190  $\text{cm}^{-1}$ , 1640-1680  $\text{cm}^{-1}$ . For blood plasma samples the intensity of 1080  $\text{cm}^{-1}$ , 1160  $\text{cm}^{-1}$  and 1450  $\text{cm}^{-1}$  Raman bands is 15-20% higher for processing by baseline als algorithm unlike using the polynomial approximation. Analysis of 1700-2000  $\text{cm}^{-1}$  spectral region was not performed, since shape of spectra in this region is mostly associated with contribution from the optical filtering module.

Analysis of Fig. 2 helps to conclude, that raw spectra of whole blood processed by two methods Raman bands become more informative due to the elimination of autofluorescence and the appearance of characteristic bands that are hardly recognizable on a raw spectra. Positions of whole blood Raman bands coincide on the entire spectral range. Herewith, the maximum intensity difference on the 1650  $\text{cm}^{-1}$  band does not exceed 25%.

### 3.2. Raman spectra of blood plasma and whole blood

To compare the Raman bands of blood plasma and whole blood a data normalization using standard normal variate (snv) method was performed. Fig. 3 shows the normalized averaged spectra of blood plasma and whole blood for all 45 tested samples.

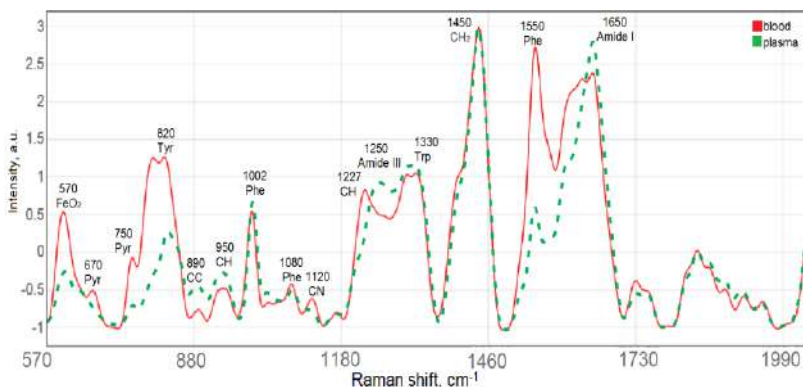


Fig. 3. Normalized averaged pure Raman spectra of blood plasma and whole blood processed by polynomial approximation method (Pyr- pyrrole, Phe- phenylalanine, Trp- tryptophan, Tyr-tyrosin).

Fig. 3 demonstrates that the Raman peaks intensities of blood plasma and whole blood coincide at 1002  $\text{cm}^{-1}$  and 1450  $\text{cm}^{-1}$  bands. The common band on 1002  $\text{cm}^{-1}$  is phenylalanine, which corresponds to a protein amino acid, the precursor of all nutrients [14]. The Raman peak on 1450  $\text{cm}^{-1}$  (deformation vibrations of  $-\text{CH}_2$  group) is present in both spectra of blood

plasma and whole blood. On the spectral ranges of  $570\text{--}800\text{ cm}^{-1}$  and  $1460\text{--}1600\text{ cm}^{-1}$ , figure shows that intensity of the Raman spectra of whole blood is higher than the intensities for blood plasma. This fact is caused by hemoglobin presence in the whole blood [14]. This is an iron protein, present in erythrocytes [15]. Hemoglobin common Raman peaks are  $570\text{ cm}^{-1}$ ,  $820\text{ cm}^{-1}$ ,  $670\text{ cm}^{-1}$ ,  $750\text{ cm}^{-1}$ ,  $1227\text{ cm}^{-1}$  and  $1550\text{ cm}^{-1}$ . The peak on  $570\text{ cm}^{-1}$  corresponds to the deformation vibrations of  $\text{FeO}_2$  group of hemoglobin. Pyrrole, one of the hemoglobin components have strong Raman peaks at  $670\text{ cm}^{-1}$  and  $750\text{ cm}^{-1}$  bands. The  $1227\text{ cm}^{-1}$  band corresponds to deformation vibrations of CH group of hemoglobin [16]. The Raman bands on  $820\text{ cm}^{-1}$  and  $1550\text{ cm}^{-1}$  correspond to the vibration of tyrosine and phenylalanine, which contribute to the composition of the protein components of blood plasma and whole blood. Since these amino acids are parts of hemoglobin, the intensity of whole blood tyrosine and phenylalanine Raman bands is 70-75% above than their Raman intensities in the blood plasma.

### 3.3. PLS analysis of Raman spectra of blood plasma and whole blood

The VIP-scores of Raman spectra matrix of the plasma and whole blood samples for the constructed regression model of the total protein concentration prediction are shown in Fig. 4.

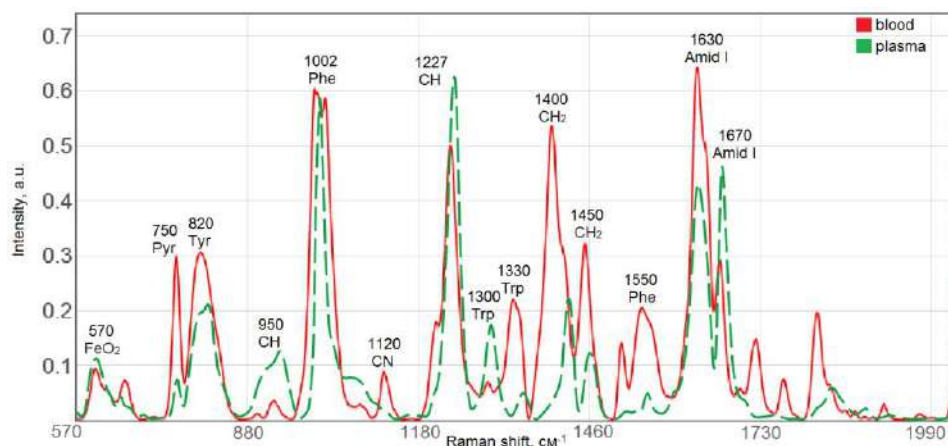


Fig. 4. VIP-scores for PLS multivariate statistical model (Pyr- pyrrole, Phe- phenylalanine, Trp- tryptophan, Tyr-tyrosin).

Fig. 4 demonstrates that the most Raman peaks of plasma and whole blood spectra are coincide on the full spectral range. The most informative spectral bands for whole blood are  $570\text{--}700\text{ cm}^{-1}$ ,  $1120\text{ cm}^{-1}$ ,  $1550\text{ cm}^{-1}$ ; and these peaks are not observed in the blood plasma Raman spectra. These bands are associated with hemoglobin groups. The Raman spectra of plasma and whole blood include multiple peaks in  $970\text{--}1040\text{ cm}^{-1}$ ,  $1370\text{--}1500\text{ cm}^{-1}$  and  $1580\text{--}1710\text{ cm}^{-1}$  bands, and these peaks are mixed in single peaks, observed in registered spectra:  $1002\text{ cm}^{-1}$ ,  $1450\text{ cm}^{-1}$  and  $1650\text{ cm}^{-1}$ . In our study the VIP distribution allows for evaluation of the Raman spectra. It doubles the spectral band, herewith increasing the Raman peaks informativeness. Analysis of obtained results makes it possible to draw a conclusion that the peaks of VIP distribution for the constructed regression model of the total protein concentration predict coincide with Raman peaks shown in Fig.3. Herewith differences are observed in the intensity amplitude of the spectral bands. As shown in Fig. 4 the largest values of VIP-scores corresponds to peaks on spectral bands of  $1002\text{ cm}^{-1}$ ,  $1227\text{ cm}^{-1}$ ,  $1440\text{ cm}^{-1}$  and  $1630\text{ cm}^{-1}$ . The chosen bands correspond to albumin and globulin [17, 18], whose concentration predominates in plasma and whole blood.

## 4. Conclusion

The current study demonstrates analysis of the plasma and whole blood Raman spectra obtained by two mathematical approaches: polynomial approximation and asymmetric least squares. The maximum differences in the Raman bands intensities did not exceed 20-25% for both approaches.

The carried out research allowed for differences detection between blood plasma and whole blood Raman spectra. The main differences in the spectral characteristics of the tested samples are observed in  $670\text{ cm}^{-1}$ ,  $750\text{ cm}^{-1}$ ,  $1120\text{ cm}^{-1}$  and  $1550\text{ cm}^{-1}$  bands. These bands are associated with hemoglobin bonds, such as pyrrole and  $\text{FeO}_2$  vibrations.

The VIP-scores calculation makes it possible to define the most informative spectral bands for total proteins analysis  $1002\text{ cm}^{-1}$ ,  $1227\text{ cm}^{-1}$ ,  $1400\text{ cm}^{-1}$ ,  $1630\text{ cm}^{-1}$  corresponding to albumin and globulin fractions.

## Acknowledgments

This research was supported by the Ministry of Education and Science of the Russian Federation.

## References

- [1] Hanlon EB, Manoharan R, Koo TW, Motz JT, Fitzmaurice M, Kramer JR, Itzkan I, Dasar RR, Feld MS. Prospects for in vivo Raman spectroscopy. Luxembourg: Phys. Med. Biol. 2000; 45(2): 59.
- [2] Premasiriti WR, Lee JC, Ziegler LD. Surface-Enhanced Raman Scattering of Whole Human Blood, Blood Plasma, and Red Blood Cells: Cellular Processes and Bioanalytical Sensing. Luxembourg: J. Phys. Chem. B. 2012; 116(31): 9376.

- [3] Artemyev DN, Bratchenko IA, Khristoforova JA, Lykina AA, Myakinin OO, Kuzmina TP, Zakharov VP, Davydkin I.L. Blood proteins analysis by Raman spectroscopy method. *Izbrannye Trudy*. – Luxembourg: Proceedings of SPIE-The International Society for Optical Engineering 2016; 98887:1Y–1.
- [4] Annika MK, Tae-Woong E, Oh J, Hunter M. Blood analysis by Raman spectroscopy. *United States: Optics letters* 2004; 27: 2004.
- [5] Dingari NC, Horowitz GL, Kang JW, Dasari RR, Barman I. Raman Spectroscopy Provides a Powerful Diagnostic Tool for Accurate Determination of Albumin Glycation. *Francisco: PLoS ONE* 2012; 7: 2.
- [6] Castiglioni C, Tommasini M, Zerbi G. Raman spectroscopy of polyconjugated molecules and materials: confinement effect in one and two dimensions. *United States: Philos. Trans. A Math. Phys Eng. Sci*, 2004; 1824: 2469.
- [7] Lykina AA, Artemyev DN, Bratchenko IA. Analysis of albumin Raman scattering registration efficiency from different volume and shape cuvette. *United States: JBPE* 2017; 2: 3157.
- [8] Esbensen KH. *Multivariate Data Analysis*. New Jersey: In Practice 4-th Ed. 2000.
- [9] Eilers PHC, Boelens HFM. *Baseline Correction with Asymmetric Least Squares Smoothing*. United States: Leiden University Medical Centre 2005.
- [10] Zeng H, Lui H, McLean DI. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *United States: Applied spectroscopy* 2007; 61: 1225.
- [11] Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *United States: Journal of Chemometrics* 2015; 10: 528.
- [12] Regula A, Majzner K, Marzes KM, Kaczor A, Pilarczyk M, Baranska M. Raman spectroscopy of proteins: a review. *United States: J. Raman Spectroscopy* 2017; 44: 1061.
- [13] Atkins CG, Buckley K, Blades MW, Turner RFB. *Raman Spectroscopy of Blood and Blood Components*. United States: *Applied spectroscopy* 2017; 5: 767.
- [14] Gelder JDe, Gussem KDe, Vandenabeele P, Moens L. Reference database of Raman spectra of biological molecules. *United States: J. Raman Spectroscopy* 2007; 9: 1133.
- [15] Das TK, Counture M, Ouellet Y, Guertin M, Rousseau DL. Simultaneous observation of the O—O and Fe—O<sub>2</sub> stretching modes in oxyhemoglobins. *United States: PNAS* 2011; 5: 479.
- [16] Casella M, Lucotti A, Tommasini M, Zerbi G. Raman and SERS recognition of  $\beta$ -carotene and haemoglobin fingerprints in human whole blood. *United States: Spectrochimica Acta Part A* 2011; 5: 915.
- [17] Uzunbajakava N, Lenfereink A, Kraan Y, Willekens B, Greve J, Otto C. Nonresonant Raman Imaging of Protein Distribution in Single Human Cells. *Luxembourg: Biopolymers* 2003; 72: 1.
- [18] Artemyev DN, Zakharov VP, Davydkin IL, Khristoforova JA, Lykina AA, Konyukhov VN, Kuzmina TP. Measurement of human serum albumin concentration using Raman spectroscopy setup. *Luxembourg: Opt. Quant Electron* 2016; 48: 337.

# Intelligent learning and testing system for students training in the problem area of nanotechnology and microsystem engineering

D. Lyapunov<sup>1,4</sup>, A. Yankovskaya<sup>1,2,3,5</sup>, Y. Dementyev<sup>1</sup>, K. Negodin<sup>1</sup>

<sup>1</sup>National Research Tomsk Polytechnic University, 30, Lenin Ave., 634050, Tomsk, Russia

<sup>2</sup>National Research Tomsk State University, 36, Lenin Ave., 634050, Tomsk, Russia

<sup>3</sup>Tomsk State University of Control Systems and Radioelectronics, 40, Lenin Ave., 634050, Tomsk, Russia

<sup>4</sup>Research Institute of Automation and Electromechanics, 53, Belinskiy St., 634034, Tomsk, Russia

<sup>5</sup>Tomsk State University of Architecture and Building, 2, Solyanaya Sq., 634003, Tomsk, Russia

---

## Abstract

For the students learning and training in the field of nanotechnology and microsystem engineering within the framework of blended learning paradigm we need high quality content; efficient learning technology; means of students motivation; evaluation tools. We propose an intelligent learning and testing system based on mixed diagnostic tests for effective comprehension of a number of subjects within the problem area. The system allows to provide effective comprehension of a number of subjects and to form the primary competences from the students point of view, revealing their future occupation preferences. During the learning process the students within small groups (not less than 4 students) solve the problems of development, modelling and design of microsystem devices. They also investigate the market needs, consider the opportunities of macroscopic sensors and actuators exchange on their microsystem analogs.

*Keywords:* Intelligent learning and testing system; nanotechnology; microsystem engineering; learning technology; blended learning; mixed diagnostic test; competences; multidisciplinary course

---

## 1. Introduction

A relevance of intelligent learning and testing systems (ILTS) design for students training in different problem areas is well acknowledged [1]. Currently the need of ILTS in such interdisciplinary field as nanotechnology and microsystem engineering is justified by including the field in the list of critical technologies of the Russian Federation [2]. Domestic industrial enterprises need highly qualified specialists in the problem area. The need is caused by the rapid development of nanotechnologies and the progress of new metamaterials fabrication. The materials under consideration can possess unique mechanical, electrical, magnetic, thermal and optical properties [3].

Current trends, challenges and news in the problem area are published in the Journal of Nano- and Microsystem Engineering [4] and in a number of foreign journals.

The specialty education on the program “Nanotechnology and Microsystem Engineering” is provided in such Russia’s privileged institutions as Moscow State Technical University of Radioengineering, Electronics and Automation, Kazan Federal University, Saint-Petersburg State Polytechnic University, Tomsk State University of Control Systems and Radioelectronics, Siberian Federal University and others.

We should note that the problem area “Nanotechnology and Microsystem Engineering” is a multidisciplinary one. Therefore, the learner should possess the competences in the following disciplines: chemistry, physics, material science, electrical engineering, electronics, thermal engineering, optics, mathematics, and others. These competences are essential for research of nano- and microscale phenomena taking into account the latest technological achievements.

For effective learning and training in the problem area the students need: high quality online content; effective learning technologies; means of motivation; cognitive graphic tools (CGT) for learning outcomes evaluation and timely feedback at each state of learning process.

We propose an intelligent learning and testing system (ILTS) based on mixed diagnostic tests (MDT), aimed at effective comprehension of a number of disciplines in the problem area of nanotechnology and microsystem engineering. Moreover, the learner will be able to construct an individual learning trajectory, to enhance his/her strengths, and fill the knowledge gaps.

## 2. Peculiarities of the problem area “Nanotechnology and Microsystem Engineering”

Before we go any further, we consider the inherent features of the problem area. Most essential of them are:

- 1) Rapid development of nanotechnologies, microsystem components modelling and design methods.
- 2) Specialists orientation primarily in scientific research.
- 3) Lack of standards on some nanotechnology materials and products.
- 4) Microsystem components miniaturization trend (Moor’s law is still valid).

To increase the industrial release of nanotechnology products we should:

- 1) provide advanced training of the students and the specialists in the problem area;
- 2) design specialized equipment for nanomaterials and microsystem products fabrication;
- 3) fulfill a database on modern materials and microsystem constructions;
- 4) design the testing devices for checking industrially released microsystems.

Semi-empirical engineering methods of modelling and design, borrowed from the adjacent disciplines (material science, electrical engineering, electronics, thermodynamics, and others), should be supplemented by new methods of microsystem devices modelling and design [5]. Using the methods the future specialists will be able to construct promising microsystem devices, taking into account detailed structural analysis of the object under development, regularities of its response, prediction of its functional characteristics.

Students learning endeavor in any field of interest, especially when we deal with such field as nanotechnology, is primarily connected with their motivation to comprehend the learning course completely. The need of students' motivation is described in detail in [6,7].

A goal setting performed by the student contributes greatly the motivation issues. An effective goal setting is proved to be the guaranteed comprehension of the discipline under study, thus moving towards the goal [6]. Unfortunately, the learning outcomes declared in the syllabi of the majority of disciplines do not correspond to the students personal goals and preferences. Therefore, at the initial stage of the learning process we must provide students with high quality lecture material along with online content, aimed to increase students interest. It can help to correct students personal goals at the early stage of the learning process. Thus, the students will be motivated to implement systematic and progressive steps to master the discipline under study.

We develop the ILTS for the advanced mastering the discipline taking into account the individual features and needs of each learner and the peculiarities of the problem area of nanotechnology and microsystem engineering.

### 3. Basic Terms and Definitions

Herein we introduce some terms and definitions used in the future chapters of the paper.

A learner is a human who is learning, e.g. the University student.

A teacher is a human who is managing and/or supporting the learning process.

Diagnostic test (DT) is a set of features differentiating any pairs of objects, which belong to different patterns.

Unconditional diagnostic test (UDT) is characterized by simultaneous presentation of all the intrinsic features (questions) of the object under study during the decision-making process.

Conditional diagnostic test (CDT) is a test, in which each subsequent feature (question) depends on the previous features (questions) represented to a learner.

Mixed diagnostic test being a compromise between unconditional and conditional components [8].

### 4. Basics of Intelligent Learning and Testing Intelligent System Construction

We have been developing the ILTS since late 2010th [1,8,9]. An evolution of knowledge acquired during a learning course, represented by a semantic web, was introduced in the paper [9]. We demonstrate the block diagram of the ILTS in Fig. 1.

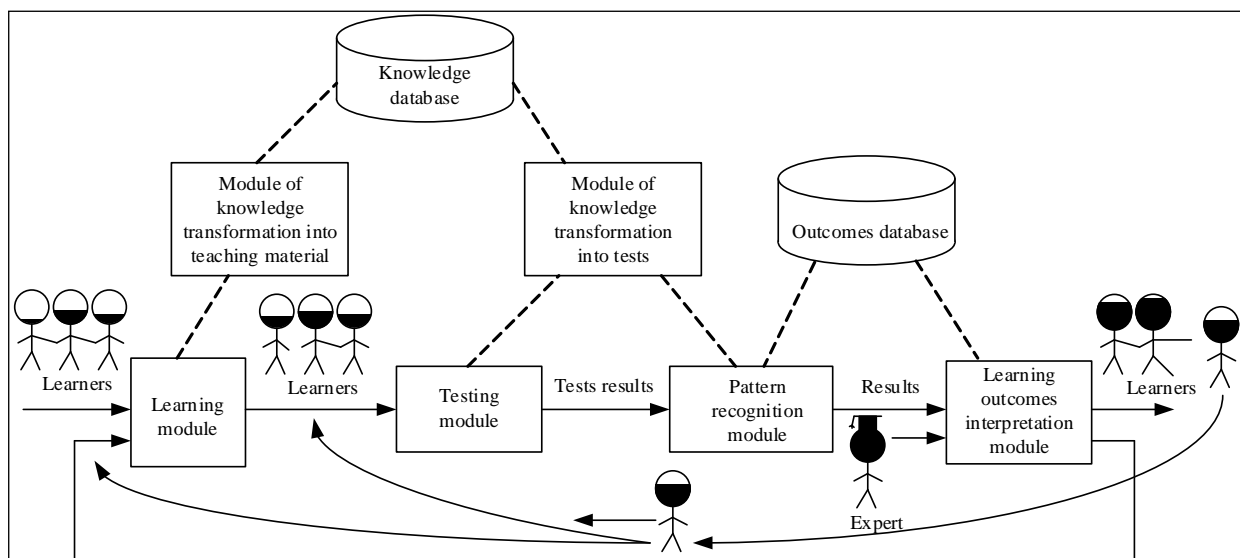


Fig. 1. A block diagram of the intelligent learning and testing system.

The algorithm of ILTS we subdivided into 7 subsequent steps.

Step 1. During a learning process a student comprehends the learning materials subsequently. The learning materials on a particular topic are represented by a text with an interactive multimedia content. We use the semantic model of knowledge representation to store the data while learning [9]. Learning module is responsible for this step.

Step 2. The ILTS acquires and stores the results of UDT. Steps 2-4 are performed via testing module.

Step 3. If the learner succeeded in the UDT, ILTS switches him/her to CDT. In this case the system defines each subsequent question depending on the answer on the previous one.



Step 4. The results of all MDT components (UDT and CDT) are recorded in the results database. We use high level of detalization to get additional data for the learning course improvement in the future. The ILTS generates Learner Action Card (LAC) for each student.

Step 5. After each learning module the LAC is converted into a set of evaluation indicators, such as: a) theory knowledge level; b) problems solving skills; c) laboratory work performance evaluation. Thus, the ILTS suggests the future learning trajectory for the student using pattern recognition module.

Step 6. After finishing the learning course the ILTS represents to each student the evaluated knowledge, LAC interpretation and evaluation indicators calculated. Students compare the learning materials comprehended with knowledge and competences acquired during the course. They reveal their knowledge gaps and a lack of skills by concurrent consideration of the LAC and the semantic web of the learning course. If there is some revealed knowledge gap or a lack of competence in the problem area, the student has an option to return to the step 1 and to practice learning materials and tests one more time. In this case the learner is involved in the decision-making process, based on analyzing of the learning outcomes via cognitive graphic tools (CGT) and revealing the knowledge gaps and the future challenges. As a result, the student constructs his/her individual learning trajectory. The present step is provided by learning interpretation module.

Step 7. When the testing procedure is finished, the learner is considered to comprehend the entire learning course as well as interdisciplinary connection between the course modules. In this case the ILTS calculated the total evaluation expressing the mean value of all the tests results.

We organized the sequence of the learning material representation so as each subsequent learning module is connected with the previous one. Each learning module is designed in accordance with a reflection cycle [10], represented in Fig. 2 by a directed cyclic diagram.

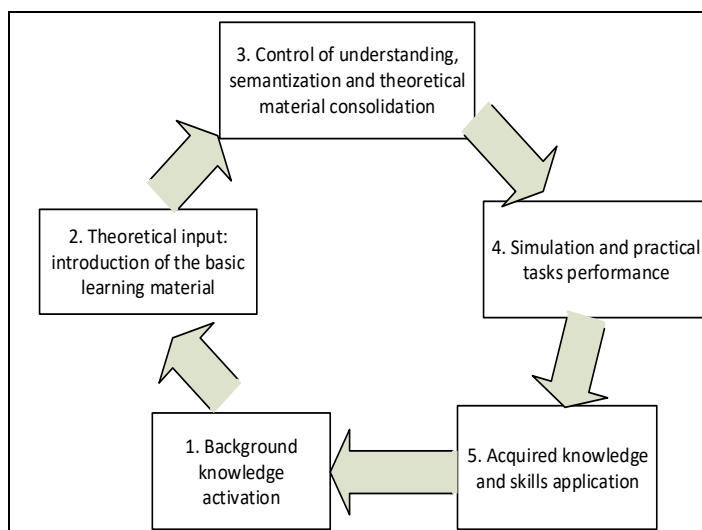


Fig. 2. Reflection cycle.

## 5. Learning Outcomes Interpretation

In this chapter we use some fragments from the papers [11,12], wherein the usage of 2-simplex prism CGT is entirely described for students' learning outcomes interpretation. We have used the advantages of the 2-simplex prism for students performance evaluation during the course "Selected Chapters of Electronics", which includes the module "Microelectromechanical Systems". An example of individual learning trajectory construction based on the MDT and 2-simplex prism CGT is shown in Fig. 3.

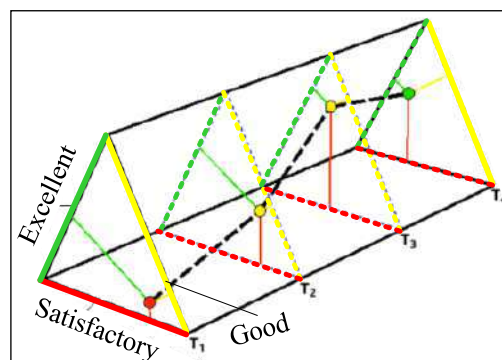


Fig. 3. Learning outcomes evaluation using 2-simplex prism cognitive graphic tool.

The results of each of the four tests are represented as points (the small circles of different colors) within 2-simplexes (cross-sections of the 2-simplex prism). Prism's faces correspond to evaluation indicators (grades): 1) "excellent"; 2) "good"; 3) "satisfactory". The distance from the base of the 2-simplex prism to the 2-simplex (equilateral triangle) under consideration

corresponds to the time interval from the beginning of the learning process to the time of the corresponding testing. The dashed line within the 2-simplex prism shows the knowledge level evolution of the student based on the test results at the time moments T1, T2, T3 and T4. We note that illustrations presented in Fig. 3 and 4 were obtained using visualization library, which is currently under development and is available via the link [13].

We observe in Fig. 3 that the student had got a grade close to “satisfactory” at the input testing at time moment T1. Although the grade is between “satisfactory” and “good”. Then, having analyzed the test result, the student had set a goal to increase his/her evaluation indicators. During the 2nd testing the learner demonstrated better result (the grade is close to “good”) at time moment T2. Then, at time moment T3 the testing result corresponds to the intermediate grade between “good” and “excellent”. And, finally, at time moment T4 the student achieved the goal, which he/she had set on the previous stages of the learning process (the grade is close to “excellent”).

Desired area of development of the learner is chosen based on the MDT results represented by CGT 2-simplex. In Fig. 4 we represent the 2-simplex CGT. The patterns in the form of points (the circles of small radii of different color) correspond to competences of the four students after finishing the learning course.

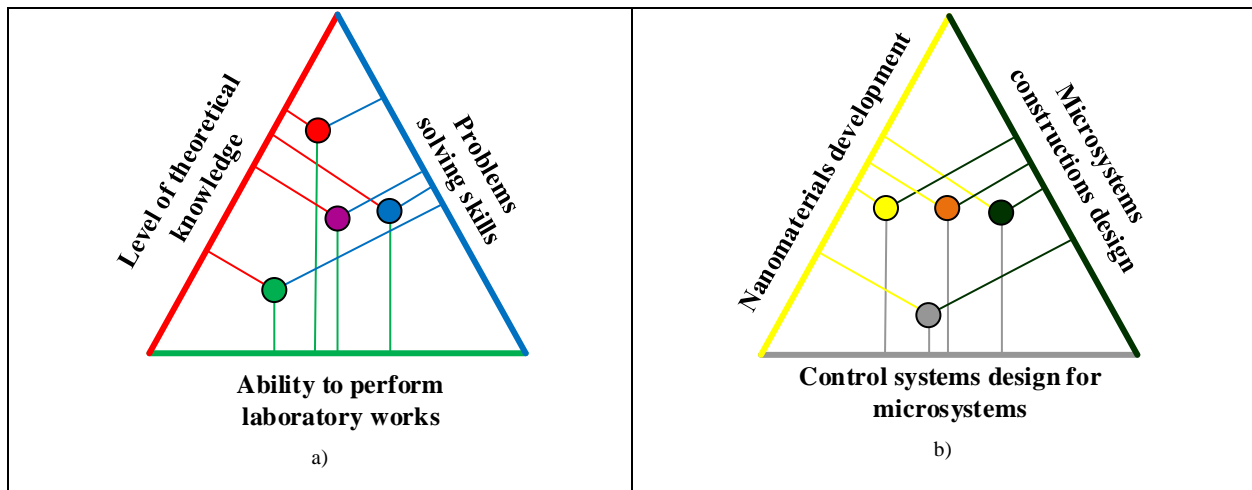


Fig. 4. 2-simplexes indicating the competences of the four students after finishing the learning course: a) according to knowledge, skills and abilities; b) according to individual aptitude to future occupation.

Each point in Fig. 4 corresponds to the balance state between the competences gained (level of knowledge, problems solving skills and ability to perform laboratory works) of the one learner. Consider the perpendicular connecting the point under study (e.g. the green one) and the side of the equilateral triangle (e.g. the side, which relates to the ability to perform laboratory works). The length of the perpendicular (green line) characterizes the students' ability under consideration. The smaller is this length, the higher is the student's evaluation of the corresponding competence. The green point (see Fig. 4, a) corresponds to testing results of the student, who gained the laboratory works performance ability in the best way, whose theoretical knowledge is quite good, but the problems solving skills are lacking. The red point characterizes the student, who knows the theory and gained problems solving skills at sufficient level, but whose ability to perform the laboratory works is not sufficient. The blue point expresses the testing results of the student who copes with problem solving in the best way, but needs improving in other areas. The violet point corresponds to the testing results of the student who achieved a balance between the competences under consideration.

Considered competences evaluations are the basis for further goal setting, taking into account all strong points and weaknesses of a particular student. The ILTS is aimed to reveal the gaps in competences and to propose the actions, which will help to eliminate them in the future.

We illustrate the testing results of the four students in Fig. 4, b. The test was constructed to reveal individual aptitude towards future occupation. Each point shown in Fig 4, b demonstrates the aptitude to the future occupation for each of four students. The yellow point corresponds to the testing result of the student, who is apt to conduct research in the field of material science to get the new materials. The dark green point corresponds to the learning outcomes of the student, who is good in microsystems constructions design. The grey point corresponds to the aptitude of microsystem devices control systems design. Finally the orange point corresponds to the student who achieved a balance between the components and have wide range of aptitudes.

Test results analysis gives an information, which is used to construct a learning trajectory. It facilitates the competence improvement in such areas as: 1) scientific research aimed at new technologies development of microsystems fabrication; 2) practical activities in the field of microsystems design; 3) teaching in the nanotechnology and microsystem engineering problem area.

## 6. Conclusion

Intelligent Learning and Testing System (ILTS) proposed is designed for the learning process efficacy increase in such multidisciplinary fields as nanotechnology and microsystem engineering, and also in the data analysis research. Learning outcomes analysis will allow to effectively design the learning courses. The courses under development are based on the mixed diagnostic tests. By using the ILTS we take into account individual peculiarities of the students during their learning activities,

reveal different types of regularities in the learning process, correspond the learner's individual preferences and learning outcomes. The ILTS provides a balanced learning material representation on every stage of learning process.

We use the semantic technologies to compare the learning material represented and the knowledge acquired for each learner providing filling the knowledge gaps if necessary. The feedback at every stage of learning process in the form of semantic web provides timely correction and repetition of uncomprehended material. In addition, the student reveals his/her field of competence, i.e. the most "sweet" course modules, thus, constructing the individual learning trajectory.

To the present time, a number of modules on the topic "Microsystem Engineering" were implemented into the e-learning course "Selected Chapters of Electronics". The learning outcomes of the students, who participated in the course, give grounds to believe that using the ILTS proposed we are able to motivate students to achieve the goals set by them in the early stages of the learning process. We propose the use of ILTS for bachelor training program "Nanotechnology and Microsystem Engineering" for disciplines: "Introduction to Nanotechnology", "Microsystem Technology", "Microelectromechanical Systems", "Digital Control Systems" and others.

## Acknowledgements

The research was funded by RFBR grant (project No. 16-07-00859a). The authors are grateful to the junior research fellow, assistant of the Tomsk State University of Control Systems and Radioelectronics, Artem V. Yamshanov for the cognitive graphic tools software design and implementation for the data analysis. We are also grateful to the 3<sup>rd</sup> year students (group 5G4B) major in Electric Power Engineering who participated in the learning course "Selected Chapters of Electronics" for their patience, persistence and desire to learn.

## References

- [1] Yankovskaya AE, Semenov ME, Yamshanov AV, Semenov DE. Cognitive means in learning and testing systems based on mixed diagnostic tests. *Artificial Intelligence and Decision Making* 2015; 4: 51–61.
- [2] A list of critical technologies of the Russian Federation. URL: <http://www.kremlin.ru/supplement/988> (01.02.2017).
- [3] Galochkin V. Introduction to nanotechnology and nanoelectronics. Samara: Povolzhsky State University of Telecommunications and Informatics, 2013; 367 p. (in Russian).
- [4] Interdisciplinary, scientific, technique and production journal "Nano- and Microsystems Technology". URL: <http://www.microsystems.ru/> (02.02.2017).
- [5] Lyapunov D. MEMS capacitance converters: Constructions, materials, research issues. Saarbrücken : LAP LAMBERT Academic Publishing GmbH & Co. KG, 2012; 139 p. (in Russian).
- [6] Dirksen J. Design for how people learn. Berkeley: New Readers, 2012; 272 p.
- [7] LeFever L. The Art of Explanation. New Jersey: Wiley and Sons, 2012; 378 p.
- [8] Yankovskaya AE. Design of Optimal Mixed Diagnostic Test With Reference to the Problems of Evolutionary Computation. Proceedings of the 1st International Conference on Evolutionary Computation and Its Applications, Moscow, 1996; 292–297.
- [9] Yankovskaya AE, Shurygin YuA, Yamshanov AV, Krivdyuk NM. Determination of the level of acquired knowledge on the training course presented by the semantic network. Proceedings of International Conference "Open Semantic Technologies for Intelligent Systems" (OSTIS-2015), Minsk, 2015; 331–338.
- [10] Plekhanova MV, Prohorets EK. Modeling of electronic courses based on the reflective cycle. *International Journal of Applied and Fundamental Research* 2015; 5: 600–604.
- [11] Yankovskaya AE, Yamshanov AV, Krivdyuk NM. 2-simplex Prism – a Cognitive Tool for Decision Making and Its Justification in Intelligent Dynamical Systems. *Machine Learning and Data Analysis* 2015; 1(14): 1930–1938.
- [12] Yankovskaya AE, Demytyev YuN, Lyapunov DY, Yamshanov AV. Intelligent Information Technology in Education. *Information Technologies in Science, Management, Social Sphere and Medicine (ITSMSSM-2016)*. Atlantis Press Publishing 2016; 17–21.
- [13] 2-simplex Prism Cognitive Graphic Tool. URL: <http://cogntool.tsuab.ru/250demos/2-simplex-prediction/> (02.02.2017).

# Sensitive detection of Nitrogen Dioxide using gold nanoparticles decorated Single Walled Carbon Nanotubes

Sunil Kumar<sup>1</sup>, Vladimir Pavelyev<sup>1</sup>, Prabhash Mishra<sup>1</sup>, Nishant Tripathi<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The modification of carbon nanotubes (CNTs) could enhance their surface and electric properties. To this purpose, we explore the impact of a thin layer of gold (Au) on the surface of single wall carbon nanotubes (SWCNTs). SWCNTs have been grown by Chemical Vapor Deposition (CVD) method and decorated with gold nanoparticles were investigated as gas sensitive materials for detecting nitrogen dioxide (NO<sub>2</sub>) at room temperature. Surface morphology and microstructure of Au-SWCNT have been characterized by FE-SEM and Raman Spectroscopy. Using the present collective approaches, the improvement in the detection of NO<sub>2</sub> gas using Au-modified nanotubes is explained. However, Au-modified SWCNT gas sensors exhibited better performances compared to pristine SWCNTs. These changes in resistance and the shift of the Fermi level just after NO<sub>2</sub> exposure was probably due to adsorption of NO<sub>2</sub> molecules on the surface of Au-SWCNTs. Surface modification of nanotubes with understanding of sensing ability at atomic level opens the new way to design a selectivity gas sensor.

*Keywords:* carbon nanotubes; nanostructured materials; nanotechnology; functionalization; sensitivity; stability

---

## 1. Introduction

The main feature of individual SWNT sensors, besides their small size is that they operate at room temperature with higher sensitivity. SWNTs possess several properties that are very essential for gas sensors. They have all their atoms on the surface, endowing them with the highest specific surface area possible together with graphene. Therefore, all the carbon atoms in the nanotube can, in principle, interact with the analytic gas, while simultaneously supporting charge transport in the device. Thus, adsorbates and electrostatic charges and dipoles close to the nanotube can greatly impact charge transport. At the same time, the carbon nanotube lattice is held together by strong sp<sup>2</sup> C-C bonds, which provide the necessary chemical stability to the carbon nanotube. An individual SWNTs sensor can be used to detect different types of molecules [1].

Detecting gas molecules is basic to environmental monitoring [2], control on chemical processing [3], space mission [4], agricultural and medical applications [1]. This type of device is very important because there are many gases which are harmful to organic life, such as humans and animals. One of the gases to be verified is nitrogen dioxide (NO<sub>2</sub>). Even in small concentrations, it irritates the respiratory tract in large concentration causes pulmonary edema. NO<sub>2</sub> create disturbance mainly in the airways and lungs, but also causes changes in blood composition, in particular, reduces the content of haemoglobin in blood. At low concentration of only 0.23 mg/m<sup>3</sup>, one feels the presence of this gas, but its adverse effects observed in healthy individuals at concentrations of NO<sub>2</sub> in all 0.56 mg/m<sup>3</sup>, which is four times lower than the detection threshold. People with chronic lung diseases experience difficulty in breathing even at a concentration of 0.38 mg/m<sup>3</sup>. Among all harmful gasses, NO<sub>2</sub> is a well-known toxic gas and air pollutant and monitoring its concentration is crucial for air quality monitoring. Prolonged exposure to low concentration of NO<sub>2</sub> capable of causing several health hazards such as coronary artery disease as well as stroke [5-6]. The sensitivity of SWNTs towards NO<sub>2</sub> at atmospheric temperature as reported [1] is particularly interesting. The sensing of NO<sub>2</sub> is important to monitor environmental pollution resulting from combustion or automotive emission [7-8]. In recent times, the accidents in the oil, coal, gas industries has been increases, which claim the lives of hundreds of people. Every year many people lose their life due to hazardous gas leakage [1-4].

Many research groups have discussed sensing mechanism of NO<sub>2</sub> based on CNT. In order to improves the sensing performance, and more challengingly, how to improve sensitivity of sensor for different gas species. One promising way is the functionalization [4, 9] of carbon nanotubes. Many characteristics of CNTs are superior to most other materials. Thus, for example, Young's modulus, which depends on the diameter and chirality of a CNT defect, can reach 1.8TPa, while when the conventional carbon fibres, it is comparable to 800GPa. The bulk compressibility of CNTs is quite high and amounts to 0.024GPa<sup>-1</sup>. If bent CNT also exhibit exceptional flexibility, their electrical conductivity depends on the magnetic field induction [10]. The magnetic properties of CNTs are remarkably different from the properties of diamond and graphite. The first measurements of the magnetic susceptibility showed that it greatly decreases with decreasing temperature of 300K. CNTs exhibit anisotropy magnetic property. With these properties, CNTs have broad application prospects, but their successful use is necessary to deal with some problems [2, 4, 10]. For example CNT through the possession of large surface energy, tend to form agglomerates, reaching up to tens or hundreds of micrometres. This leads to deterioration of the properties of CNTs in comparison with those that would be typical for homogeneous distribution. Solution to this problem can be achieved using various methods. CNT mechanical processing time must also be limited; since it increases the density of surface defects is increased [11]. Therefore, in addition to mechanical processing methods use the chemical treating CNTs to achieve more efficient dispersibility and impart additional properties. For example, using metal catalysts in the form of nanoparticles to decorate CNT, promotes the interaction with specific gas species. In this experiment CNTs have been functionalize by gold decoration.

Existing gas sensors are based on metal oxide semiconductor. However they have a low sensitivity, high operating temperature and reaction time and substantial recovery. To ensure effective monitoring of air quality status it is necessary to improve the characteristics of gas sensors that can detect danger in advance. Development of NO<sub>2</sub> sensors based on carbon nanotubes due to their unique properties will provide an opportunity to find a solution to these critical problems. To increase the sensitivity and selectivity to specific gas, as well as their reliability in various condition. The extraordinary property of SWNTs towards NO<sub>2</sub> sensing attracts not only academicians but also industrials to make low power NO<sub>2</sub> gas sensor. In present work, we are trying to solve above mentioned problems, for same, SWNTs grown sample is decorated with gold nanoparticles and also we have done detailed study on various effect of Au decoration on sensor characteristics.

## 2. Experiment

SWNTs used in this sensor have been grown by standard Chemical vapor Deposition (CVD) technique [10,13, 14]. CVD technique is one of the best technology for CNTs growth on silicon wafers.

We grow SWNTs on 5X5 mm chromium coated silicon wafer by standard CVD method [10, 13, 15]. Deposited SWNTs are decorated by gold. Gold is coated over sample by sputtering system. After that two electrodes are made by standard lithography technique as shown in Fig.1.

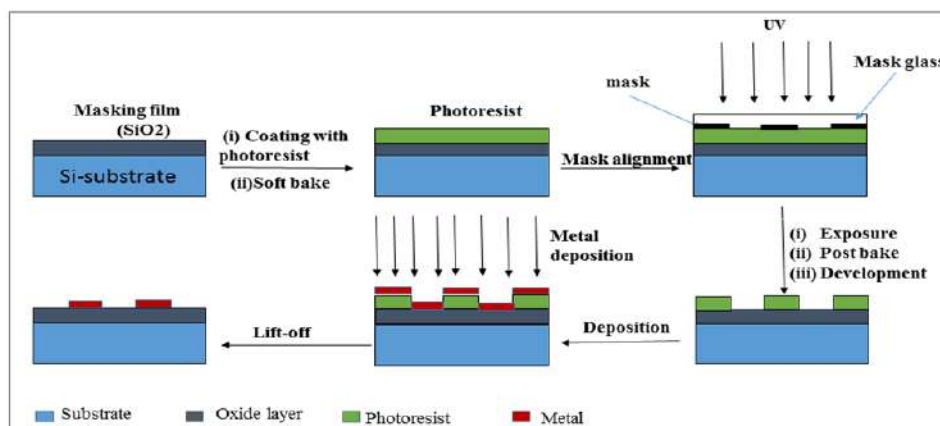


Fig.1. Phases of lithographical process.

Formation of sensor electrode has the following successive processes: lithography, deposition and etching. Typical lithography process includes a set of operations that can be divided into three phases (Fig.1):

- Forming a continuous uniform layer of resist on the substrate surface;
- Once the surface has been coated with photoresist, the substrate is exposed to UV light;
- Once exposed, the substrate is immersed in a developer solution.

NO<sub>2</sub> sensor research work carried out based on a CNT in a special chamber with one side connected to the gas distributor and on the other with the release into the environment. The gas supply comes from the two cylinders: the first bottle contains only air, the second air cylinder + NO<sub>2</sub> concentration of 100ppm. The camera also has outputs for connection of an oscilloscope, multimeter that allows you to measure the change in resistance of the sensor in real time. Restoring the sensor is carried out by exposure to UV radiation. The flow of UV rays sent directly to the camera cell by limiting their distribution area. The calculation and measurement of the concentration of nitrogen dioxide (NO<sub>2</sub>) to obtain the experimental data; regulation of the inlet gas concentration is done by standard mass flow controller and change in the resistance is measured by using multimeter/oscilloscope.

## 3. Results and discussion

Figure 2 shows the scanning electron microscopy (FESEM) image of pristine SWNTs grown over silicon substrate, in which we clearly observe a dense horizontal network of SWNTs over all substrate. The present SWNTs on substrate also verify by Raman spectroscopy (Fig.4). In Raman spectra, a sharp peak in the range of 200cm<sup>-1</sup> to 300cm<sup>-1</sup> is verifying the existence of SWNTs on silicon substrate. Figure 3 shows the FESEM image of Au decorated SWNTs surface, where we can see non-uniform particles of gold is distributed on every CNT. First we had done sensing experiment without UV supported recovery. And we found that the recovery time is more than 12 hours, which is impractical and does not meet all the tasks to be performed by the sensor. For the functional operation of the sensor it is necessary to its full recovery after each cycle of gas exposure. To expedite this process, we need to give the adsorbed gas molecules enough energy to break chemical bonds and their desorption from the surface of the CNTs. To achieve such an effect is possible by heating or exposing the sensor with UV exposure. Exposure to UV light is more advantageous way compared with heating, since, firstly, quantum energy UV radiation allows strong enough to destroy the chemical bonds, thereby accelerating the desorption process several times; Second, importantly, the use of UV lamps easier to operate [15-20]. After that we performed a series of experiment to monitor the response of the sensor with different concentrations of NO<sub>2</sub>, followed by reduction by means of UV radiation (see Fig.5 to Fig. 6). To see the various

effect of gold decoration on sensing property firstly we perform the sensing experiment on pristine SWNTs with the concentration of NO<sub>2</sub> is 40ppm level and we found initial resistance Ri = 65.06KΩ. After the start of gas supply to the resistance test chamber starts to decrease gradually. The response of the sensor is a 1 ~ 3 seconds. After 5 minutes the gas flow was stopped, the camera only did the air flow and also produces ultraviolet light. Almost immediate increase in resistance was noted. Full recovery of the sensor to the initial position was 4 minutes 30 seconds. Now same experiment was repeated for gold decorated SWNTs sample with kept all sensing parameter same as before.

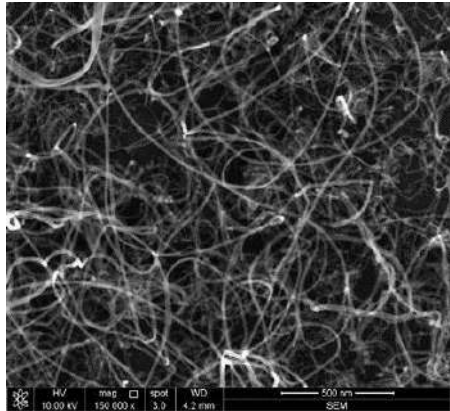


Fig. 2. FESEM image of pristine SWNTs.

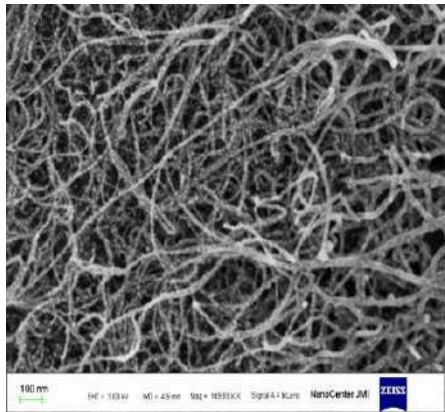


Fig. 3. FESEM image of gold decorated SWNTs.

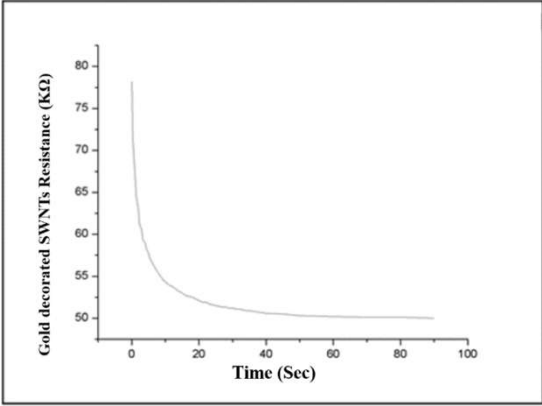


Fig. 4. Raman spectra of pristine SWNTs.

It is observed that initial resistance Ri = 78.2KΩ sharply decreased to drag reduction occurred more rapidly than without gold sample. Comparison between NO<sub>2</sub> gas sensor without gold coated and with gold coated has been shown in the Fig.6 The sensitivity for each case can be calculated by formula:

$$S = \frac{R_0 - R_{NO_2}}{R_0} \times 100\% \tag{1}$$

Where -S is the sensitivity of the sensor; R<sub>0</sub> is the sensor resistance before you start working; R<sub>NO<sub>2</sub></sub> is the resistance of the sensor at the end of the experiment. And we found sensor sensitivity for pristine type sensor approximately 30percent and for gold decorated sensor around 38 percent approximately. The comparison between both type of sensor also shown in Fig. 6 and it is clearly observe from figure that gold decorated sensor have better sensitivity as compare to pristine SWNTs sensor. The possible reason for better sensitivity is that gold decorated CNTs have larger surface area as compare to pristine CNTs and hence the area for gas molecules interaction with sample is also larger.



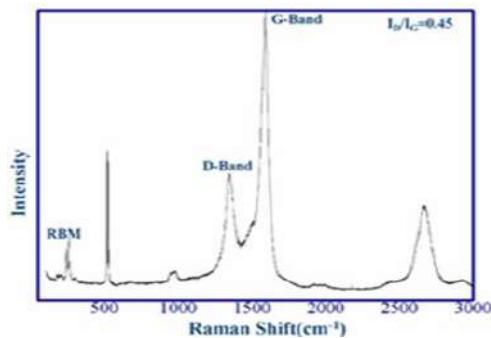


Fig. 5. Resistance Vs time for Au decorated SWNTs sensor.

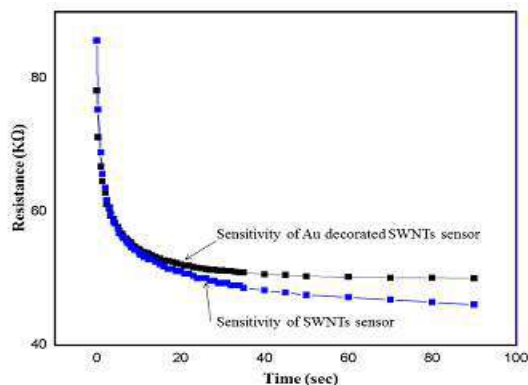


Fig. 6. Plot of the comparison between response of SWNT NO<sub>2</sub> sensor and Au-modified SWNT NO<sub>2</sub> sensor. Au-modified SWNT NO<sub>2</sub> sensor showed increase in sensitivity as compared with without gold coated SWNTs NO<sub>2</sub> sensor.

#### 4. Conclusion

The conclusion of our work is that we successfully developed a good quality NO<sub>2</sub> sensor. Based on the results of observations it can be argued that the sensor has an almost instantaneous reaction rate to the feed gas and the selected recovery technique using UV radiation has advantages over previous technologies by small time and ease of use. The change in sensitivity of SWNT sensor is induced by the coating of Au layer. The chemical pattern clearly demonstrates a significantly higher sensitivity of the Au-modified SWNT sensor compared with the un-functionalized SWNT sensor for NO<sub>2</sub> gas.

#### Reference

- [1] Kong J. Nanotube Molecular Wires as Chemical Sensors. *Science* 2000; 287(5453): 622–625.
- [2] Zanolli Z, Leghrib R, Felten A, Pireaux J-J, Llobet E, Charlier J-C. Gas sensing with au-decorated carbon nanotubes. *ACS Nano* 2011; 5(6).
- [3] Zeng Q, Luna J, Bayazitoglu Y, Wilson K, Imam MA, BarreraEV. Metal Coated Functionalized Single-Walled Carbon Nanotubes for Composites Application. *Mater. Sci. Forum* 2007; 561–565: 655–658.
- [4] Sun Y-P, Fu K, Lin Y, Huang W. Functionalized Carbon Nanotubes: Properties and Applications. *Acc. Chem. Res.* 2002; 35(12): 1096–1104.
- [5] Vallejos S, Gracia I, Chmela O, Figueras E, Hubálek J, Cané C. Chemoresistive micromachined gas sensors based on functionalized metal oxide nanowires: Performance and reliability. *Sensors Actuators B Chem.* 2016; 235: 525–534.
- [6] Suehiro J, Zhou G, Hara M. Fabrication of a carbon nanotube-based gas sensor using dielectrophoresis and its application for ammonia detection by impedance spectroscopy. *J. Phys. D. Appl. Phys.* 2003; 36(21): L109–L114.
- [7] Tran TH, Lee J-W, Lee K, Lee YD, Ju B-K. The gas sensing properties of single-walled carbon nanotubes deposited on an aminosilane monolayer. *Sensors Actuators B Chem.* 2008; 129(1): 67–71.
- [8] Penza M, Rossi R, Alvisi M, Signore MA, Cassano G, Dimaio D, Pentassuglia R, Piscopiello E, Serra E, Falconieri M. Characterization of metal-modified and vertically-aligned carbon nanotube films for functionally enhanced gas sensor applications. *Thin Solid Films* 2009; 517(22).
- [9] Derycke V, Auvray S, Borghetti J, Chung C-L, Lefèvre R, Lopez-Bezanilla A, Nguyen K, Robert G, Schmidt G, Anghel C, Chimot N, Lyonnais S, Streiff S, Campidelli S, Chenevier P, Filoramo A, Goffman MF, Goux-Capes L, Latil S, Blase X, Triozon F, Roche S, Bourgoin J-P. Carbon nanotube chemistry and assembly for electronic devices. *Comptes Rendus Phys.* 2009; 10(4): 330–347.
- [10] Mishra P, Pavelyev VS, Patel R, Islam SS. Resistive sensing of gaseous nitrogen dioxide using a dispersion of single-walled carbon nanotubes in an ionic liquid. *Mater. Res. Bull.* 2016; 78: 53–57.
- [11] Meng L, Fu C, Lu Q. Advanced technology for functionalization of carbon nanotubes. *Prog. Nat. Sci.* 2009; 19(7): 801–810.
- [12] Lee K, Lee J-W, Dong K-Y, Ju B-K. Gas sensing properties of single-wall carbon nanotubes dispersed with dimethylformamide. *Sensors Actuators B Chem.* 2008; 135(1): 214–218.
- [13] Tripathi N, Mishra P, Joshi B, Islam SS. Precise control over physical characteristics of Carbon Nanotubes by differential variation of Argon flow rate during Chemical Vapor Deposition processing: A systematic study on growth kinetics. *Mater. Sci. Semicond. Process* 2015; 35: 207–215.
- [14] Tripathi N, Mishra P, Joshi B, Islam SS. Catalyst free, excellent quality and narrow diameter of CNT growth on Al<sub>2</sub>O<sub>3</sub> by a thermal CVD technique. *Phys. E Low-dimensional Syst. Nanostructures* 2017; 62: 43–47.
- [15] Penza M, Rossi R, Alvisi M, Cassano G, Serra E. Functional characterization of carbon nanotube networked films functionalized with tuned loading of Au nanoclusters for gas sensing applications. *Sensors Actuators, B Chem.* 2009; 140(1).
- [16] Brahim S, Colbern S, Gump R, Grigorian L. Tailoring gas sensing properties of carbon nanotubes. *J. Appl. Phys.* 2008; 104(2): 24502.
- [17] Mishra P, Harsh, Islam SS. Trace level ammonia sensing by SWCNTs (network/film) based resistive sensor using a simple approach in sensor development and design. *Int. Nano Lett.* 2013; 3(1): 46.
- [18] Peng N, Zhang Q, Chow CL, Tan OK, Marzari N. Sensing mechanisms for carbon nanotube based NH<sub>3</sub> gas detection. *Nano Lett.* 2009; 9(4): 1626–1630.
- [19] Huang XJ, Choi YK. Chemical sensors based on nanostructured materials. 2007; 122(2): 659–671.
- [20] Van PTH, Thanh NH, Van Quang V, Van Duy N, Hoa ND, Van Hieu N. Scalable fabrication of high-performance NO<sub>2</sub> gas sensors based on tungsten oxide nanowires by on-chip growth and RuO<sub>2</sub>-functionalization. *ACS Appl. Mater. Interfaces* 2014; 6(15): 12022–12030.

# Physicochemical properties of submicron and nanoscale particles of Ga and AlGa alloy obtained by laser ablation in a liquid

V.S. Kazakevich<sup>1</sup>, P.V. Kazakevich<sup>1</sup>, P.S. Yaresko<sup>1</sup>, D.A. Kamynina<sup>1,2</sup>

<sup>1</sup>Samara branch of P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Novo-Sadovaya 221, 443011, Samara, Russia

<sup>2</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

Optical absorption spectra of gallium nanoparticles synthesized by laser ablation in 2-propanol, tetrahydrofuran, ethyl alcohol, liquid nitrogen and argon were obtained. A shift of the maximum and broadening of the absorption band of Ga, Al nanoparticles and AlGa alloy due to fast aggregation during the substitution cryogenic liquid in a colloid on the liquid at room temperature process were detected. When the AlGa nanoparticles were moved from the liquid argon medium to distilled water, a chemical reaction with the evolution of gaseous hydrogen was observed. The dependence of the evolved gas volume on the percentage ratio of metals in the AlGa film obtained by the vacuum deposition method was constructed. In the case of laser ablation of Ga in ethyl alcohol, the formation of gallium core / shell nanoparticles was fixed.

*Keywords:* laser ablation; nanoparticles; optical absorption spectra; gallium; thin films; cryogenic liquid; hydrogen

---

## 1. Introduction

At the present time, interest in gallium is due to its special chemical and physical properties, such as a strong tendency to form a Ga-Ga bond in solids and molecules, a low melting point of 303 K (29.8 °C), expansion of the volume upon freezing, and also the possibility of creating on its basis numerous technologically significant alloys and compounds [1]. The transition from macroscopic dimensions to nanoparticles leads to dependence of the melting temperature on the particle size - with decreasing diameter, the melting temperature also decreases [2]. The use of these properties can lead to the creation of logical information recording elements based on interphase transitions induced by optical radiation inside the gallium nanoparticle [3]. Gallium nanoparticles are the best option in terms of the energy required to change the phase state. Also, gallium can be used in alternative energy sources, as a component of the AlGa alloy, which prevents of the aluminum oxidation in the air medium. It is known, that the interaction of aluminum with water leads to the gaseous hydrogen emission, but under atmospheric conditions such a reaction is impossible, since oxidation of the aluminum surface occurs. The use of an oxidation resistant AlGa alloy is one of the ways to solve this problem [4]. The transition to alloyed nanoparticles can lead to an increase in the useful yield of the hydrogen evolution reaction by increasing the effective area of interaction of aluminum with distilled water.

One of the methods to obtain such nanoobjects is the method of laser ablation in liquid media. Unlike chemical methods, it is possible to obtain particles with a wide size distribution and completely free of the reaction products. In the case of chemical interaction of the target with the surrounding medium, laser radiation can initiate chemical processes. Therefore, it is promising to use inert cryogenic liquids.

The change in one of the parameters of laser ablation (the source of laser radiation, the target or the liquid in which the process takes place) will affect on the final products. The formation of nanoparticles and differences in their form, size and degree of aggregation can be registered by using absorption spectra.

Therefore, the aim of this work was to determine the effect of the liquid medium in which the ablation takes place on the optical properties of gallium nanoparticles and also to consider changes in optical spectra associated with the rapid aggregation of Ga, Al nanoparticles and AlGa alloy during the substitution in a colloid of a cryogenic liquid (liquid nitrogen) to a liquid at room temperature.

## 2. Experimental technique

For the synthesis of nanoparticles, a standard scheme of the laser ablation method in liquid, supplemented with a special cuvette for working with liquid nitrogen or argon, which prevents the liquid from boiling around the target, was used [5]. The radiation of an Nd: YAG laser with a wavelength of 1064 nm, a pulse repetition rate of 20 Hz, and pulse duration of 250 ps was focused on the target surface. As liquids were used: glycerol, ethyl alcohol, 2-propanol, tetrahydrofuran, liquid nitrogen and liquid argon. The thickness of the liquid layer above the target surface was 5 mm. Surface treatment took place both in a stationary mode - laser radiation was focused at one point of the target, and in the scanning mode - a cuvette with a sample, by means of motorized tables Standa moved relative to the stationary laser beam. For the analysis of obtained particles by scanning electron microscopy, a titanium foil was placed in the cuvette with the target during irradiation to precipitate the ablation products.

The obtained colloids were analyzed by the LOMO spectrophotometer SF-56. The measurement range is 190-1100 nm, the spectral resolution is 0.3 nm. Since the design of this instrument does not provide for the analysis of cryogenic sols, in both parts of the experiment a technique for replacing cryogenic liquid in a colloid with a liquid at room temperature was used [6].

The first series of experiments consisted in obtaining colloids of gallium particles to further determine the optical absorption spectra associated with plasmon resonance. The target was a plate of gallium (99.99%) 2 mm thick. The energy of laser



radiation in the pulse was 15 mJ, and the laser fluence varied from 20 to 400 J / cm<sup>2</sup>. Irradiation was performed in a stationary mode for 30 minutes.

To compare the optical characteristics of Ga nanoparticles obtained by laser ablation with the optical characteristics of gallium particles synthesized by other methods [7, 8], glycerol, distilled water, ethyl alcohol and isopropyl alcohols were used as liquid media. However, at laser ablation in room temperature fluids the probability of formation of microaggregates from gallium nanoparticles increases. This is due to the fact that the melting temperature of the target is close enough to room temperature and the removed material does not have time to crystallize. Therefore, the next stage was the use of liquid nitrogen. In [6], differences in the formation of a liquid nitrogen colloid droplet during the overflow into cuvettes filled with different liquids were shown. Therefore, in this work, the colloid of Ga nanoparticles was divided into two equal volumes, after which one part was transferred to ethyl alcohol and the other to be compared to isopropyl alcohol.

In the second series of experiments, a thin AlGa film was ablated in liquid nitrogen and liquid argon media. The production of this film was carried out using a vacuum universal station (VUS-5), by spraying aluminum (99.99%) and gallium (99.99%) onto the surface of the slide. As vaporizers, graphite rods were chosen. On the evaporators aluminum and gallium in the proportions determined by laboratory scales Electronic balance B 2104 were placed:

- 99% Al, 1% Ga
- 97% Al, 3% Ga
- 95% Al, 5% Ga
- 93% Al, 7% Ga
- 90% Al, 10% Ga

The energy of laser radiation in a single pulse was 0.3 mJ, and the laser fluence at the samples surface was 0.11 J / cm<sup>2</sup>. It was selected in such a way that the glass substrate did not break down. Irradiation occurred in the scanning mode. The treatment area was 30 mm<sup>2</sup>. To compare the optical absorption spectra of alloyed AlGa nanoparticles obtained in liquid nitrogen or liquid argon, the particles were transferred to distilled water.

In the same way, a thin aluminum film and a thin gallium film were prepared and irradiated in a liquid argon medium, followed by the replacement of argon in the colloid by H<sub>2</sub>O.

Visualization and elemental analysis of ablation products deposited on the titanium foil from the colloid were carried out using a scanning electron microscope Carl Zeiss Evo 50 equipped with a nitrogen-free energy dispersive detector X-Max 80 (EDX).

### 3. Results and discussion

Figure 1a shows the absorption spectra of gallium nanoparticles obtained by laser ablation in ethyl and isopropyl alcohols. It can be seen that the spectra have the same absorption band in the region from 262 to 280 nm with local peaks at 267 and 275 nm. This can be attributed to the fact that both liquids have practically the same density - 789 and 786 kg / m<sup>3</sup>, respectively. According to the published data [7], this parameter of the medium has a significant effect on the optical properties of metallic nanoparticles. It is important to note that the absorption lines of gallium particles synthesized in isopropyl alcohol are shifted to the long-wavelength region of the spectrum in the present paper in compare with the data obtained in [7].

The spectra of nanoparticles obtained in glycerol and water are characterized by a broad absorption band. In the case of glycerol, this may be due to the formation of aggregates of nanoparticles in a viscous medium. And the process of ablation in water is characterized by the formation of oxides. In alcohols, the absorption band is much narrower. This difference can be explained by the fact that due to the high activity of the surface of the metallic particles, they bind to the solvent molecules. This leads to decrease of the particles aggregation probability. In glycerol, the absorption band of 220-300 nm is characterized by two absorption maxima at 224 nm and 260 nm, respectively. In water, the absorption band is shifted by 245-307 nm due to oxidation and has a maximum at 272 nm.

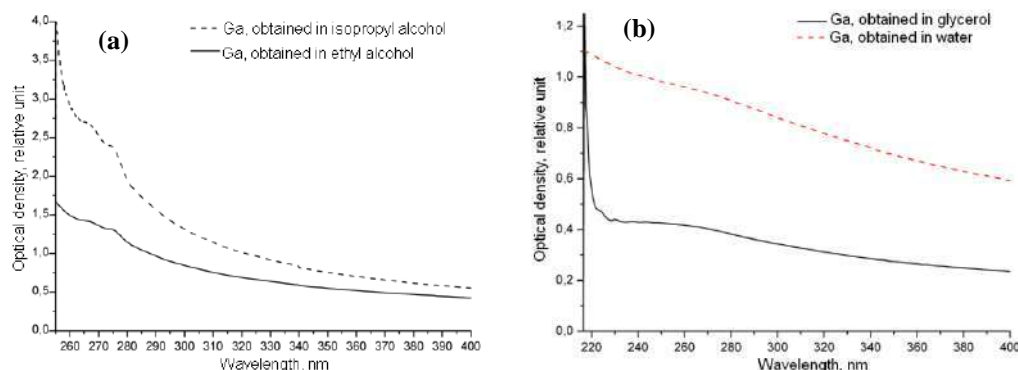


Fig.1. The absorption spectra of gallium nanoparticles obtained (a) in isopropyl and ethyl alcohol; (b) in glycerol and distilled water.

By scanning electron microscopy of gallium particles obtained in ethyl alcohol spherical structures with characteristic dimensions from 80 to 800 nm were revealed. In a number of cases, elongated shape gallium structures with a thickness of 50 to 100 nm, repeating the contours of the particles, were found on the titanium foil surface. Apparently, such structures are

fragments of the shell of nanoparticles. The formation of such structures is most often associated with the formation of bubbles at the interaction of laser radiation with the target [9, 10].



Fig.2. SEM image of gallium particles obtained by laser ablation in ethyl alcohol.

Figure 3 shows the absorption spectra of Ga nanoparticles obtained in liquid nitrogen. The colloid was divided into two volumes. In the first volume, the cryogenic liquid was replaced by isopropyl alcohol, and in the second - by ethanol. In the case of replacement with isopropyl alcohol, the absorption band falls on the interval 215-290 nm with maxima at 231 and 280 nm. And when substituted for ethanol, the absorption band (205-280 nm) and its maxima (212 and 248 nm) are shifted to the short-wavelength region of the spectrum.

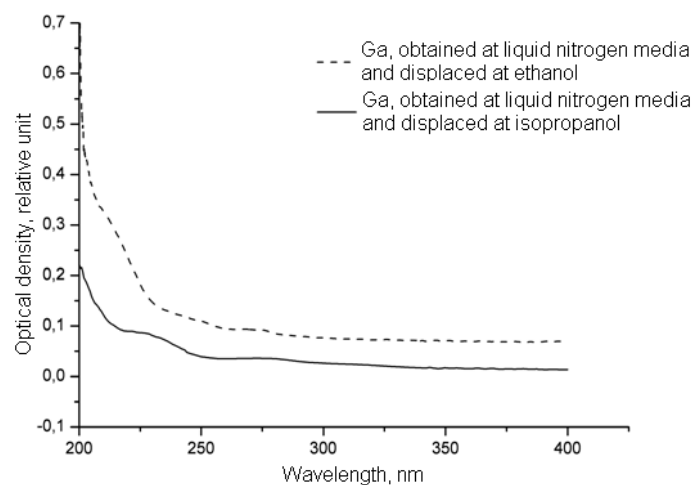


Fig.3. Absorption spectra of Ga nanoparticles obtained in liquid nitrogen and transferred to isopropyl and ethyl alcohols.

SEM - analysis of the initial target in the form of the thin AlGa film, obtained by the vacuum deposition method, is shown in Fig. 4. The film thickness was 600 nm. The evaporation temperature of Ga is 2420 °C, and the evaporation temperature of Al is 2380 °C. However, it should be noted that at the initial moment of deposition of metals on the glass substrate surface, the gallium concentration exceeds the concentration of aluminum. On the evaporation process may influence the presence on Al of an oxide film, whose evaporation temperature is 3000 °C.

A comparison of the optical absorption spectra of alloyed AlGa nanoparticles obtained in liquid nitrogen and liquid argon is presented on Figure 5. For this, cryogenic liquids in colloids were replaced by distilled water. In both cases, an absorption band from 210 to 300 nm with peaks at 224 and 267 nm is observed. The difference in spectra lies in the fact that the particles obtained in the liquid nitrogen medium have one more, broad, absorption band in the range from 300 to 700 nm.

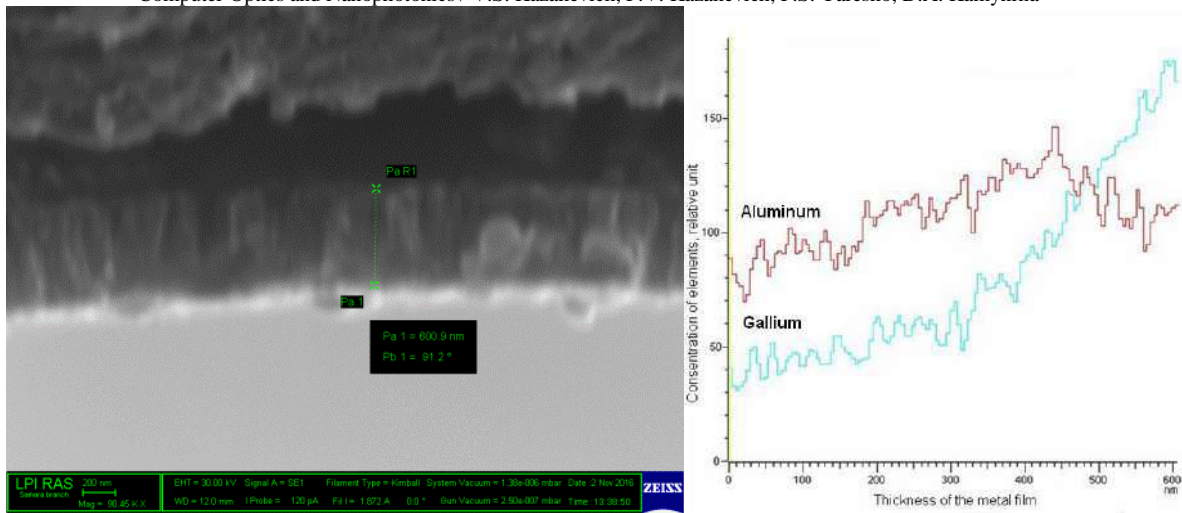


Fig.4. SEM-image and elemental analysis of the thin AlGa film by the thickness.

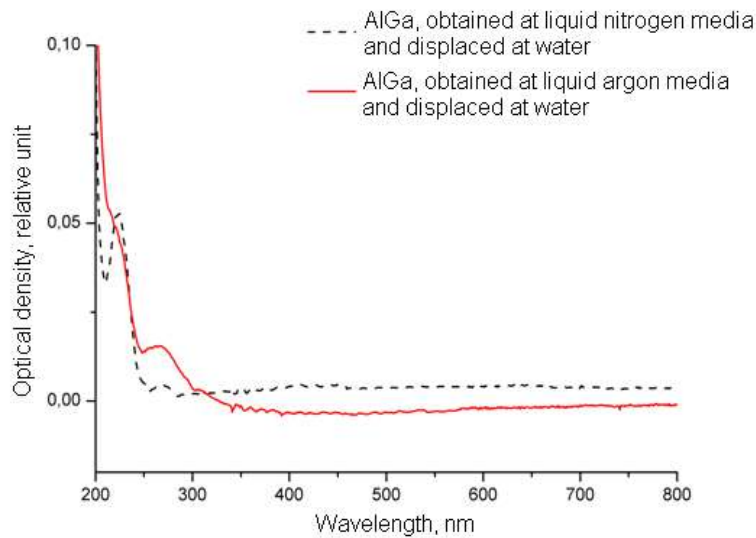


Fig.5. Absorption spectra of AlGa nanoparticles obtained in liquid nitrogen and liquid argon, replaced in water.

During the replacement of AlGa nanoparticles from liquid argon to water, active gas evolution was observed. When the open flame was brought on, a rapid ignition of the gas with a characteristic pat was occurred. Therefore, it can be argued that hydrogen gas is released as a result of the chemical reaction  $2Al + 6H_2O = 2Al(OH)_3 + 3H_2$ . Nanoparticles of Al synthesized in liquid nitrogen and transferred to water do not enter into this reaction, since, apparently, the formation of aluminum nitride occurs.

Figure 6a shows a comparison of the absorption spectra of AlGa nanoparticles obtained in liquid argon during the ablation of thin films with different percentages of metals, after being replaced by water. With a change in the composition of the film, the absorption bands and their maxima remain the same, only small changes in the optical density are observed. The maximum optical density was recorded for 95% Al and 5% Ga. The dependence of the volume of the evolved gas on the percentage of metals in the target is shown in Fig. 6b. The greatest volume of gas yield, 8 milliliters, is accounted for 95 percent of aluminum and 5 percent of gallium.

Absorption spectra of Al and Ga nanoparticles obtained in a liquid argon medium after the replacement of argon in a colloid by  $H_2O$  are shown in Figure 7. Aluminum nanoparticles have an absorption band from 210 to 250 nm, and gallium particles have an absorption band from 250 to 300 nm.

By using a scanning electron microscope, images and an elemental analysis of micron and submicron particles synthesized in liquid argon and liquid nitrogen during the ablation of thin AlGa films with followed replacement of the cryogenic liquid in the colloid on water were obtained. According to elemental analysis, the presence of both gallium and aluminum was found in the composition of nanoparticles (Fig. 8a, b). The presence of nitrogen on the spectrum in the case of ablation in liquid nitrogen indicates the possible formation of nitrides of the used metals.

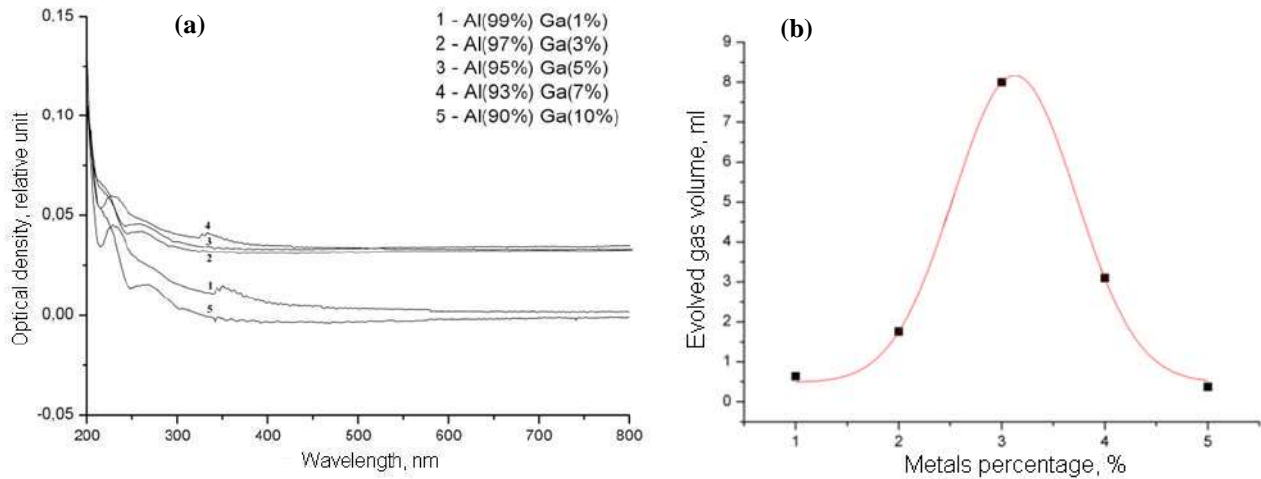


Fig.6. (a) Absorption spectra of AlGa nanoparticles obtained in liquid argon during the ablation of thin films with different percentages of metals after the replacement of argon by the water; (b) Graph of the dependence of the evolved hydrogen volume on the percentage ratio of metals in the target.

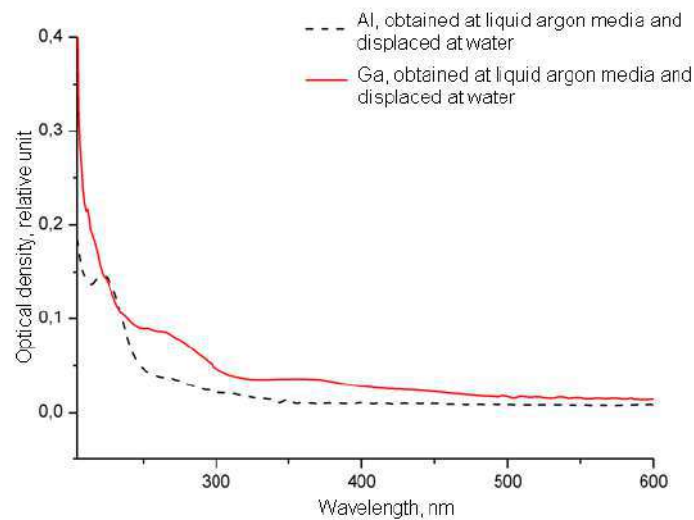


Fig.7. Optical absorption spectra of Al and Ga nanoparticles obtained in a liquid argon medium with followed argon replacement in a colloid on H2O.

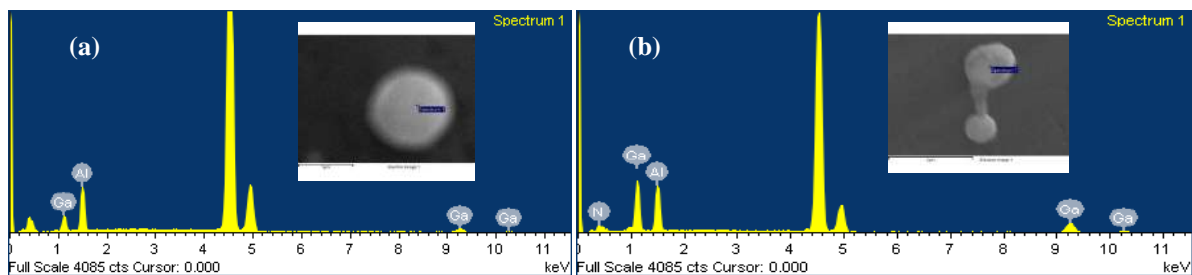


Fig.8. Data of the energy-dispersion analysis of micron and submicron nanoparticles obtained by laser ablation of AlGa target (a) in a liquid argon medium, (b) in a liquid nitrogen medium with followed replacement in a colloid of a cryogenic liquid on water.

#### 4. Conclusion

In the present work, micro- and nanoparticles Ga, Al, AlGa by laser ablation in liquid media were synthesized. Shell fragments of gallium particles were found. Optical absorption spectra of Ga nanoparticles obtained in glycerol, water, isopropyl alcohol, ethanol and liquid nitrogen are shown. In the case of liquid nitrogen, the absorption spectra of the particles were obtained after replacing the cryogenic liquid on isopropyl and ethyl alcohols. The absorption spectra of AlGa particles synthesized in liquid argon and liquid nitrogen were also obtained. Information about the optical absorption spectra of Ga nanoparticles obtained at various parameters is promising from the point of view of creating gallium logic information recording elements [3].

The technique proposed for applying thin AlGa films that are not oxidized in air and their further laser ablation in an inert cryogenic liquid can be used in the development of alternative methods for producing hydrogen. The optimal percentage of Al

and Ga in the composition of these films (19: 1) was selected, at which the maximum yield of hydrogen gas was observed after irradiation in liquid argon with following replacement on water.

## References

- [1] Hunderi O, Ryberg R. Band structure and optical properties of gallium. *Phys. F: Met. Phys.* 1974; 4: 2084–2095.
- [2] Bandin AE, Beznosyuk SA. Dependence of the nanoparticles melting temperature on its shape in terms of titanium nanoparticles. *Izvestiya of Altai State University* 2011; 3-2: 127–130.
- [3] Soares BF, Jonsson F, Zheludev NI. All-Optical Phase-Change Memory in a Single Gallium Nanoparticle. *Physical Review Letters* 2007; PRL 98: 153905. DOI: 10.1103/PhysRevLett.98.153905.
- [4] Woodall MJ, Jeffrey TZ, Charles RA. Power Generation from Solid Aluminium. United States Patent Application, 2008.
- [5] Kazakevich VS, Kazakevich PV, Yaresko PS, Kamynina DA. Laser ablation of gold in liquid argon. *Fizicheskoe Obrazovanie v VUZah* 2016; 22: 23–28.
- [6] Kazakevich VS, Kazakevich PV, Yaresko PS, Nesterov IG. Production of colloidal gold in various liquids using a laser ablation in liquid nitrogen technique. *Proceedings of the Samara Scientific Center of the Russian Academy of Sciences* 2012; 14: 268–272.
- [7] Meléndrez MF, Cárdenas G, Arbiol J. Synthesis and characterization of gallium colloidal nanoparticles. *Journal of Colloid and Interface Science* 2010; 346: 279–287.
- [8] Kang M, Saucer TW, Warren MV, Wu JH, Sun H. Surface plasmon resonances of Ga nanoparticle arrays. *Appl. Phys. Lett.* 2012; 101: 081905. DOI: 10.1063/1.4742328.
- [9] Yan ZJ, Bao RQ, Wright RN, Chrisey DB. Hollow nanoparticle generation on laser-induced cavitation bubbles via bubble interface pinning. *Appl. Phys. Lett.* 2010; 97: 124106.
- [10] Yan ZJ, Zhao Q, Chrisey DB. Structural evolution of hollow Al<sub>2</sub>O<sub>3</sub> particles formed on excimer laser-induced bubbles. *Mater. Chem. Phys.* 2011; 130: 403–408.

# Nanocrystalline Silicon and Silicon Carbide Optical Properties

Daria Lizunkova<sup>1</sup>, Natalya Latukhina<sup>1</sup>, Victor Chepurinov<sup>1</sup> and Vyacheslav Parandin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoe shosse, Samara, 443086, Russia

---

## Abstract

Porous silicon possesses a wide range of the unique properties and has good perspectives for photo-sensitive structures for a solar cells new generation. Due to the developed pore system, the area of absorbing surface increases, and also the increased sensitivity expands into the short-wavelength region due to the increased energy band gap of silicon nano-particles and silicon nano-filaments on the walls of the pores. Carbonization of the surface layer of porous silicon makes absorption even more effective. In this case, the spectrum of the solar cell expands into the short-wavelength region due to absorption of the high-energy photons in the wide gap material (SiC). In this study, layers of nanocrystalline porous silicon and porous silicon carbide are used as wide-gap material layers in photosensitive structures. The spectral characteristics of the specular reflectance of these materials are investigated.

*Keywords:* porous silicon, photoluminescence, rare earth elements,, photoelectric converters, silicon carbide, reflection coefficient

---

## 1. Introduction

The use of porous silicon (por-Si) as a sensitive layer in multilayer heterostructures of silicon photoelectric converters makes it possible to significantly increase the efficiency of energy conversion [1]. A promising sensitive layer of a photoelectric converters is a layer with silicon nanocrystals, as well as layers of wide-band materials. At the same time, the absorption spectrum of the photoconductivity spectrum expands into the short-wavelength region due to the quantum-size increase in the width of the band gap of silicon in nanocrystals and due to the absorption of high-energy photons in the wide-band material. An effective system of silicon nanocrystals can be a layer of porous silicon, since the pore walls are a disordered system of quantum wells, filaments and quantum dots [2]. In addition, due to the developed pore system, the area of the absorbing surface of the photodetector increases significantly. However, a number of existing problems prevents the use of porous silicon in the photoelectric converters. This low reproducibility of results due to uncontrolled factors of the technological process, instability of the PC parameters due to the reagent remaining in its pores, as well as its high electrical resistance. The solution to these problems can be the creation of a porous layer locally on the surface with seeds of pore formation, as well as the use of a stabilizing coating, which can be a wide-band silicon carbide semiconductor. The aim of this work was to study the photoelectric properties of samples of multilayer photosensitive structures with a porous layer locally created on the working surface and a stabilizing coating of silicon carbide. The porous layer was created on silicon substrates with textured and ground surfaces. The seeds of pore formation on the ground and textured surfaces are the depressions of the microrelief, where the electric-field intensity is maximum, so a porous layer on such surfaces is formed locally [3]. Samples with a polished surface have served as tested ones. Some samples were carbidized, that is, an epitaxial layer of silicon carbide was created on the surface of the porous layer, so that the samples were Si / SiC heterostructures with a large area of the absorbing surface.

## 2. Experimental technique

To create a porous layer, silicon plates were subjected to electrochemical etching in a vertical cell in water-alcohol solutions of hydrofluoric acid.



Carbideization of the samples leading to the formation of SiC / Si heterostructures was carried out by gas-transfer endotaxy in a hydrogen stream in a vertical reactor with cold walls using a graphite container [4].

The measurements included the measurement of the specular reflection coefficient and the structures photoluminescence. The photoluminescence spectra were measured by excitation with an ultraviolet (330 nm) laser at the room temperature for samples with a porous layer doped with rare-earth elements (REE) such as erbium or ytterbium.

The spectral dependences of the reflection coefficients were studied using a Shimadzu UV-2450 spectrophotometer with a prefix 206-14046. The measurement range was 0.3 - 1  $\mu\text{m}$ , the measurement step and the spectral width of the monochromator slit were 2 nm, the scanning rate was slow. The angle of radiation incidence having an elliptical polarization of about 3: 1 - 4: 1 was  $5^\circ$  with an aperture of not more than  $5^\circ$ . The radiation receiver of the Shimadzu UV-2450 spectrophotometer is a photoelectric multiplier. This causes significant noise measurements of the instrument in the near infrared region.

### 3. Morphology

Figure 1 shows SEM images of transverse cleavages of samples with a porous layer formed on the polished (a) and textured (b) surfaces. On the SEM image of the surface of the textured layer in the region of the junction of the pyramids, it is clearly seen that the porous layer was formed predominantly in the depression of the relief (the darker areas in Fig. 1, c). The sample with a textured surface has undergone carbideization, as a result of which nanowires of carbon clearly visible on the cleaved surface were formed in some regions of the surface. The porous layer thickness of this sample was 12.55  $\mu\text{m}$ .

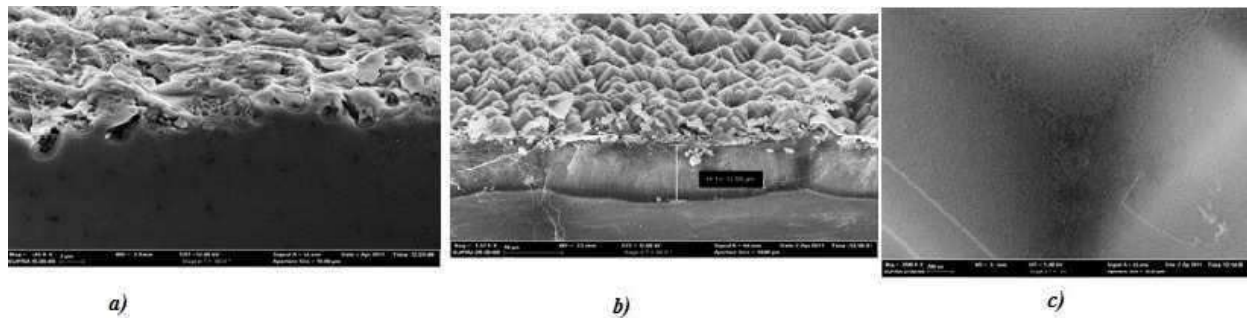


Figure 1: SEM images of transverse cleavages of samples with a porous layer formed on (a) the porous layer is formed on ground surfaces; (b) the porous layer is formed on textured surfaces; (c) the image of a textured surface in the area on the pyramids joint, where a porous layer was formed.

### 4. Spectral dependences of the reflection coefficient

The spectral dependences of the reflection coefficients were studied using a SHIMADZU UV-2450PC spectrophotometer in the wavelength range from 0.3 to 1  $\mu\text{m}$  on different types of the working surface of the samples. Samples with a porous surface (past electrochemical etching), as well as samples that have undergone carbideization, have entered the measuring group. KDB-3 silicon plates with textured, ground or polished surfaces without pores were used as test samples. We can see that the formation of a porous layer significantly reduces the reflection coefficient, while the course of the curves of the spectral dependences remains almost unchanged, which is explained by the local nature of pore formation. An exception is a sample with a textured surface that was etched for 5 minutes, its reflectivity coefficient in the short-wave part of the spectrum is noticeably lower than in others ones. This is explained by pore formation dynamics on such surface. At the initial stage of etching, silicon nanocrystals are formed almost over the entire surface of the textured layer, and with further etching some of them located on the walls of the pyramids dissolve, and the formation of the porous layer only occurs in the relief depressions [5]. A similar course of the curve for

the spectral dependence of the reflection coefficient is also observed for carbided samples with a "failure" of the reflection coefficient in the 250 - 300 nm x-band. It is explained by the absorption of light in nanocrystals of wide-band silicon carbide [6].

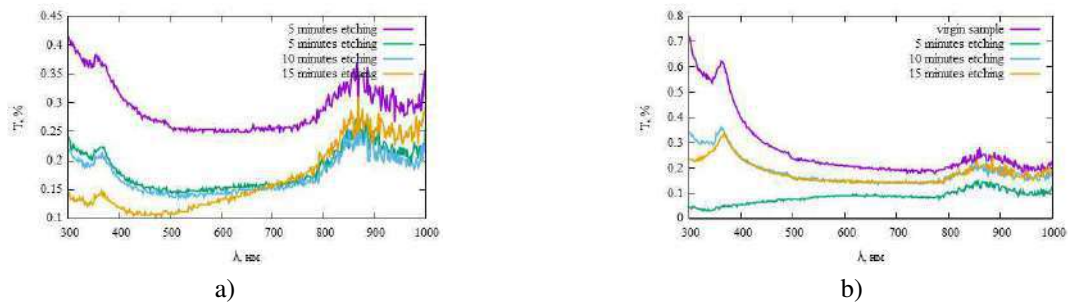


Figure 2: Spectral dependences of the reflection coefficients of samples with a porous layer formed at different etching time on: (a) ground surface; (b) textured surface.

## 5. Spectral dependences of photoluminescence

Figure 3 shows the photoluminescence spectra of porous silicon samples doped with ytterbium (a) and erbium (b), where narrow peaks of the spectral maxima of ytterbium emission at 980 nm and erbium emission at 1550 nm are clearly visible. Since the mechanism of photoluminescence of REE ions in a solid silicon matrix is based on the recombination of an exciton generated by radiation in a silicon nanocrystal, the presence of sufficiently intense peaks of PL of ytterbium and erbium in the spectra of the samples under study confirms the presence of a sufficiently large concentration of nanocrystals in porous layers [7]. In figure 3, a wide band of 550 - 750 nm corresponding to the emission spectrum of silicon nanocrystals is also visible, as well as a laser pumping peak at 370 nm.

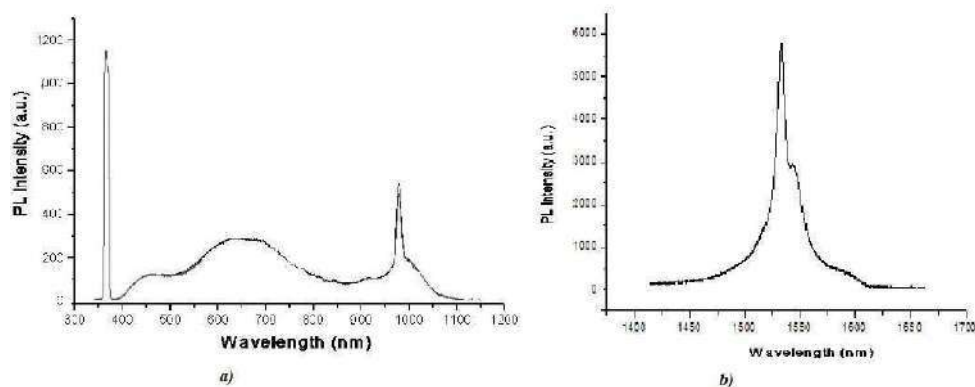


Figure 3: Photoluminescence spectra of samples with a layer of porous silicon formed on a textured surface: a) doped with ytterbium; b) doped with erbium.

## 6. A reflecting surface model

The working surface of the photosensitive structures can be represented as a consisting of two components one with different reflection coefficients: textured and porous. The microrelief on the textured surface of silicon is an etching polyhedron in the form of regular tetragonal pyramids with lateral faces that are natural surfaces of a single crystal and an angle at the apex of  $70.5^\circ$ . The textured surface reduces optical losses due to the total effect of multiple reflection of the incident beam from the frontal surface and multiple total internal reflection from the back and side



surfaces. The trajectories of light rays on an idealized textured surface with the refractive index of the medium  $n=1$  are shown in Figure 4. The light normally incident to the surface undergoes several reflections, as a result of which the intensity of the reflected light will decrease as a power of multiplicity. The nature of the interaction of the incident radiation with such a surface will strongly depend on the relationship between the wavelength and the geometric dimensions of the relief. While the geometric dimension of the microrelief exceeds the wavelength of the radiation, the laws of geometric optics operate, that is, multiple effects reflection take place here. If the height of the relief is comparable with the wavelength or much less than the latter, then with respect to this radiation the surface is perfectly smooth and manifests itself as highly reflective. When reflecting from the microrelief surfaces in the area where the radiation "feels" the microrelief along with the mirror component of the reflected light, there is also a diffusely scattered constituent. Measurements made in the wavelength range  $0.5-1.2 \mu\text{m}$  give the value  $R = 5-7\%$ . The application of multilayer antireflection coatings makes it possible to reduce the reflection coefficient almost to zero [8].

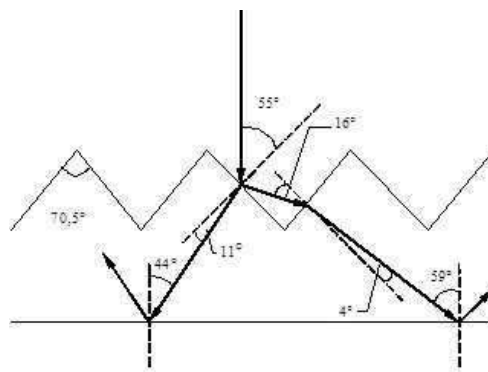


Figure 4: A trajectory of the light rays on the idealized textured surface of the solar cell with refractive indices of the medium  $n = 1$  and  $n_{Si} = 3.8$ .

According to the data of [9], the reflection coefficient can be determined from the formula:

$$R = (1 - \delta)r_1r_2 + \delta r_1r_2r_3, \tag{1}$$

Where  $\delta$  is a part of the secondary reflected light,  $r_1, r_2, r_3$  are reflection coefficients from the successive faces, depending on the light wavelength. Since the faces of the pyramids are atomically smooth surfaces oriented along the crystallographic plane (111), the numerical values of these coefficients can take the values of the reflection coefficients from the polished silicon surface for a given wavelength. The reflectivity of the surface of porous silicon depends strongly on its porosity. With a large degree of porosity, the reflection coefficient of the porous layer can be practically zero throughout the visible range of wavelengths of the incident radiation.

Let us consider, for convenience of calculation, an ideal textured surface, averaging the heights of the pyramids:

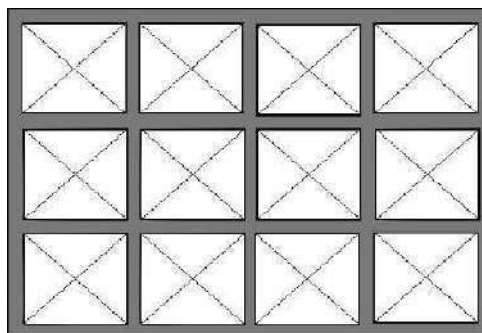


Figure 5: A schematic representation of the ideal surface (top view)

The height of these pyramids was  $2.26 \mu\text{m}$ , base -  $3.23 \mu\text{m}$ ; the width of the regions of the porous layer (gray in Figure 5) is  $1.25 \mu\text{m}$ . Calculations of the reflection coefficient were carried out for a surface, 56.97% of the area occupied by a textured surface and 43.03% of the surface area of a porous layer (Fig. 6).

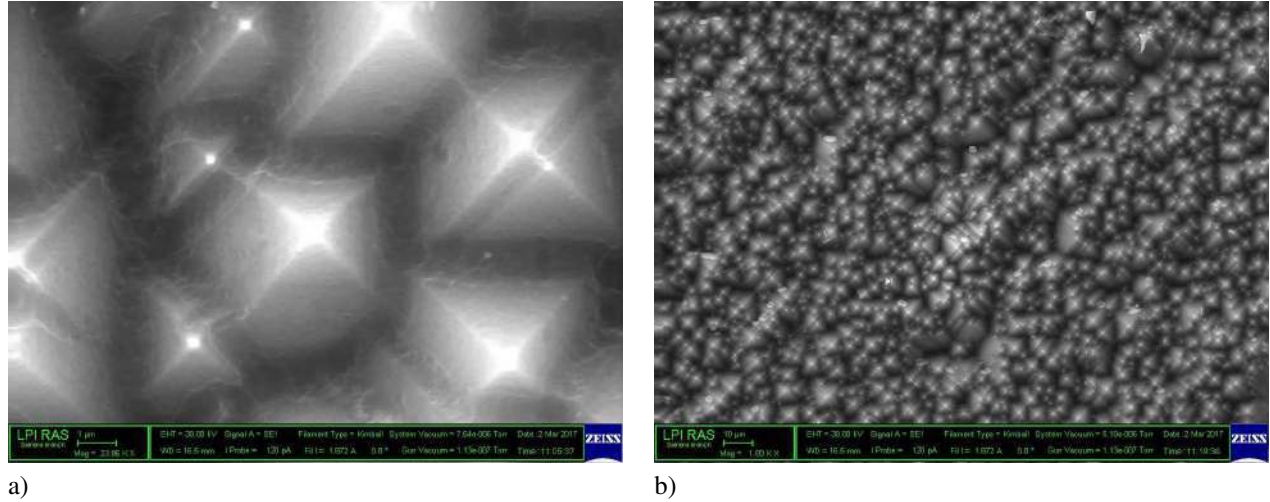


Figure 6: SEM images of the simulated surface on a different scale

The numerical values of the area shares were determined using the program JMicroVision v1.27 (Fig. 7).

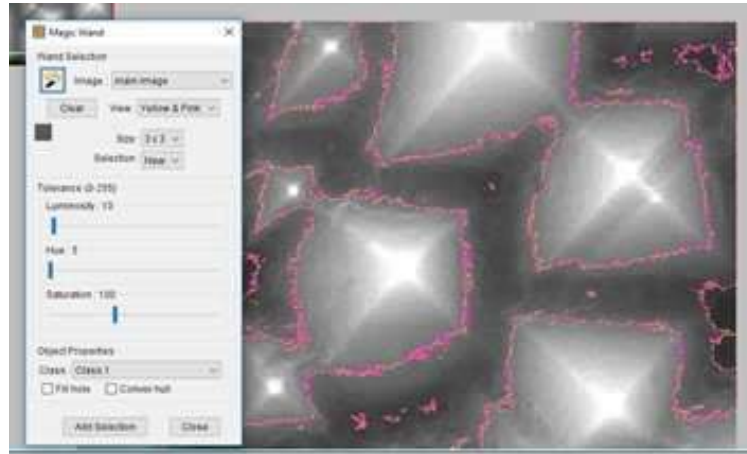


Figure 7: Isolation of regions with different reflectivity on the investigated surface

The simulation results are shown in Fig. 8. The reflection coefficient  $R$  is calculated at an angle of incidence of light to the first face =  $54^\circ 40'$ , to the opposite face =  $16^\circ$ , a small part of the secondary reflected light (= 10%) returns to the first face at an angle of  $86^\circ 40'$ . The calculation is made for wavelengths of  $0.4\text{--}0.75 \mu\text{m}$ . When taking into account the illumination of different faces, the reflection coefficient from the textured part of the surface is calculated by the formula:

$$R = 0.35[r_1 r_2 (1 - \delta) + \delta r_1 r_2 r_3] + 0.65 r_1' r_2', \quad (2)$$

Indices and strokes for  $r$  correspond to the notation for the angles of incidence. The coefficient of reflection from the porous part of the surface was assumed to be zero throughout the investigated wavelength range.

Analyzing the obtained data, we can say that the reflection coefficient, experimentally determined, at a given degree of filling is much lower than the theoretical one. This can be explained by the fact that the real faces of the

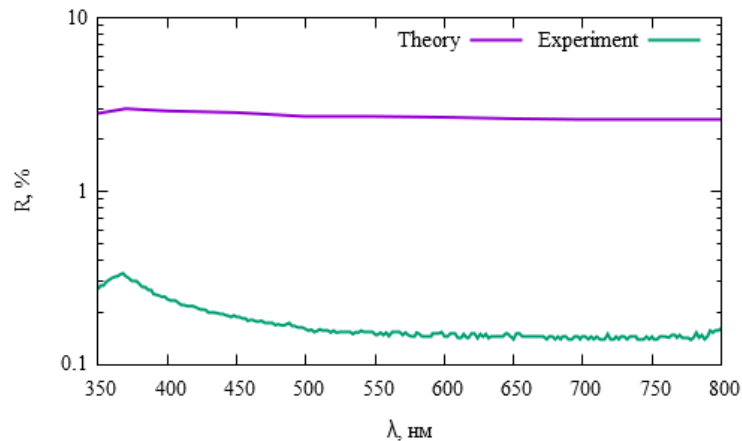


Figure 8: Spectral dependence of the reflection coefficient of samples with a fillability of 43% of the surface by a porous layer, both theoretical and experimental curves

pyramids are not atomically smooth. As a result of electrochemical etching, a weakly expressed relief appears on them. It significantly reduces their reflection coefficient, which is also observed in the works of other authors [10]. It should be noted that the values of the experimental reflection coefficient for a textured silicon surface with a porous layer, obtained in [10], are much higher than those ones obtained in this work. It may be explained by a significantly smaller fraction of the surface occupied by the porous layer.

## 7. Conclusion

Thus, the conducted studies show that porous silicon, created on a textured or polished surface, is an effective system of nanocrystals. Its application in a photoconductive device makes it possible to increase significantly the fraction of absorbed radiation in the spectral range 400-1000 nm and to increase the photocurrent.

## References

- [1] Latukhina, N. Efficient silicon solar cells for space and ground-based aircraft./ N. Latukhina , A. Rogozin , G. Puzyrnaya , D. Lizunkova , A. Gurtovb, S. Ivkov // *ProcediaEngineering*. 2015.- Vol. 104. -p. 157-161
- [2] Latukhina, N.V. New prospects of old materials: silicon and silicon carbide / N.V. Latukhina, V.I. Chepurinov, G.A. Pisarenko // *Electronics of the NTB*. - 2013. - 4 (00126) - p.104-110.
- [3] Sokolov, V.I. Some characteristics of porous silicon (reflection, scattering, refractive index, microhardness) / V.I. Sokolov, A.I. Shelnyh. // *JETP letters*. -2008-V.34, . 5. - p. 34-39.
- [4] Latukhina, N.V. Photosensitive Heterostructures on the Basis of Nanocrystal Porous Silicon/N. V. Latukhina, A. S. Rogozhin, S. Saed, V. I. Chepurinov// *Russian Microelectronics*, 2016, Vol. 45, Nos. 89, pp. 613618
- [5] Gosteva E.A. Investigation of the coefficient in silicon structures with different porosity ./ E.A. Gosteva, V.V. Starkov, Yu. N. Parkhomenko // *Nanostructured materials and conversion devices for solar energy: a collection of proceedings of the IV All-Russian Scientific Conference (September 29-30, 2016, Cheboksary)*, 2016, p.59-63
- [6] Kirsanov, N. Yu. Multilayer Photosensitive Structures Based on Porous Silicon and Rare-Earth-Element Compounds: Study of Spectral Characteristics/ N. Yu. Kirsanov , N. V. Latukhina, D. A. Lizunkova, G. A. Rogozhina, and M. V. Stepikhova// *Semiconductors*, 2017, Vol. 51, No. 3, pp. 353356
- [7] Sokolov S.A. Photoluminescence of Rare Earth Ions (Er<sup>3+</sup>, Yb<sup>3+</sup>) in a Porous Silicon Matrix/ S. A. Sokolov, R. Rsslhuber, D.M. Zhigunov, N.V. Latukhina, V.Yu. Timoshenko // *Thin Solid Films*, 2014, V.562, p. 462-466
- [8] Gorbach T.Ya. Selective properties of an anisotropically etched surface/ T.Ya. Gorbach, S.V. Svechnikov, N.V. Kotova, E.V. Podlisny // *Optoelectronics and Semiconductor Technology*, 1986. V.10. - p. 649.
- [9] Borodina N.M. Silicon photoconverters with a textured surface and their properties / N.M. Borodina, A.K. Zaitseva, E.A. Marasanova, A. A. Polisman // *Helio Technique*, 1982. 3, p.6-11.
- [10] Hyukyong Kwon. Investigation of Antireflective Porous Silicon Coating for Solar Cells/ Hyukyong Kwon, Jaedoo Lee, Minjeong Kim, and Soohong Lee// *International Scholarly Research Network ISRN Nanotechnology*, V. 2011, p.1-4

# Nanocrystalline Silicon and Silicon Carbide Optical Properties

Daria Lizunkova<sup>1</sup>, Natalya Latukhina<sup>1</sup>, Victor Chepurinov<sup>1</sup> and Vyacheslav Parandin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoe shosse, Samara, 443086, Russia

---

## Abstract

Porous silicon possesses a wide range of the unique properties and has good perspectives for photo-sensitive structures for a solar cells new generation. Due to the developed pore system, the area of absorbing surface increases, and also the increased sensitivity expands into the short-wavelength region due to the increased energy band gap of silicon nano-particles and silicon nano-filaments on the walls of the pores. Carbonization of the surface layer of porous silicon makes absorption even more effective. In this case, the spectrum of the solar cell expands into the short-wavelength region due to absorption of the high-energy photons in the wide gap material (SiC). In this study, layers of nanocrystalline porous silicon and porous silicon carbide are used as wide-gap material layers in photosensitive structures. The spectral characteristics of the specular reflectance of these materials are investigated.

*Keywords:* porous silicon, photoluminescence, rare earth elements,, photoelectric converters, silicon carbide, reflection coefficient

---

## 1. Introduction

The use of porous silicon (por-Si) as a sensitive layer in multilayer heterostructures of silicon photoelectric converters makes it possible to significantly increase the efficiency of energy conversion [1]. A promising sensitive layer of a photoelectric converters is a layer with silicon nanocrystals, as well as layers of wide-band materials. At the same time, the absorption spectrum of the photoconductivity spectrum expands into the short-wavelength region due to the quantum-size increase in the width of the band gap of silicon in nanocrystals and due to the absorption of high-energy photons in the wide-band material. An effective system of silicon nanocrystals can be a layer of porous silicon, since the pore walls are a disordered system of quantum wells, filaments and quantum dots [2]. In addition, due to the developed pore system, the area of the absorbing surface of the photodetector increases significantly. However, a number of existing problems prevents the use of porous silicon in the photoelectric converters. This low reproducibility of results due to uncontrolled factors of the technological process, instability of the PC parameters due to the reagent remaining in its pores, as well as its high electrical resistance. The solution to these problems can be the creation of a porous layer locally on the surface with seeds of pore formation, as well as the use of a stabilizing coating, which can be a wide-band silicon carbide semiconductor. The aim of this work was to study the photoelectric properties of samples of multilayer photosensitive structures with a porous layer locally created on the working surface and a stabilizing coating of silicon carbide. The porous layer was created on silicon substrates with textured and ground surfaces. The seeds of pore formation on the ground and textured surfaces are the depressions of the microrelief, where the electric-field intensity is maximum, so a porous layer on such surfaces is formed locally [3]. Samples with a polished surface have served as tested ones. Some samples were carbidized, that is, an epitaxial layer of silicon carbide was created on the surface of the porous layer, so that the samples were Si / SiC heterostructures with a large area of the absorbing surface.

## 2. Experimental technique

To create a porous layer, silicon plates were subjected to electrochemical etching in a vertical cell in water-alcohol solutions of hydrofluoric acid.

Carbide formation of the samples leading to the formation of SiC / Si heterostructures was carried out by gas-transfer endotaxy in a hydrogen stream in a vertical reactor with cold walls using a graphite container [4].

The measurements included the measurement of the specular reflection coefficient and the structures photoluminescence. The photoluminescence spectra were measured by excitation with an ultraviolet (330 nm) laser at the room temperature for samples with a porous layer doped with rare-earth elements (REE) such as erbium or ytterbium.

The spectral dependences of the reflection coefficients were studied using a Shimadzu UV-2450 spectrophotometer with a prefix 206-14046. The measurement range was 0.3 - 1  $\mu\text{m}$ , the measurement step and the spectral width of the monochromator slit were 2 nm, the scanning rate was slow. The angle of radiation incidence having an elliptical polarization of about 3: 1 - 4: 1 was  $5^\circ$  with an aperture of not more than  $5^\circ$ . The radiation receiver of the Shimadzu UV-2450 spectrophotometer is a photoelectric multiplier. This causes significant noise measurements of the instrument in the near infrared region.

### 3. Morphology

Figure 1 shows SEM images of transverse cleavages of samples with a porous layer formed on the polished (a) and textured (b) surfaces. On the SEM image of the surface of the textured layer in the region of the junction of the pyramids, it is clearly seen that the porous layer was formed predominantly in the depression of the relief (the darker areas in Fig. 1, c). The sample with a textured surface has undergone carbidisation, as a result of which nanowires of carbon clearly visible on the cleaved surface were formed in some regions of the surface. The porous layer thickness of this sample was 12.55  $\mu\text{m}$ .

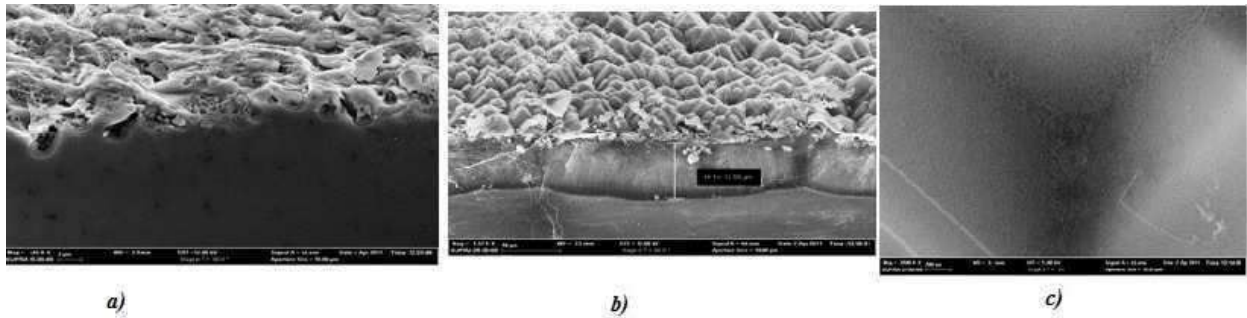


Figure 1: SEM images of transverse cleavages of samples with a porous layer formed on (a) the porous layer is formed on ground surfaces; (b) the porous layer is formed on textured surfaces; (c) the image of a textured surface in the area on the pyramids joint, where a porous layer was formed.

### 4. Spectral dependences of the reflection coefficient

The spectral dependences of the reflection coefficients were studied using a SHIMADZU UV-2450PC spectrophotometer in the wavelength range from 0.3 to 1  $\mu\text{m}$  on different types of the working surface of the samples. Samples with a porous surface (past electrochemical etching), as well as samples that have undergone carbidization, have entered the measuring group. KDB-3 silicon plates with textured, ground or polished surfaces without pores were used as test samples. We can see that the formation of a porous layer significantly reduces the reflection coefficient, while the course of the curves of the spectral dependences remains almost unchanged, which is explained by the local nature of pore formation. An exception is a sample with a textured surface that was etched for 5 minutes, its reflectivity coefficient in the short-wave part of the spectrum is noticeably lower than in others ones. This is explained by pore formation dynamics on such surface. At the initial stage of etching, silicon nanocrystals are formed almost over the entire surface of the textured layer, and with further etching some of them located on the walls of the pyramids dissolve, and the formation of the porous layer only occurs in the relief depressions [5]. A similar course of the curve for

the spectral dependence of the reflection coefficient is also observed for carbided samples with a "failure" of the reflection coefficient in the 250 - 300 nm x-band. It is explained by the absorption of light in nanocrystals of wide-band silicon carbide [6].

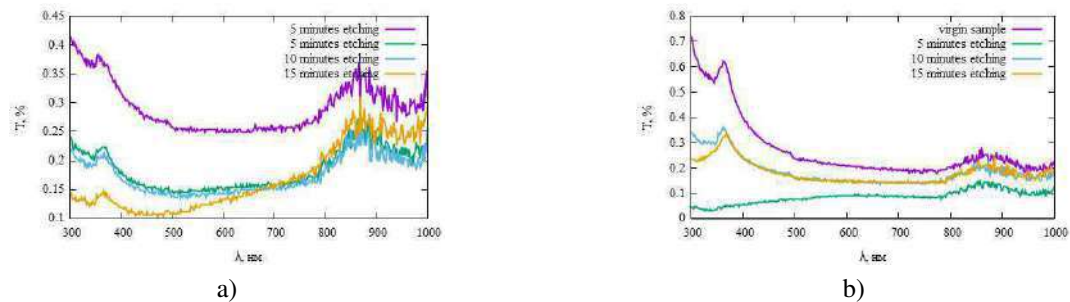


Figure 2: Spectral dependences of the reflection coefficients of samples with a porous layer formed at different etching time on: (a) ground surface; (b) textured surface.

## 5. Spectral dependences of photoluminescence

Figure 3 shows the photoluminescence spectra of porous silicon samples doped with ytterbium (a) and erbium (b), where narrow peaks of the spectral maxima of ytterbium emission at 980 nm and erbium emission at 1550 nm are clearly visible. Since the mechanism of photoluminescence of REE ions in a solid silicon matrix is based on the recombination of an exciton generated by radiation in a silicon nanocrystal, the presence of sufficiently intense peaks of PL of ytterbium and erbium in the spectra of the samples under study confirms the presence of a sufficiently large concentration of nanocrystals in porous layers [7]. In figure 3, a wide band of 550 - 750 nm corresponding to the emission spectrum of silicon nanocrystals is also visible, as well as a laser pumping peak at 370 nm.

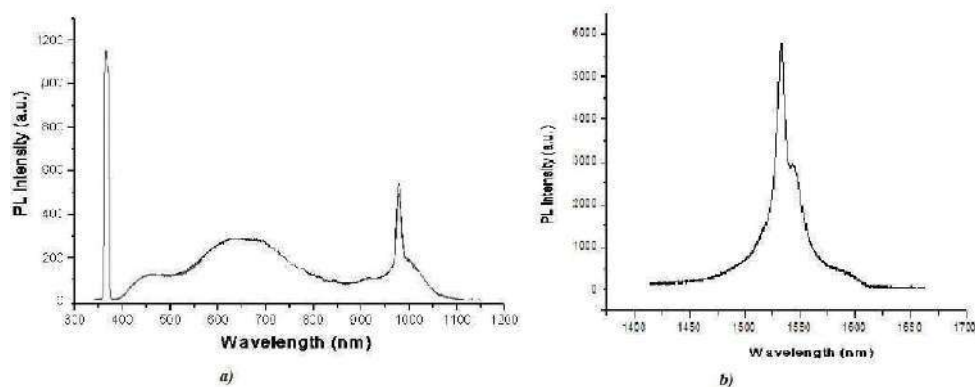


Figure 3: Photoluminescence spectra of samples with a layer of porous silicon formed on a textured surface: a) doped with ytterbium; b) doped with erbium.

## 6. A reflecting surface model

The working surface of the photosensitive structures can be represented as a consisting of two components one with different reflection coefficients: textured and porous. The microrelief on the textured surface of silicon is an etching polyhedron in the form of regular tetragonal pyramids with lateral faces that are natural surfaces of a single crystal and an angle at the apex of  $70.5^\circ$ . The textured surface reduces optical losses due to the total effect of multiple reflection of the incident beam from the frontal surface and multiple total internal reflection from the back and side

surfaces. The trajectories of light rays on an idealized textured surface with the refractive index of the medium  $n=1$  are shown in Figure 4. The light normally incident to the surface undergoes several reflections, as a result of which the intensity of the reflected light will decrease as a power of multiplicity. The nature of the interaction of the incident radiation with such a surface will strongly depend on the relationship between the wavelength and the geometric dimensions of the relief. While the geometric dimension of the microrelief exceeds the wavelength of the radiation, the laws of geometric optics operate, that is, multiple effects reflection take place here. If the height of the relief is comparable with the wavelength or much less than the latter, then with respect to this radiation the surface is perfectly smooth and manifests itself as highly reflective. When reflecting from the microrelief surfaces in the area where the radiation "feels" the microrelief along with the mirror component of the reflected light, there is also a diffusely scattered constituent. Measurements made in the wavelength range  $0.5-1.2 \mu\text{m}$  give the value  $R = 5-7\%$ . The application of multilayer antireflection coatings makes it possible to reduce the reflection coefficient almost to zero [8].

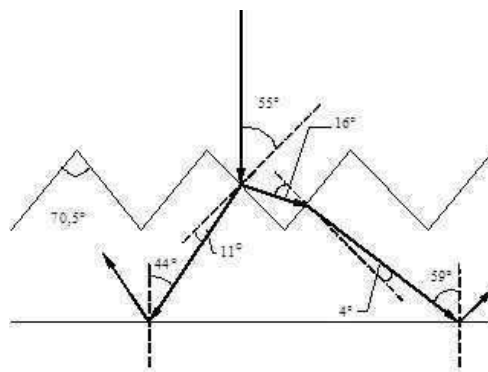


Figure 4: A trajectory of the light rays on the idealized textured surface of the solar cell with refractive indices of the medium  $n = 1$  and  $n_{Si} = 3.8$ .

According to the data of [9], the reflection coefficient can be determined from the formula:

$$R = (1 - \delta)r_1r_2 + \delta r_1r_2r_3, \tag{1}$$

Where  $\delta$  is a part of the secondary reflected light,  $r_1, r_2, r_3$  are reflection coefficients from the successive faces, depending on the light wavelength. Since the faces of the pyramids are atomically smooth surfaces oriented along the crystallographic plane (111), the numerical values of these coefficients can take the values of the reflection coefficients from the polished silicon surface for a given wavelength. The reflectivity of the surface of porous silicon depends strongly on its porosity. With a large degree of porosity, the reflection coefficient of the porous layer can be practically zero throughout the visible range of wavelengths of the incident radiation.

Let us consider, for convenience of calculation, an ideal textured surface, averaging the heights of the pyramids:

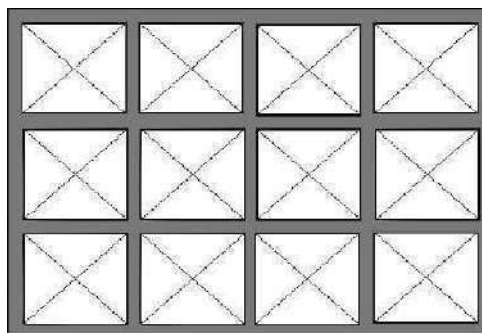


Figure 5: A schematic representation of the ideal surface (top view)



The height of these pyramids was  $2.26 \mu\text{m}$ , base -  $3.23 \mu\text{m}$ ; the width of the regions of the porous layer (gray in Figure 5) is  $1.25 \mu\text{m}$ . Calculations of the reflection coefficient were carried out for a surface, 56.97% of the area occupied by a textured surface and 43.03% of the surface area of a porous layer (Fig. 6).

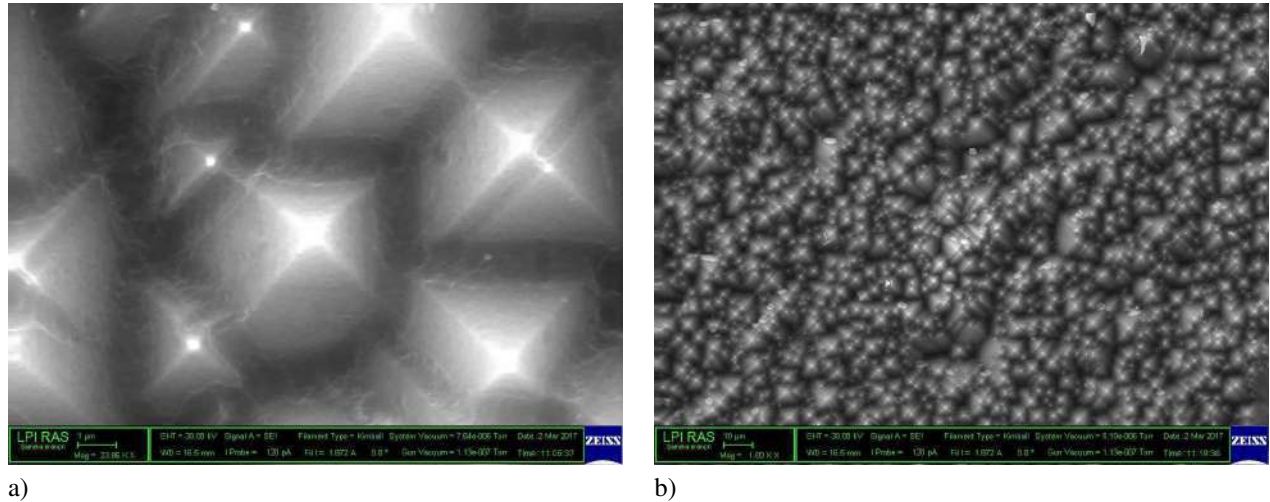


Figure 6: SEM images of the simulated surface on a different scale

The numerical values of the area shares were determined using the program JMicroVision v1.27 (Fig. 7).

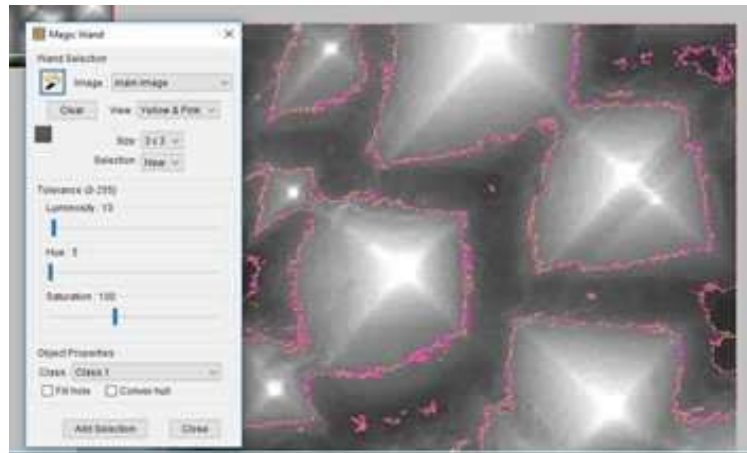


Figure 7: Isolation of regions with different reflectivity on the investigated surface

The simulation results are shown in Fig. 8. The reflection coefficient  $R$  is calculated at an angle of incidence of light to the first face =  $54^\circ 40'$ , to the opposite face =  $16^\circ$ , a small part of the secondary reflected light (= 10%) returns to the first face at an angle of  $86^\circ 40'$ . The calculation is made for wavelengths of  $0.4\text{--}0.75 \mu\text{m}$ . When taking into account the illumination of different faces, the reflection coefficient from the textured part of the surface is calculated by the formula:

$$R = 0.35[r_1 r_2 (1 - \delta) + \delta r_1 r_2 r_3] + 0.65 r_1' r_2', \quad (2)$$

Indices and strokes for  $r$  correspond to the notation for the angles of incidence. The coefficient of reflection from the porous part of the surface was assumed to be zero throughout the investigated wavelength range.

Analyzing the obtained data, we can say that the reflection coefficient, experimentally determined, at a given degree of filling is much lower than the theoretical one. This can be explained by the fact that the real faces of the



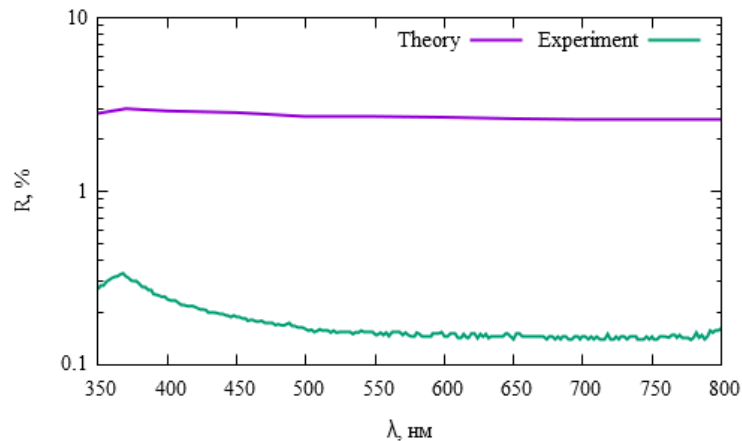


Figure 8: Spectral dependence of the reflection coefficient of samples with a fillability of 43% of the surface by a porous layer, both theoretical and experimental curves

pyramids are not atomically smooth. As a result of electrochemical etching, a weakly expressed relief appears on them. It significantly reduces their reflection coefficient, which is also observed in the works of other authors [10]. It should be noted that the values of the experimental reflection coefficient for a textured silicon surface with a porous layer, obtained in [10], are much higher than those ones obtained in this work. It may be explained by a significantly smaller fraction of the surface occupied by the porous layer.

## 7. Conclusion

Thus, the conducted studies show that porous silicon, created on a textured or polished surface, is an effective system of nanocrystals. Its application in a photoconductive device makes it possible to increase significantly the fraction of absorbed radiation in the spectral range 400-1000 nm and to increase the photocurrent.

## References

- [1] Latukhina, N. Efficient silicon solar cells for space and ground-based aircraft./ N. Latukhina , A. Rogozin , G. Puzyrnaya , D. Lizunkova , A. Gurtovb, S. Ivkov // *ProcediaEngineering*. 2015.- Vol. 104. -p. 157-161
- [2] Latukhina, N.V. New prospects of old materials: silicon and silicon carbide / N.V. Latukhina, V.I. Chepurinov, G.A. Pisarenko // *Electronics of the NTB*. - 2013. - 4 (00126) - p.104-110.
- [3] Sokolov, V.I. Some characteristics of porous silicon (reflection, scattering, refractive index, microhardness) / V.I. Sokolov, A.I. Shelnih. // *JETP letters*. -2008-V.34, . 5. - p. 34-39.
- [4] Latukhina, N.V. Photosensitive Heterostructures on the Basis of Nanocrystal Porous Silicon/N. V. Latukhina, A. S. Rogozhin, S. Saed, V. I. Chepurinov// *Russian Microelectronics*, 2016, Vol. 45, Nos. 89, pp. 613618
- [5] Gosteva E.A. Investigation of the coefficient in silicon structures with different porosity ./ E.A. Gosteva, V.V. Starkov, Yu. N. Parkhomenko // *Nanostructured materials and conversion devices for solar energy: a collection of proceedings of the IV All-Russian Scientific Conference (September 29-30, 2016, Cheboksary)*, 2016, p.59-63
- [6] Kirsanov, N. Yu. Multilayer Photosensitive Structures Based on Porous Silicon and Rare-Earth-Element Compounds: Study of Spectral Characteristics/ N. Yu. Kirsanov , N. V. Latukhina, D. A. Lizunkova, G. A. Rogozhina, and M. V. Stepikhova// *Semiconductors*, 2017, Vol. 51, No. 3, pp. 353356
- [7] Sokolov S.A. Photoluminescence of Rare Earth Ions (Er<sup>3+</sup>, Yb<sup>3+</sup>) in a Porous Silicon Matrix/ S. A. Sokolov, R. Rsslhuber, D.M. Zhigunov, N.V. Latukhina, V.Yu. Timoshenko // *Thin Solid Films*, 2014, V.562, p. 462-466
- [8] Gorbach T.Ya. Selective properties of an anisotropically etched surface/ T.Ya. Gorbach, S.V. Svechnikov, N.V. Kotova, E.V. Podlisny // *Optoelectronics and Semiconductor Technology*, 1986. V.10. - p. 649.
- [9] Borodina N.M. Silicon photoconverters with a textured surface and their properties / N.M. Borodina, A.K. Zaitseva, E.A. Marasanova, A. A. Polisman // *Helio Technique*, 1982. 3, p.6-11.
- [10] Hyukyong Kwon. Investigation of Antireflective Porous Silicon Coating for Solar Cells/ Hyukyong Kwon, Jaedoo Lee, Minjeong Kim, and Soohong Lee// *International Scholarly Research Network ISRN Nanotechnology*, V. 2011, p.1-4

# The combination of Raman spectroscopy and Autofluorescence analysis for estimation of blood and urine homeostasis

L.A. Shamina<sup>1</sup>, I.A. Bratchenko<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

In this study we measured spectral features of blood and urine by Raman spectroscopy and autofluorescence analysis. Analysis of specific spectra allows for identification of informative spectral bands proportional to components whose content is associated with body fluids homeostasis changes at various pathological conditions. In general, the developed approach of body fluids analysis provides the basis of a useful and minimally invasive method of pathologies screening.

*Keywords:* Raman spectroscopy; autofluorescence; biofluidity homeostasis; blood; urine; pathology

---

## 1. Introduction

The pathological conditions provoke alterations in body fluids homeostasis; therefore it is possible to use the component composition analysis of urine, blood, saliva and other body fluids for pathologies detection such as cancer [1]. Presently, the biochemical analyses are widely used for the body fluids cancer diagnosis. In addition to the laboratory methods today a variety of physical and chemical methods of analysis may be successfully utilized for the study of the body fluids composition. Physical methods have such advantages as simplicity of sample preparation, wide dynamic range and great versatility in comparison with chemical methods of analysis. Therefore, body fluids analysis with optical methods can become a successful alternative to existing laboratory methods. Raman Spectroscopy (RS) and autofluorescence (AF) analysis allow for the homeostasis changes detection in the body fluids at the molecular level [2]. These techniques are successfully used in different branches of clinical medicine and in the experimental studies of the body fluids composition for the various locations cancer detection. The aim of this work is to study the spectral features of blood and urine from patients with tumors for identification criteria that may allow to estimate the homeostasis changes and the tumor presence.

## 2. Materials and methods

### 2.1. Experimental setup

Study of the body fluids spectral features was performed with the experimental setup shown in Fig. 1. The excitation of collected spectra was performed by the laser module LuxxMaster LML-785.0RB-04 (central wavelength 785 nm). The fiber-optic Raman probe RPB785 allows for focusing of the exiting radiation, collecting and filtering of the scattered radiation. The collected signal was decomposed into a spectrum using a high-resolution Shamrock SR-500i-D1-R spectrograph with integrated cooled up to -65°C digital camera ANDOR DU416A-LDC-DD. Tested body fluids were placed in the PMMA cuvette with an aluminum coating. The cuvette geometry (depth 6.5 mm, radius of deepening curvature 19 mm) was optimized to match the working distance of probe focusing lens. The Raman probe was normally positioned on the axis of the deepening; a detailed description of the utilized experimental setup is presented in [3]. The utilized spectrograph with a grating of 600 slits/mm allows for recording the spectrum of the tested substance in 780-950 nm area divided by three spectral ranges; for each single spectral range the exposure time was 20 seconds. A sequential recording of three spectra for each tested sample was performed. The final spectrum was received from averaging of all three recorded spectra. The total time of the final spectrum recording was 3 minutes.

### 2.2. Samples preparation

The standardized collection of blood and urine samples from patients of Samara Regional Clinical Oncology Dispensary was performed. Collected samples were placed in sterile test-tubes and were stored at +2 + 4°C before the analysis. Analysis of collected body fluids was performed within 60 h after sample collection. Patients of Samara Regional Clinical Oncology Dispensary with malignant tumors or benign tumors were enrolled in this study. Patients with systemic diseases and patients taking any medical antitumor drugs were excluded from the study.

### 2.3. Spectra processing

Recorded spectra were processed by the method proposed by Zeng et al [4] for AF and Raman signals separation. The processing of experimental data was performed on the bases of regression analysis. Definition of spectrum informative bands during the regression model constructing was performed by the analysis of the variable importance in projection (VIP) [5]. The

VIP distribution makes it possible to define the most informative spectral bands in the blood and urine spectra specific for patients with lung cancer.

### 3. Results

#### 3.1. Spectral characteristics of blood

Currently a set of biochemical methods for analyzing body fluids aimed to detect a pathological process is used in laboratory diagnostics. Biochemical methods make it possible to perform a quantitative estimation of certain organic and mineral components level, as well as that of enzymes and hormones and to detect their deviations from the norm [6]. Qualitative estimation of the level of indicators included in the standard biochemical analysis is possible when studying the spectral characteristics of body fluids. Figure 1 shows the common Raman spectrum of blood.

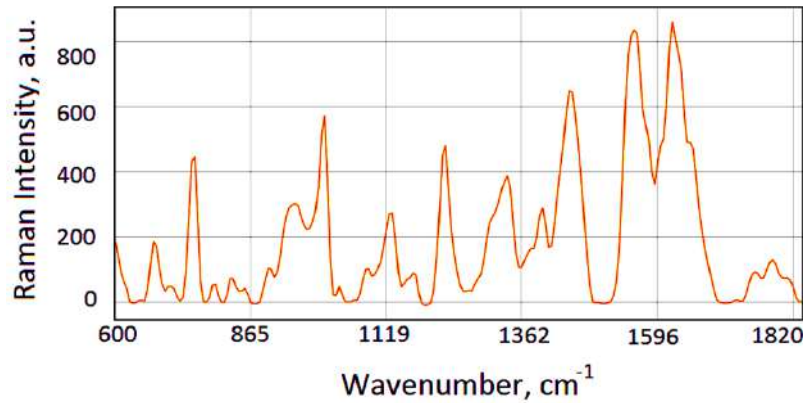


Fig.1. Common Raman spectrum of blood.

Human body fluids have a complex chemical composition; shape of body fluids Raman spectra and certain spectral bands intensities are due to the contribution of molecular vibrations of several components. Therefore, the analysis of the intensity ratios of certain bands of the body fluid spectrum allows to assess homeostasis changes and to obtain the information about the functional state and possible pathologies of internal organs. The first step in our study is comparison of the obtained body fluid spectral characteristics with the spectral bands associated with components analyzed in standard biochemical analysis. Table 1 shows the observed vibrational bands of human blood Raman spectra. Here general organic components of biochemical blood test are marked with (\*) symbol; experimental peaks observed in our study are in bold type.

Table 1. Observed vibrational bands of human blood Raman spectra. Here general organic components of biochemical blood test are marked with (\*) symbol; experimental peaks observed in our study are in bold type.

Wavenumber, cm <sup>-1</sup>	Substances	References
493	Bilirubin*	[7]
<b>679</b>	<b>Creatinine*</b>	[8]
<b>756</b>	<b>L-tryptophan</b> (is part of several organic components)	[9]
<b>829</b>	<b>Collagen</b>	[10]
846	Creatinine*	[8]
<b>941</b>	<b>Protein*</b>	[11]
<b>1002</b>	<b>Protein*</b>	[11]
	<b>Urea*</b>	[8]
	<b>Hemoglobin*</b>	[12]
<b>1128</b>	<b>Glucose*</b>	[13]
<b>1225</b>	<b>L-tryptophan</b> (is part of several organic components)	[9]
<b>1336</b>	<b>Protein*</b>	[11]
	<b>Bilirubin*</b>	[7]
<b>1451</b>	<b>Protein*</b>	[11]
	<b>Bilirubin*</b>	[7]
1500	Bilirubin*	[7]
<b>1556</b>	<b>Hemoglobin*</b>	[12]
	<b>Fibrin</b>	[12]
<b>1623</b>	<b>Hemoglobin*</b>	[12]

Analysis of Table 1 allows for defining that the blood spectral characteristics recorded by the utilized experimental setup contain intensity peaks proportional to the main organic components determined by biochemical analysis. However, the homeostasis state estimation only by certain peaks provides insufficient information for the detection of the pathologies such as cancer, since the presented corresponding components have low specificity for certain cancer localization detection. In addition, the tumor cells metabolism and its interaction with the microenvironment are complex; tumor-associated metabolic changes may significantly differ for various tumors [14]. It should be stressed that human body fluids have a complex chemical composition, and the blood spectral characteristics are due to the presence of a large number components including ones with partially

overlapping spectra. Particularly, as shown in Table 1, the spectral characteristics of proteins, urea and hemoglobin simultaneously contribute to the band 1002  $\text{cm}^{-1}$ ; the intensities of the bands 1336  $\text{cm}^{-1}$  and 1451  $\text{cm}^{-1}$  are proportional to the changes of proteins and bilirubin; hemoglobin and fibrin contribute to the band 1556  $\text{cm}^{-1}$ . Moreover, the intensity of the bands 756  $\text{cm}^{-1}$  and 1225  $\text{cm}^{-1}$  is proportional to the concentration of L-tryptophan, which is part of several organic components. Thus, the detection of the position and intensity of Raman peaks may not be sufficient for detection of pathology by body fluid spectral analysis. A more detailed analysis of the body fluid component composition is required for pathology identification; therefore, it is necessary to detect and evaluate the spectral properties of those chemical components that are characteristic of a particular pathology. In this case, obtaining statistically reliable information is possible by using multidimensional processing of the full body fluid spectrum. For this purpose, the experimental data were processed on the bases of discriminant analysis method with regression on latent structures (PLS-DA). VIP allows for evaluation of individual variables from the predictors block influence on the PLS model. The higher the VIP-score of an individual variable is, the more significant it is in model construction. VIP-scores of Raman spectra matrix of blood samples are shown in Fig. 2.

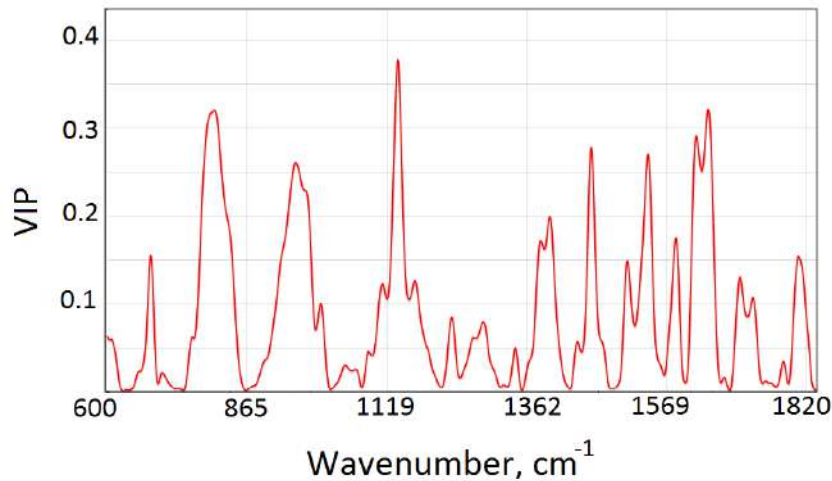


Fig. 2. VIP-scores of blood samples Raman spectra matrix.

Analysis of Fig. 2 allows for defining the most informative spectral bands for identification the features of blood Raman spectra in case lung cancer growth. The multivariate analysis of the experimental data allows to establish that the informative criteria for the lung cancer detection are changes in the Raman bands intensity 790-820  $\text{cm}^{-1}$  (glutathione), 1135-1140  $\text{cm}^{-1}$  (mannose), 946-970  $\text{cm}^{-1}$  (proteins), 1465-1475  $\text{cm}^{-1}$  (lipids, proteins) and 1640-1660  $\text{cm}^{-1}$  (proteins, phospholipids) [3].

### 3.2. Spectral characteristics of urine

Approximation curves of urine AF for tested samples are shown in Fig. 3.

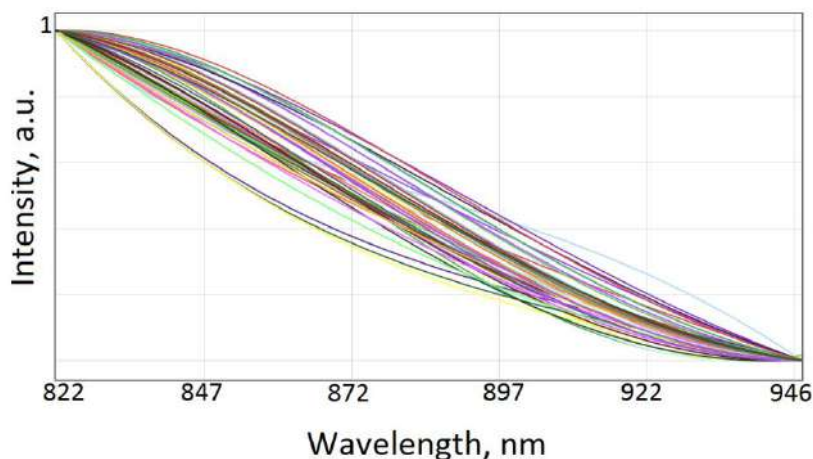


Fig. 3. Polynomial approximation of urine samples AF.

The porphyrins (nitrogen-containing pigments) accumulates in sites of active cells division and excretes with urine. Alterations in the AF urine spectrum reflect changes and metabolic imbalance of porphyrins. Therefore, the AF intensity of urine can be used as an informative criterion of oncopathology growth. The AF spectrum of porphyrins has features in red and near-infrared spectral ranges, so the excitation of the AF spectra by 785 nm laser allows to evaluate the presence of porphyrins in the tested sample. Analysis of the VIP-scores of urine samples Raman spectra matrix allows to determinate that for oncopathology the most specific changes in urine homeostasis are associated with the spectral bands 1000-1015  $\text{cm}^{-1}$  (urea),

1525-1560 cm<sup>-1</sup> (tryptophan, proteins), and 1690-1705 cm<sup>-1</sup> (pyruvate). An example use of a RS and AF combination for the analysis of body fluids in the detection of oncopathology was demonstrated in [3].

#### 4. Discussion and conclusions

Detected AF and RS spectra features may be the basis of the method for pathologies detection and become an alternative to available detection techniques of pathological conditions using laboratory methods for body fluids analysis. Besides, in future studies it is possible to analyze the correlation of the body fluids spectral characteristics and the biochemical studies results, which may allow to expand the description of the various components contribution to the observed spectral bands. In addition to studied spectral properties of urine and blood, it is also possible to use additional body fluids as research objects for non-invasive diagnostics of various pathologies. The developed approach may become the basis for the non-invasive method of cancer screening, for example, when used in Lab-on-a-chip (LOC) systems [15]. The advantages of such systems are portability, small amount of the tested sample and high efficiency. Utilizing the LOC system may simplify the current experimental setup and replace an expensive spectrometer and a cooled digital camera with a less costly portable spectrometer. Improving the recorded signal quality is also possible due to the utilizing higher quality optical elements (filters), as we used low cost Raman Probe in this research. Moreover, it is possible to increase the accuracy of detection of particular pathology by preallotment of certain markers from tested samples. This approach was demonstrated by Feng et al [16]. For this purpose it is expedient to use, for example, microfluidics technologies and chromatography. Utilizing such technologies allows for sequential chemical selection of body fluids components including the stages of sample separation into different fractions, mixing the intermediate products and their transfer to various reaction microchambers. However, in case of body fluids microdose spectra collection the recorded signal quality may decrease. Improvement of the Raman signal collection is possible with the application of Surface-enhanced Raman spectroscopy (SERS). SERS allows to achieve an improvement of registered Raman signal by several orders and is successfully used in the analysis of various biological material microdoses. Thus, the LOC system, including microfluidics and SERS technologies will improve the obtained spectra quality and increase the informativeness of the analysis. In general, the development of LOC system based on proposed method may provide the opportunity of human body fluids precise analysis for accurate screening of pathologies and detection of microorganisms in body fluids.

#### Acknowledgments

This research was supported by the Ministry of Education and Science of the Russian Federation and the program U.M.N.I.K.

#### References

- [1] Peedell C. Concise Clinical Oncology. Elsevier Health Sciences, 2005; 395 p.
- [2] Tuchin V. Handbook of Optical Biomedical Diagnostics. SPIE Press Book, 2002; 1410 p.
- [3] Shamina L, Bratchenko IA, Artemyev DN, Myakinin OO, Moryatov AA, Kaganov OI, Orlov AE, Kozlov SV, Zakharov VP. Raman and autofluorescence analysis of human body fluids from patients with malignant tumors. *J. of Biomedical Photonics & Eng.* 2017; 3(2).
- [4] Zeng H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Applied Spectroscopy* 2007; 61(11): 1225–1232.
- [5] Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics* 2015; 29(15): 528–536.
- [6] Glick D. Methods of Biochemical Analysis. John Wiley & Sons, 2009; 540 p.
- [7] Celis F, Campos-Vallette MM, Gómez-Jeria JS, Clavijo RE, Jara GP, Garrido C. Surface-enhanced Raman scattering and theoretical study of the bilichromes biliverdin and bilirubin. *Spectroscopy Letters* 2016; 49(5): 336–342.
- [8] de Almeida ML, Saatkamp CJ, Fernandes AB, Pinheiro AL, Silveira L Jr. Estimating the concentration of urea and creatinine in the human serum of normal and dialysis patients through Raman spectroscopy. *Lasers Med Sci.* 2016; 31(7): 1415–1423.
- [9] Gelder J, de Gussem K, Vandenabeele P, Moens L. Reference database of Raman spectra of biological molecules. *J. Raman Spectrosc* 2007; 38(9): 1133–1147.
- [10] Lin D, Pan J, Huang H, Chen G, Qiu S, Shi H, Chen W, Yu Y, Feng S, Chen R. Label-free blood plasma test based on surface-enhanced Raman scattering for tumor stages detection in nasopharyngeal cancer. *Scientific Reports* 2016; 11(4): 2590–2594.
- [11] Artemyev DN, Bratchenko IA, Khristoforova YuA, Lykina AA, Myakinin OO, Kuzmina TP, Davydkin IL, Zakharov VP. Blood proteins analysis by Raman spectroscopy method. Proc. SPIE 9887. Biophotonics: Photonic Solutions for Better Health Care V, 2016.
- [12] Boyd S, Bertino MF, Seashols SJ. Raman spectroscopy of blood samples for forensic applications. *Forensic Science International* 2011; 208(1-3): 124–128.
- [13] Shao J, Lin M, Li Y, Li X, Liu J, Liang J, Yao H. In Vivo Blood Glucose Quantification Using Raman Spectroscopy. *PLoS One* 2012; 7(10).
- [14] Pavlova N, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. *Cell Metabolism* 2016; 23(1): 27–47.
- [15] Ashok PC, Singh GP, Tan KM, Dholakia K. Fiber probe based microfluidic Raman spectroscopy. *J. Opt Express* 2010; 18(8): 7642–7649. DOI: 10.1364/OE.18.007642.
- [16] Feng S, Zheng Z, Xu Y, Lin J, Chen G, Weng C, Lin D, Qiu S, Cheng M, Huang Z, Wang L, Chen R, Xie S. A Noninvasive Cancer Detection Strategy Based on Gold Nanoparticle Surface-enhanced Raman Spectroscopy of Urinary Modified Nucleosides Isolated by Affinity Chromatography. *Biosensors & Bioelectronics* 2017; 91: 616–622.

# Deposition of Zinc Oxide thin film layer with the help of modified sputtering system

S.A. Fomchenkov<sup>1,2</sup>, S.D. Poletaev<sup>1,2</sup>

<sup>1</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

<sup>2</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The process of magnetron sputtering of dielectric zinc oxide (ZnO) films at a constant current source, was studied. It is demonstrated that this method of dielectric films deposition makes it possible to obtain high-quality coatings and layers that meet the requirements for creating multilayer diffractive optical elements.

*Keywords:* magnetron sputtering; thin films; diffraction optical elements; Zinc Oxide

---

## 1. Introduction

In recent years, the development of diffractive optical elements (DOE) have attracted the researchers due to the prospects of their use in optical signal and image processing systems, including in computational optics. In addition to the traditional use of DOE as spectral selectors, a significant number of DOE types have been developed so far, allowing many other functions such as multiplication and beam formation, optical signal distribution through processing channels, wave front formation, etc [1-3]. As a rule, the optical characteristics of such multilayer elements depend on many factors such as their structure, materials used and their refractive indices, the order and ratio of layer thicknesses and micro-relief, internal or surface [4,5].

The basis of such elements are optically transparent in the visible or infrared range, dielectric alternating films deposited on an optical quality substrate, for example, quartz. Various methods for the deposition of films can be used to create such structures: vacuum thermal deposition [6,7], electron-beam sputtering, and many others. But the creation of optical elements requires high quality and accuracy of the results. Therefore, the most preferred method is magnetron sputtering. This method allows the film to be sprayed at a high speed, with a low pressure of the working gas in the chamber, which allows obtaining very pure structures [8].

In this paper, we investigated the possibility of obtaining such structures using a magnetron sputtering installation with a constant current source "Caroline D12A", modernized for the purpose of sputtering dielectric targets.

## 2. The object of the study (Model, Process, Device, Sample preparation etc.)

The installation of magnetron sputtering "Caroline D12A" is designed for the deposition of conductive films, so it has a constant current source. The unit is equipped with four positions for the installation of targets, which allows the deposition of four different materials in one operating cycle. This embodiment does not permit the dispersion of dielectric films since charge accumulation occurs on the dielectric target, which contradicts the principle of magnetron sputtering. And the power supply to the magnetron results in the start-up and quick stopping of the spraying in a very short time. To solve this problem, the power supply was replaced by a high-frequency generator with an operating frequency of 13.56 MHz and a maximum output power of 1 kW. This solution allows us to accumulate a charge for one half-period of the signal being supplied and to take it off during the second half-period.

The biggest challenge during the modernization of Caroline D12 A was the development and optimization of the matching device between the generator and the magnetron. As a result of the work done, the optimal ranges and ratios of the components of the matching device, namely a tunable coil and a tunable capacitor, were selected. This allowed us to achieve an optimal alignment with allowance for unrecoverable losses (15%).

In this work, ZnO target was used for the demonstration of the sputtering process. The refractive index of the deposited layer was measured with the help of ellipsometer as shown in Fig. 1.

In the course of the work, the optimal parameters for the deposition of a zinc oxide film on a quartz substrate were obtained. The power supplied from the signal generator  $P = 500$  W, an argon flow rate  $Q(\text{Ar}) = 2.0$  l/h,  $Q(\text{O}_2) = 0.7$  l/h, the residual pressure in the chamber  $p = 5 \times 10^{-4}$  Pa, substrate heating temperature was  $t = 120$  °C, the drum rotation speed  $V = 11$  RPM. The distance from the target to the substrate was about 20 cm.

As a result, high-quality thin film of zinc oxide was obtained on quartz substrate. The scratch test indicated that the adhesion of the film to the substrate was high. The optical and physical properties of the thin film was measured with the help of films "Ellipsometer M2000DI". Consequently, the deposition rate was 10 nm/min.

For optical applications, the layers should be homogenous over the entire surface of the substrate. To study these properties, the surface of the film was examined using a Zygo NewView 7300 white light interferometer. The area of the investigated surface was  $351 \mu\text{m} \times 263 \mu\text{m}$ . Figure 2 (a) shows the resulting three-dimensional surface model. The scatter of heights along the

investigated surface is demonstrated in Figure 2 (b). The topography of the surface with an altitude indication is shown in Figure 2 (c). On the basis of surface studies using an interferometer, it can be observed that the height of the surface does not differ by more than 1 nm, which indicates a high uniformity of the film.

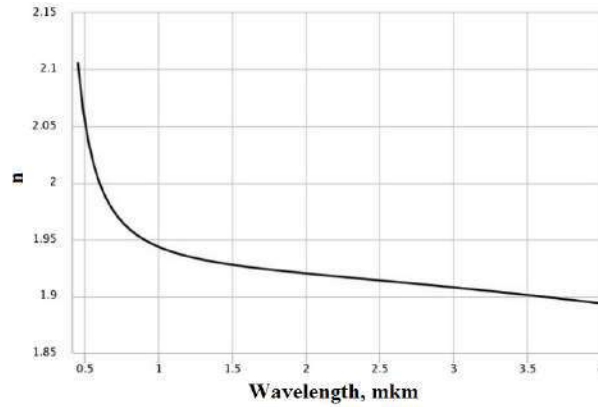


Fig. 1. The dependence of the refractive index (n) of ZnO on the wavelength.

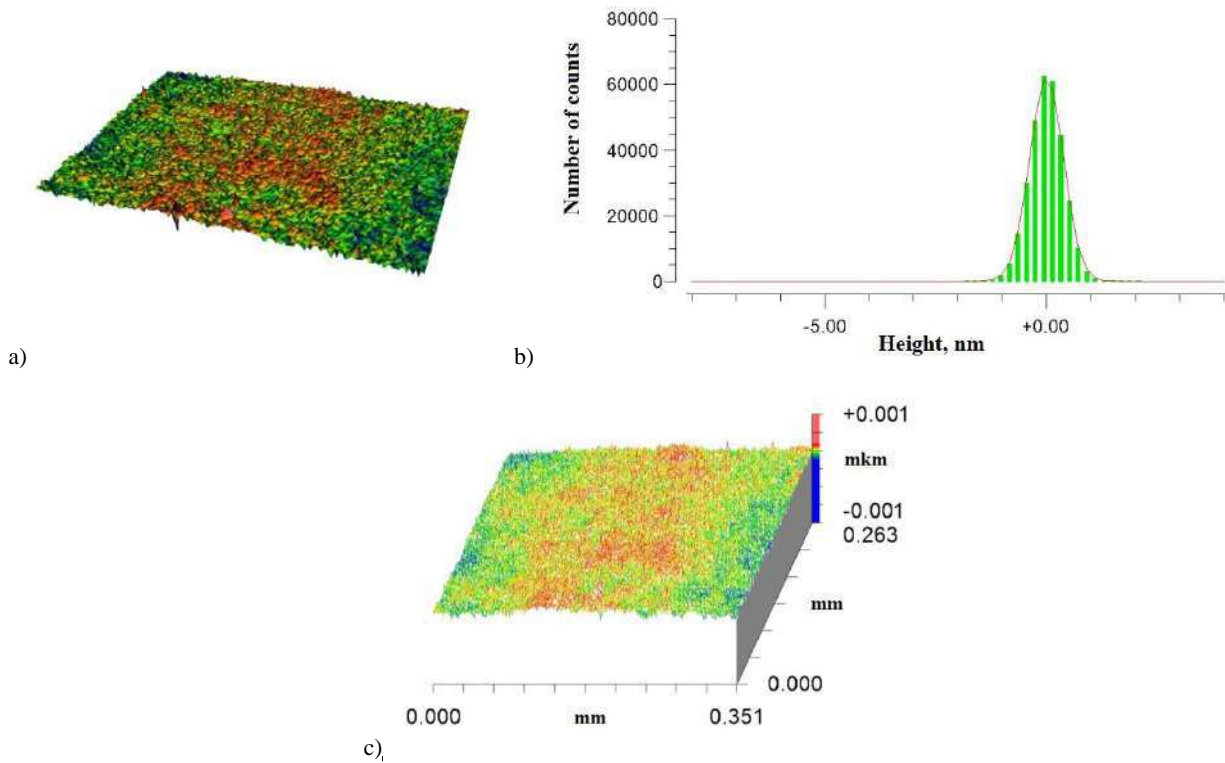


Fig. 2. The surface morphology of zinc oxide (ZnO) film measured with the help of white light interferometer, Zygo NewView 7300: (a) a three-dimensional model of the surface, (b) dispersion of the height of the indicator portion of the surface under study, (c) the surface topography.

### 3. Conclusion

In this work, the modernization of the magnetron sputtering unit was carried out. The deposition of ZnO thin film was demonstrated by using the optimized parameters of the matching device and the deposition modes.

The height of the deposited layer helped determine the deposition rate at a given mode. The layers were smooth and homogeneous which makes it possible to use ZnO along with other dielectric materials for the fabrication of multilayer diffractive optical elements.

Moreover, our aim is to fabricate optical filters with an alternating layers of dielectric thin films with a high and low refractive index.

### Acknowledgements

This work was partially funded by the Ministry of Education and Science of the and Russian Foundation for Basic Research grant No. 16-29-11744.

**References**

- [1] Danilov OB, Sidorov AI. Controllable diffractive optical elements with a vanadium dioxide film. *Journal of Technical Physics* 1999; 69(11): 91–96.
- [2] Bykov DA, Dokolovich LL. Diffraction of an optical beam on a Bragg grating with a defect layer. *Computer Optics* 2014; 38(4): 590–597.
- [3] Butt MA, Fomchenkov SA, Ullah A, Habib M, Ali RZ. Modeling of multilayer dielectric filters based on TiO<sub>2</sub> / SiO<sub>2</sub> and TiO<sub>2</sub> / MgF<sub>2</sub> for fluorescence microscopy imaging. *Computer Optics* 2016; 40(5): 674–678. DOI: 10.18287/2412-6179-2016-40-5-674-678.
- [4] Butt MA, Fomchenkov SA. Thermal effect on the optical and morphological properties of TiO<sub>2</sub> thin films obtained by annealing a Ti metal layer. *Journal of the Korean Physical Society* 2017; 70(2): 169–172.
- [5] Parinin VD, Karpeev SV, Tukmakov KN, Volodkin BO. Tunable diffraction grating with transparent indium-tin oxide electrodes on a lithium niobate x-cut crystal. *Computer Optics* 2016; 40(5): 685–688. DOI: 10.18287/2412-6179-2016-40-5-685-688.
- [6] Verma P, Pavelyev VS, Volodkin BO, Tukmakov KN, Reshetnikov AS, Andreeva TV, Fomchenkov SA, Khonina SN. Design, simulation, and fabrication of silicon-on-insulator mems vibratory decoupled gyroscope. *Computer Optics* 2016; 40(5): 668–673. DOI: 10.18287/2412-6179-2016-40-5-664-668-673.
- [7] Butt MA, Fomchenkov SA. Thermal effect on the optical and morphological properties of TiO<sub>2</sub> thin films obtained by annealing a Ti metal layer. *Journal of the Korean Physical Society*. 1: 607–612.
- [8] Fabrication of silicon slanted grating by using modified thermal deposition technique to enhance fiber-to-chip coupling (Conference Paper)



# Approximation of optical signals by the vortex eigenfunctions of the double finite Hankel transform

M.S. Kirilenko<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

This study considers a class of double finite Hankel transforms that describes the transmission of a vortex optical signal through a two-lens system with limited aperture radii in the object and spectral flats. The eigenfunctions of the given transform are calculated for different orders of vortex  $m$ . The functions obtained are an orthonormal system of functions, which can help expand an unspecified limited optical allocation with high accuracy. Approximation of optical signals without radial symmetry is also performed in this paper.

*Keywords:* vortex optical beams; eigenfunctions; finite Hankel transform; approximation

---

## 1. Introduction

The current use rate of optical fiber in terms of time and frequency characteristics tends to the bandwidth limit [1]. However, the requirements to increasing the volume of global traffic are constantly growing. For the purpose of ensuring the correspondence of communication networks to ever-growing bandwidth requirements, additional approaches are considered for the multiplexing of optical fiber channels. One of such approaches is mode division multiplexing (MDM) [2,3]. Special advantage for increasing the bandwidth of the information channel is achieved with the help of optical beams with an orbital angular momentum and an infinite number of available quantum states [4]. The significant success of such method of channels multiplexing has already been demonstrated in optical fibers [5] and in free space [6,7]. For the purpose of forming and analysis of vortex beams, diffractive optical elements are used [8,9], and lens systems are used to place them into optical fiber [10, 11].

Transmission of a vortex laser beam of the  $m^{\text{th}}$  order through the spherical lens can be described using the Hankel transform of the  $m^{\text{th}}$  order. In real lens systems, there is a spatial limit, and finite (spatial-limited) transmission operators [12, 13] are used to describe the transmission of an optical signal. Due to the spatial limit both in the object and in the spectral region, it is impossible to obtain an ideal image in a two-lens system. In order to understand how the optical signal is distorted, it is necessary to expand them according to the eigenmodes of the lens system. As such, the concept of communication modes [14, 15] is widely used. Communication modes for square apertures and Fresnel transforms are prolated angulous spheroidal functions [16,17], which are widely examined and used in optics [18-21]. Communication modes for round apertures and finite Hankel transform are circular [22] and generalized [23] spheroidal functions.

The studies [24,25] show the possibility of approximation of both one-dimensional and two-dimensional limited signals by spheroidal functions passing through the lens system without distortion.

For the purpose of superresolution development, the study [26] also examines spheroidal modes and generalized spheroidal functions, instead of which there were used the Zernike polynomials in the calculation. For the record, the Zernike polynomials have a well-defined analytical form and they are often used in analysis problems of wave front and adaptive optics [27-29]. In contrast to the Zernike basis, spheroidal functions do not have an analytic representation and they are calculated as the eigenfunctions of an operator connected with a certain optical system. Expansion in the eigenfunctions of the system makes it possible to estimate the distortion of the transmitted signal in general, i.e. to assess the quality of information transmission by the system.

This study considers the transmission of optical signals through a two-lens imaging system based on a double finite Hankel transform of the  $m^{\text{th}}$  order. We performed the calculation of an eigenfunctions set, which allows analyzing the distortion of an optical signal during transmission on the basis of approximation in functions of this set. Approximation of some signals with a good level of accuracy was performed.

## 2. Brief theoretical information

Let us consider the optical system in Figure 1. The optical beam passes through an aperture of radius  $R$  in the region  $D_1$ , at the focus distance of which the lens is located. Then in the output focal plane of the lens (in the spectral plane  $D_2$ ), it is located one more aperture of radius  $P$ . The output image of the beam is considered in the output focal plane of the second lens.

Let us consider vortex beams represented in the form:

$$f(r, \varphi) = f(r) \exp(im\varphi), \quad (1)$$

ere  $m$  is an integer number presenting the order of the optical vortex.

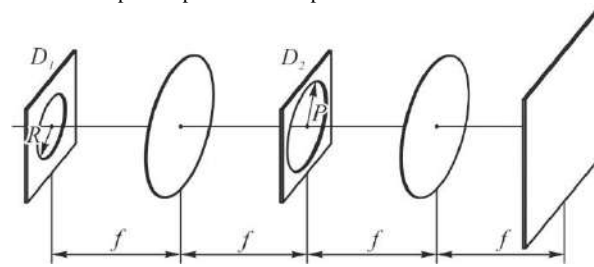


Fig. 1. Scheme of the optical system.

For vortex beams (1), the transmission of an optical signal through the two-lens system shown in Figure 1 can be written as follows:

$$H_{R,P} [f(r) \exp(im\varphi)](r', \varphi') = \frac{kP}{2\pi f} \int_0^R L(r, r'; m, P) f(r) r dr \cdot \exp(im\varphi'), \quad (2)$$

$$L(r, r'; m, P) = \int_0^{2\pi} \frac{J_1 \left( \frac{k}{f} P \sqrt{r^2 + r'^2 + 2rr' \cos \varphi} \right)}{\sqrt{r^2 + r'^2 + 2rr' \cos \varphi}} \exp(im\varphi) d\varphi, \quad (3)$$

where  $J_1$  – Bessel functions,  $k$  – wavenumber,  $f$  – focus distance of both lens. Thus, if a vortex allocation is specified at the input of a given optical system, then there will be a vortex allocation at the output, as well, while the order of the vortex  $m$  does not change.

The eigenfunctions of the expansion operator  $H_{R,P}$  represent an orthogonal system which it is used to expand optical signals that do not necessarily have a radial symmetry.

### 3. Calculation of eigenfunctions and approximation

The calculations will be performed with the following parameters:  $k / 2\pi f = 1$ ,  $R = 1$  and  $P = 5$ . Tables 1 and 2 show images of some functions and superpositions, respectively. The number  $n$  corresponds to the number of the eigenfunction in the decreasing order of the moduli of eigenvalues.

Table 1. The examples of eigenfunctions.

Indices	Amplitude	Phase
$m = 2, n = 1$		
$m = -7, n = 2$		

Table 2. The examples of superposition of eigenfunctions.

Indices	Amplitude	Phase
$m_1 = 2, n_1 = 1$ + $m_2 = -7, n_2 = 2$		

$$m_1 = 3, n_1 = 2$$

$$+$$

$$m_2 = -3, n_2 = 2$$

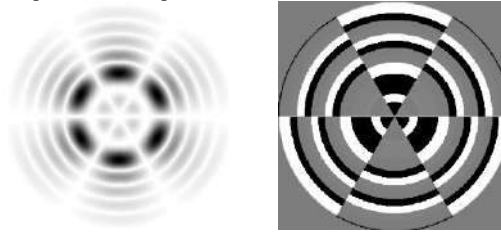

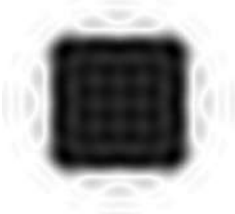






Table 3 shows which optical signals were approximated with the use of the calculated eigenfunctions. As such, there are approximation errors expressing the intensity standard deviation of an optical beam from its approximation by the eigenfunctions of the system. The expansion is performed only by those functions, which have the eigenvalue greater than 0.5 and in addition,  $|m| \leq 8$ .

Table 3. The approximation of optical signals with the eigenfunctions of the system.

Name	Signal amplitude	Approximation	Deviation
“Square”			0.0534
“Window”			0.0868
“Triangle”			0.0982

#### 4. Conclusion

The calculation of the eigenfunctions of a given optical system is a complicated computational problem, especially at large values of the order of the vortices  $m$ . The operator  $H_{R,P}$  written in the expressions (2) and (3) is self-adjoint, therefore its eigenfunctions must be real, and, consequently, their phase must be binary. However, the errors in the calculations introduce additional phase values.

Nevertheless, even in the presence of errors, the realization of the expansion of non-radial-symmetric signals along radial-vortex eigenfunctions turned out to be possible, as it was demonstrated in this study. In this case, the deviations of the approximated functions from their originals were no more than 10%.

It is worth noting that, for good approximation, the allocation of optical signals in the spatial region  $D_1$  of the first aperture should not be larger than the dimensions of the aperture itself.

#### Acknowledgement

This study was conducted with partial financial support by the Russian Foundation for Basic Research (RFBR grants 16-47-630546, 16-07-00825).

#### References

- [1] Essiambre R, Kramer G, Winzer PJ, Foschini GJ, Goebel B. Capacity limits of optical fiber networks. *J. Lightw. Technol.* 2010; 28(4): 662–701. DOI: 10.1109/JLT.2009.2039464.
- [2] Berdague S, Facq P. Mode division multiplexing in optical fibers. *Appl. Optics* 1982; 21: 1950–1955. DOI: 10.1364/AO.21.001950.

- [3] Khonina SN, Kazanskiy NL, Soifer VA. Optical Vortices in a Fiber: Mode Division Multiplexing and Multimode Self-Imaging. *Recent Progress in Optical Fiber Research*. Ed. by Dr. Moh. Yasin. Rijeka: InTech 2012; 327–352. DOI: 10.5772/2428.
- [4] Soskin MS, Vasnetsov MV. Singular optics. *Progress in Optics* 2001; 4: 219–276.
- [5] Bozinovic N, Yue Y, Ren Y, Tur M, Kristensen P, Huang H, Willner AE, Ramachandran S. Terabit-scale orbital angular momentum mode division multiplexing in fibers. *Science* 2013; 340(6140): 1545–1548. DOI: 10.1126/science.1237861.
- [6] Yan Y, Xie G, Lavery MPJ, Huang H, Ahmed N, Bao C, Ren Y, Cao Y, Li L, Zhao Z, Molisch AF, Tur M, Padgett MJ, Willner AE. High-capacity millimetre-wave communications with orbital angular momentum multiplexing. *Nature Communications* 2014; 5: 4876. DOI: 10.1038/ncomms5876.
- [7] Soifer VA, Korotkova O, Khonina SN, Shchepakina EA. Vortex beams in turbulent media: Review. *Computer Optics* 2016; 40(5): 605–624. DOI: 10.18287/2412-6179-2016-40-5-605-624.
- [8] Khonina SN, Kotlyar VV, Soifer VA, Honkanen M, Lautanen J, Turunen J. Generation of rotating Gauss-Laguerre modes with binary-phase diffractive optics. *Journal of Modern Optics* 1999; 46(2): 227–238. DOI: 10.1080/09500349908231267.
- [9] Khonina SN, Kotlyar VV, Soifer VA, Jefimovs K, Turunen J. Generation and selection of laser beams represented by a superposition of two angular harmonics. *Journal of Modern Optics* 2004; 51(5): 761–773. DOI: 10.1080/09500340408235551.
- [10] Khonina SN, Karpeev SV. Excitation and detection of angular harmonics in an optical fiber using DOE. *Computer Optics* 2004; 26: 16–26.
- [11] Karpeev SV, Khonina SN. Experimental excitation and detection of angular harmonics in a step-index optical fiber. *Optical Memory & Neural Networks (Information Optics)* 2007; 16(4): 295–300. DOI: 10.3103/S1060992X07040133.
- [12] Sneddon IN. *The Use of Integral Transforms*. New York & Boston: McGraw-Hill, 1972; 539 p.
- [13] Debnath L, Bhatta D. *Integral Transforms and their Applications*, second ed. Boca Raton, FL: Goo Chapman and Hall/CRC Press, 2007; 778 p.
- [14] Miller AR. Communicating with waves between volumes: evaluating orthogonal spatial channels and limits on coupling strength. *Applied Optics* 2000; 39(11): 1681–1699. DOI: 10.1364/AO.39.001681.
- [15] Martinsson P, Ma P, Burval A, Friberg AT. Communication modes in scalar diffraction. *Optik* 2008; 199(3): 103–111. DOI: 10.1016/j.ijleo.2006.07.009.
- [16] Slepian D, Pollak HO. Prolate spheroidal wave functions, Fourier analysis and uncertainty – I. *Bell System Technical Journal* 1961; 40(1): 43–63. DOI: 10.1002/j.1538-7305.1961.tb03976.x.
- [17] Landau HJ, Pollak HO. Prolate spheroidal wave functions, Fourier analysis and uncertainty – II. *Bell System Technical Journal* 1961; 40(1): 65–84. DOI: 10.1002/j.1538-7305.1961.tb03977.x.
- [18] Khonina SN, Kotlyar VV. Effect of diffraction on images matched with prolate spheroidal wave functions. *Pattern Recognition and Image Analysis* 2001; 11(3): 521–528.
- [19] Khonina SN, Volotovskii SG, Soifer VA. A method to compute eigenvalues of prolate spheroidal functions of zero order. *Doklady Akademii Nauk* 2001; 376(1): 30–32.
- [20] Volotovskii SG, Kazanskiy NL, Khonina SN. Analysis and development of the methods for calculating eigenvalues of prolate spheroidal functions of zero order. *Pattern Recognition and Image Analysis* 2001; 11(2): 473–475.
- [21] Khonina SN. A finite series approximation of spheroidal wave functions. *Computer Optics* 1999; 19: 65–70.
- [22] Karoui A, Moumni T. Spectral analysis of the finite Hankel transform and circular prolate spheroidal wave functions. *Journal of Computational and Applied Mathematics* 2009; 233: 315–333. DOI: 10.1016/j.cam.2009.07.037.
- [23] Yoshinobu I. Evaluation of Aberrations Using the Generalized Prolate Spheroidal Wavefunctions. *Journal of the Optical Society of America* 1970; 60(1): 10–14. DOI: 10.1364/JOSA.60.000010.
- [24] Kirilenko MS, Khonina SN. Coding of an optical signal by a superposition of spheroidal functions for undistorted transmission of information in the lens system. *Proc. SPIE* 2014; 9156: 91560J. DOI: 10.1117/12.2054214.
- [25] Kirilenko MS, Khonina SN. Calculation of eigenfunctions for imaging two-lens system with axial symmetry. *Computer Optics* 2014; 38(3): 412–417. DOI: 10.18287/0134-2452-2014-38-3-412-417.
- [26] Pich'e K, Leach J, Johnson AS, Salvail JZ, Kolobov MI, Boyd RW. Experimental realization of optical eigenmode super-resolution. *Optics Express* 2012; 20(24): 26424–26433. DOI: 10.1364/OE.20.026424.
- [27] Tyson RK. *Principles of Adaptive Optics*. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010; 314 p.
- [28] Khonina SN, Kotlyar VV, Wang Ya. Diffractive optical element matched with Zernike basis. *Pattern Recognition and Image Analysis* 2001; 11(2): 442–445.
- [29] Porfirev AP, Khonina SN. Experimental investigation of multi-order diffractive optical elements matched with two types of Zernike functions. *Proc. SPIE* 2016; 9807: 98070E. DOI: 10.1117/12.2231378.

# Development of mathematical model of laser treatment heat processes using diffractive optical elements

S.P. Murzin<sup>1</sup>, A.Yu. Tisarev<sup>1</sup>, M.V. Blokhin<sup>1</sup>, S.A. Afanasiev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

Calculation of laser beam intensity distribution in the focal plane of diffractive optical element was performed using software TracePro. To determine temperature fields occurring in the process of laser treatment of material, software of computational gas dynamics CFX version 15.0 and supercomputer "Sergey Korolev" were used. Temperature dependence of heat conductivity for the coating of the Ni-Al alloy produced by plasma spraying was determined. Alloy heat capacity was calculated based on additive rule. Besides, the temperature dependence of the absorption coefficient at CO<sub>2</sub>-laser treatment was also determined.

*Keywords:* diffractive optical elements; laser treatment; mathematical model; heat source; temperature

---

## 1. Introduction

Coating deposition on the components of gas turbine engine requires a sublayer. The sublayer is an intermediate tie coat, which compensates for differences in the coefficient of linear expansion of materials, as well as provides a higher adhesion strength. As reacting or thermoreacting nickel-aluminum powder is used as such material between the substrate and the sprayed coating. Thermally reactive powder is sprayed on the substrate by the method of plasma spraying. In this method, powder particles interact with a high-temperature plasma jet [1-3]. When the exothermically reacting powder is sprayed, its components react with the formation of new compounds and a significant amount of heat is released, that allows additionally heat the powder. To intensify the diffusion of the composite components of material, at least one component must remain in the liquid or gaseous phase.

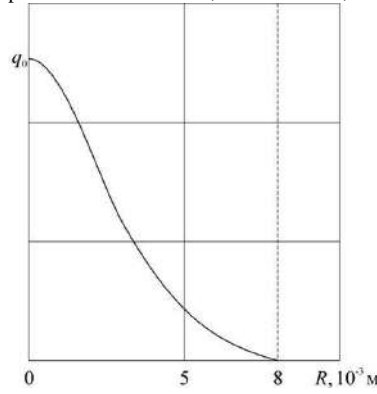
To improve coatings characteristics, the heat treatment of sprayed coating is widely used [4-6]. Heat treatment is applied to improve the coating characteristics by following factors: an increase in the contact area of the coating and the substrate; reduction in material porosity; increase in the strength of interparticle bonds [7, 8]. Laser treatment of sprayed coatings is one of the methods to improve coating properties [9-12]. Heat treatment of the sprayed coating can cause material cracking due to an increase the level of stresses during phase transformations. However, it is possible to realize laser treatment modes leading to a decrease in residual stresses. On purpose to reduce residual stresses in the absence of crack formation, the speed of laser spot moving along the substrate surface and the beam power density must be determined.

For the formation of the laser beam, various optical systems are used. However, none of them can provide the appropriate combination of such properties as the creation of the required power distribution, the concentration of the all energy of laser beam in the treatment zone of a given shape, and high reliability. The use of diffractive optical elements is promising [13-17]. Diffractive optical elements make it possible to form a predetermined beam intensity profile in the focal plane, carrying out the transformation of laser energy, chosen by calculation. The use of diffractive optical elements in the technology of laser material treatment reveals new possibilities for controlling the properties and operational characteristics of processed parts [18-21].

The aim of this work is to develop the mathematical model of laser treatment processes of the sublayer at coating deposition on the corps parts of a gas turbine engine using diffractive optical elements. It is known, temperature cycle can be a factor, which largely determines the state of processes in the treated materials. Similar approaches for mathematical model design can be used to study the formation of nanostructured materials by laser treatment.



Fig. 1. Appearance the working surface of DOE.


 Fig. 2. Beam power density distribution of a CO<sub>2</sub> slab laser Rofin DC 010.

## 2. Calculation of the laser beam intensity distribution in the DOE focal plane

The laser beam intensity distribution in the DOE focal plane has been calculated. It has the following parameters:  $f = 0.2241$  m;  $L_0 = 5.6 \cdot 10^{-3}$  m;  $R = 2.5 \cdot 10^{-2}$  m;  $r = 0.7R$ . The DOE working surface is shown in Fig. 1. The diameter of the focused beam of the Rofin DC 010 CO<sub>2</sub> slab laser is  $1.6 \cdot 10^{-2}$  m. The beam power is regulated within 10 ... 1000 W. Beam wavelength is  $\lambda = 10.6 \cdot 10^{-6}$  m. The beam power density distribution is shown in Fig 2. This distribution is close to Gaussian distribution: the beam quality parameter or beam distribution parameter is  $M^2 = 1.1$ . To change the size of laser beam, focused by DOE, it is possible to use a telescopic system of two lenses.

For calculations the software TracePro was applied. This software is designed for three-dimensional modeling of optical components surfaces, construction of the path of beams in optical systems and optical analysis. TracePro software make it possible to design the optical elements according to the equations of their surfaces. In TracePro realized the method of generalized ray tracing. Calculation of each beam incident on the surface of the optical element is performed taking into account absorption, reflection, refraction, diffraction and scattering.

The surface of the diffractive optical element is designed using the macro-language built in the TracePro software. It refers to the type of schematic programming languages that allow to compose macroprograms using loop and branch operators. To design an optical surface at  $f = 0.2241$  m;  $L_0 = 5.6 \cdot 10^{-3}$  m, the value of the polynomial was determined, in the form of which it is possible to represent the relation:

$$P_n = \frac{\int_0^{M \cos \theta} dU \int_0^{\sqrt{R^2 - U^2}} \exp\left(-\frac{U^2 + V^2}{r^2}\right) dV}{\pi f \left(g\left(\frac{R}{r}\right)\right)^*} \quad (1)$$

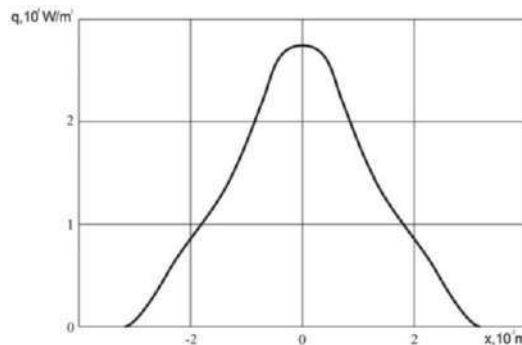
In Matlab software it was determined:

$$P_n = -2.3267 \cdot 10^{-18} \cdot x^6 + 5.7104 \cdot 10^{-15} \cdot x^5 - 5.8706 \cdot 10^{-12} \cdot x^4 + 2.4413 \cdot 10^{-9} \cdot x^3 + 1.0053 \cdot 10^{-6} \cdot x^2 + 4.4895 \cdot 10^{-5} \cdot x - 0.00086771.$$

The calculated density distribution along the axis  $Oy$  in the DOE focal plane at a beam power of  $Q = 500$  W is shown in Fig.3. We represent  $q(x, y)$  in the form of equation

$$q(\xi, \eta) = q_0 \left( a_{n_1} \bar{\xi}^{-2n} + a_{n_1-1} \bar{\xi}^{-2(n-1)} + \dots + a_2 \bar{\xi}^{-4} + a_1 \bar{\xi}^{-2} + a_0 \right) \cdot \left( b_{n_2} \bar{\eta}^{-2m} + b_{n_2-1} \bar{\eta}^{-2(m-1)} + \dots + b_2 \bar{\eta}^{-4} + b_1 \bar{\eta}^{-2} + b_0 \right) \quad (2)$$

where  $q_0$  is the power density in the center of the heat source;  $\bar{\xi} = \xi / (10^{-3} \text{ m})$ ;  $\bar{\eta} = \eta / (10^{-3} \text{ m})$  – dimensionless coordinates;  $a_{n_1}, a_{n_1-1}, \dots, a_2, a_1, a_0$ ;  $b_{n_2}, b_{n_2-1}, \dots, b_2, b_1, b_0$  – coefficients of polynomials, where  $n$  and  $m$  are integers;  $v(\xi, \eta)$  – is an additional function.


 Fig. 3. Intensity distribution along the axis  $O_y$  in the DOE focal plane at beam power  $Q = 500$  W;  $q_0 = 2,6504 \cdot 10^8$  W/m<sup>2</sup>.

We take  $a = 6.25 \cdot 10^{-3}$  m;  $b = 1.25 \cdot 10^{-3}$  m;  $n_1 = 5$ ;  $n_2 = 2$ ;  $a_5 = -4.7423 \cdot 10^{-5}$ ;  $a_4 = 1.6906 \cdot 10^{-3}$ ;  $a_3 = -2.2028 \cdot 10^{-2}$ ;  $a_2 = 0.13326$ ;  $a_1 = -0.44559$ ;  $a_0 = 1$ ;  $b_2 = 15.13$ ;  $b_1 = -7.2412$ ;  $b_0 = 1$ ;  $q_0 = 2.6504 \cdot 10^8$  W/m<sup>2</sup>,  $\nu(\xi, \eta) = 1.0$ .  $\xi = x$ ;  $\eta = y$ .

We have obtained an expression describing the power density distribution  $q(x, y)$  in the form of equation for a strip heat source at beam power of  $Q = 500$  W:

$$q(x, y) = q_0 \left( -4.7423 \cdot 10^{-5} \cdot x^{10} + 1.6906 \cdot 10^{-3} \cdot x^8 - 2.2028 \cdot 10^{-2} \cdot x^6 + 0.13326 \cdot x^4 - 0.44559 \cdot x^2 + 1 \right) \cdot \left( 15.13 \cdot y^4 - 7.2412 \cdot y^2 + 1 \right) \text{ [W/m}^2\text{];}$$

at  $q_0 = 2.6504 \cdot 10^8$  W/m<sup>2</sup>;  $x \in [-3.125 \text{ mm}; 3.125 \text{ mm}]$ ;  $y \in [-0.625 \text{ mm}; 0.625 \text{ mm}]$ .

An experimental determination of the laser beam power density distribution using DOE has been performed. To measure the power density distribution in the spot of the Rofin DC 010 CO<sub>2</sub> slab laser, a mechanical scanning method was applied. A standard power meter equipped with a square diaphragm with size of  $10^{-4} \times 10^{-4}$  m was used for these purposes. Results obtained in experimental researches correlate good with the calculated data. The relative error in determining the power density  $q$  did not exceed 5 ... 7%.

### 3. Construction of a mathematical model of the heat processes of laser treatment of the sublayer during at spraying of a triggered coating on the body parts of a gas turbine engine using DOE

To determine the temperature fields on the supercomputer "Sergey Korolev" both in the substrate and coating occurred at the laser treatment of material, software CFX 15.0 was used. To solve the problem, a finite-element model of the all coated working ring was built. To simulate cooling due to radiation and convective heat transfer, an air domain model was built. The calculation scheme is shown in Fig. 4.

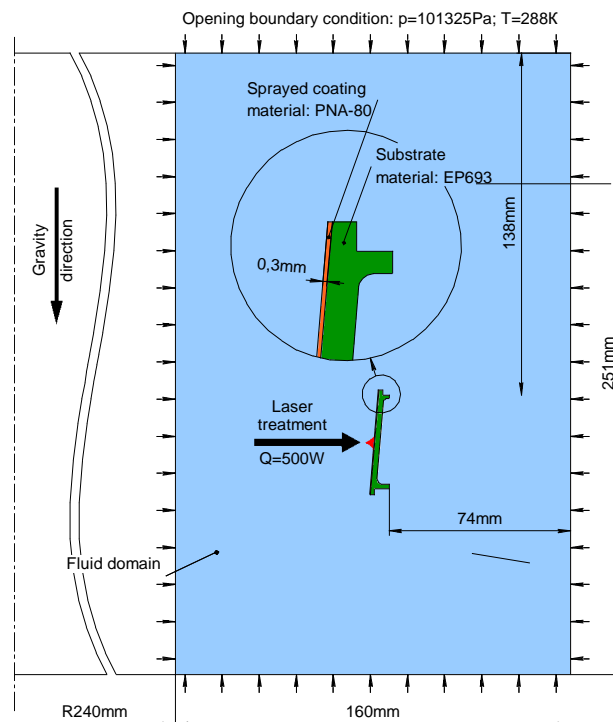


Fig. 4. Calculation scheme for determining the temperature fields in the substrate and coating.

Models of the coating and substrate for the sector at  $30^\circ$  were subdivided into hexagonal elements with an element edge size  $2.5 \cdot 10^{-4}$  m (Fig. 5a), while the remaining volume of rings consisted of tetrahedral elements. Air domain was subdivided by a tetrahedral grid as well. The region of the near-wall layer of air domain contained hexagonal elements, represented in Fig. 5b. Directly in the laser treatment area, the size of the elements of finite-volume was reduced.

Turbulence model SST was applied for these calculations. In this case the flow is characterized by low Reynolds numbers, and therefore for the correct simulation of detached flows, the values of the parameter  $y^+$  are less than 1. These values  $y^+$  were achieved by choosing the geometrical characteristics of the finite-volume elements in the near-wall layer. The radiation from the walls was taken into account by connecting the Discrete Transfer model. Radiation characteristics of the surface were described by a gray body model. The upward movement of heated air due to a reduction in its density was taken into account by the Buoyant model.

The model had the following boundary conditions. The heat flow power was 500W. Heat flow of the laser source was determined by 125 heat source points with increment  $2.5 \cdot 10^{-4}$  m. To simulate the convective cooling of the technological object, the opening air boundary was simulated in Fig. 4. In the calculation model, it was possible to change the rotational speed of the treated body. In order to determine the temperature distribution of the ring during laser treatment, the following physical properties of the coating and substrate materials have to be specified: density, heat conductivity and heat capacity. In addition, to simulate the absorption of heat flow and ring cooling, it is necessary to determine the beam characteristics of the surface.

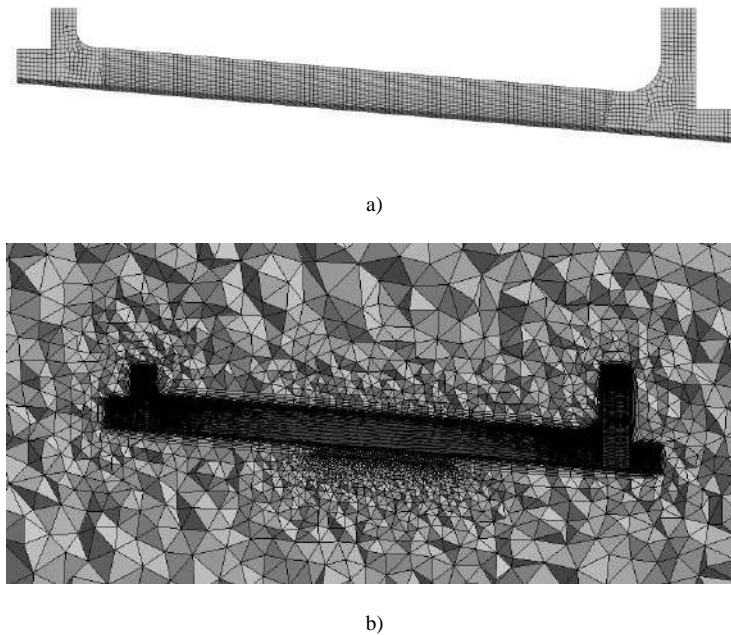


Fig. 5. Discretization of the calculated area: model of the coating and substrate (a); model of air domain (b)

At coating deposition the interaction of its constituent components is accompanied by chemical reactions that lead to the formation of material physical properties, which are different from original components. The change in the properties of materials as a function of temperature is not given with sufficient accuracy in known monographs and reference books. It led to the need for their calculation. The dependence of the heat conductivity on the temperature for coatings made of Ni-Al alloy obtained by plasma spraying in air was determined. Calculation of heat capacity Ni-Al alloy was performed based on the additivity rule. The temperature dependence of absorption coefficient under the action of CO<sub>2</sub>-laser was determined. It was necessary for estimating the amount of absorbed energy.

#### 4. Conclusion

The calculation of intensity distribution of laser beam in the focal plane of the DOE was conducted. The software complex TracePro is used for calculation, which allows to build optical elements by the equations of their surfaces. The density distribution in the DOE focal plane at beam power of 500 W was determined. An expression is obtained that describes the power density distribution in the form equation for a strip heat source. The experimental determination of the laser beam power density distribution using DOE has been performed. A CO<sub>2</sub> slab laser Rofin DC 010 was used. His the beam propagation parameter is 1.1. It is determined that the researchers of experimental studies correlate well with the calculated data. The relative error in determining the power density did not exceed 5 ... 7%.

Construction of a mathematical model of the heat processes of laser treatment of the sublayer during at spraying of a triggered coating on the body parts of a gas turbine engine using DOE is constructed. To determine temperature fields occurring in the process of laser treatment of material, software of computational gas dynamics CFX version 15.0 and supercomputer "Sergey Korolev" was used. Temperature dependence of heat conductivity for the coating of the Ni-Al alloy produced by plasma spraying was determined. Alloy heat capacity was calculated based on additive rule. Temperature dependence of the absorption coefficient at CO<sub>2</sub>-laser treatment was determined.

#### Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation as part of the Program "Research and development on priority directions of scientific-technological complex of Russia for 2014-2020" within the project RFMEFI57815X0131

#### References

- [1] Tucker RC Jr. Thermal spray coatings: Broad and growing applications. *Int. J. Powder Metall* 2002; 38(7): 45–53.
- [2] Batra U. Thermal spray coating of abradable Ni based composite. *Surf. Eng.* 2009; 25(4): 284–286.
- [3] Jin Y, Qian Z, Wang C, Yue J, Li K. Process optimization of plasma spraying Ni-based alloy coating. *Heat Treatment of Metals* 2013; 38(4): 104–108.
- [4] Molins R, Normand B, Rannou G, Hannoyer B, Liao H. Interlamellar boundary characterization in Ni-based alloy thermally sprayed coating. *Mat. Sci. Eng. A-Struct.* 2003; 351(1-2): 325–333.
- [5] Wang J-H, Friesel M, Willander M, Warren R. Microstructure of Ni-based self-fluxing alloy sprayed coating. *J. Iron Steel Res. Int.* 2005; 12(2): 56–59.



- [6] Zhang XC, Xu BS, Xuan FZ, Tu ST, Wang HD, Wu YX. Porosity and effective mechanical properties of plasma-sprayed Ni-based alloy coatings. *Appl. Surf. Sci.* 2009; 255(8): 4362–4371.
- [7] Skulev H, Malinov S, Basheer PAM, Sha W. Modifications of phases, microstructure and hardness of Ni-based alloy plasma coatings due to thermal treatment. *Surface and Coatings Technology* 2004; 185(1): 18–29.
- [8] Kromer R, Costil S, Cormier J, Courapied D, Berthe L, Peyre P, Boustie M. Laser surface patterning to enhance adhesion of plasma sprayed coatings. *Surface and Coatings Technology* 2015; 278: 171–182.
- [9] Liu F, Liu C-S, Tao X-Q, Chen S-Y. Ni-based alloy cladding on copper crystallizer surface by laser. *Dongbei Daxue Xuebao. Journal of Northeastern University* 2006; 27(10): 1106–1109.
- [10] Felgueroso D, Vijande R, Cuetos JM, Tucho R, Hernández A. Parallel laser melted tracks: Effects on the wear behaviour of plasma-sprayed Ni-based coatings. *Wear* 2008; 264(4): 257–263.
- [11] Kromer R, Costil S, Cormier J, Courapied D, Berthe L, Peyre P, Boustie M. Laser surface patterning to enhance adhesion of plasma sprayed coatings. *Surface and Coatings Technology* 2015; 278: 171–182.
- [12] Murzin SP. Formation of structures in materials by laser treatment to enhance the performance characteristics of aircraft engine parts. *Computer Optics* 2016; 40(3): 353–359. DOI: 10.18287/2412-6179-2016-40-3- 353-359.
- [13] Doskolovich LL, Khonina SN, Kotlyar VV, Nikolsky IV, Soifer VA, Uspleniev GV. Focusators into a ring. *Opt. Quant. Electron.* 1993; 25(11): 801–814.
- [14] Khonina SN, Kotlyar VV, Skidanov RV, Soifer VA. Levelling the focal spot intensity of the focused Gaussian beam. *J. Mod. Optic* 2000; 47(5): 883–904.
- [15] Doskolovich LL, Kazansky NL, Kharitonov SI, Soifer VA. A method of designing diffractive optical elements focusing into plane areas. *J. Mod. Optic* 1996; 43(7): 1423–1433.
- [16] Soifer V. *Computer Design of Diffractive Optics*. UK, USA, India, Russia: Ed., Cambridge International Science Publishing Ltd. & Woodhead Pub. Ltd., 2012; 896 p.
- [17] Kharitonov SI, Doskolovich LL, Kazanskiy NL. Solving the inverse problem of focusing laser radiation in a plane region using geometrical optics. *Computer Optics* 2016; 40(4): 439–450. DOI: 10.18287/2412-6179-2016-40-4-439-450.
- [18] Murzin SP. The research of intensification's expedients for nanoporous structures formation in metal materials by the selective laser sublimation of alloy's components. *Computer Optics* 2011; 35(2): 175–179.
- [19] Murzin SP. Local laser annealing for aluminium alloy parts. *Laser. Eng.* 2016; 33(1-3): 67–76.
- [20] Murzin SP, Balyakin VB. Microstructuring the surface of silicon carbide ceramic by laser action for reducing friction losses in rolling bearings. *Opt. Laser Technol.* 2017; 88: 96–98.
- [21] Murzin SP. Formation of nanoporous structures in metallic materials by pulse-periodic laser treatment. *Opt. Laser Technol.* 2015; 72: 48–52.

# Investigation of methods for the formation of multicolored images reconstructed with protective holograms

L.A. Nayden<sup>1</sup>, I.K. Tsyganov<sup>1</sup>, S.B. Odinkov<sup>1</sup>

<sup>1</sup>Moscow State Technical University named after N. U. Bauman, ul. Baumanskaya 2-ya, 5, 105005, Moscow, Russia

## Abstract

Dot-matrix holograms contain diffraction gratings with different periods and orientations. Grating parameters (period and orientation) are calculated according to input data, which is graphic raster file. Traditionally image uses additive color model RGB, which involves limited color range. To increase color range International Illumination Commission (ICI) color model is considered. In this paper methods for calculating the parameters of diffraction gratings with different periods and orientations for ICI graphic files are investigated and analyzed.

**Keywords:** Diffraction gratings; Colored holograms; Color chart; Dot-Matrix; Colorimetric system

## 1. Introduction

Result of the calculation dot-matrix hologram is a set of diffraction gratings. Color, displayed from hologram pixel, should be close to the color of the corresponding pixel of the input image. The following grating parameters are used to set hologram pixel color:

- The period of the holographic grating;
- Angular orientation of the holographic grating;
- Relief parameters - depth and profile type.

## 2. Methods for calculating holographic pixel parameters

The period of the holographic diffraction grating determines the wavelength of the radiation diffracted on it. The varying grating orientation angular orientation determines the angle of rotation when a certain pattern is restored. It is possible to achieve a smooth color change in the image when the hologram is rotated. Relief parameters: namely the depth and type of the profile, which determine the brightness of the radiation diffracted from a particular pixel. The brightness of diffracted radiation is determined by profile depth and type and grating area. The parameters of the diffraction gratings can be simply described in the HSB colorimetric system, where the grating period corresponds to the Hue coordinate, and the brightness of the diffracted radiation corresponds to the coordinates of the Saturation and Brightness.

The original image for rainbow holograms in most cases can be created using image editors on the computer. The result is a raster image file, in which the image is represented as a finite set of pixels [2]. Image dot form a graphic pixel with diffraction gratings located at a very small distance from each other. The perceived color of a pixel is coming result of diffraction in different holographic gratings. Pixels contain information about the color described in the RGB colorimetric system [3-4]. There are known formulas for converting color coordinates between colorimetric systems HSB and RGB. This fact allows us to calculate the parameters of holographic diffraction gratings from the input image file.

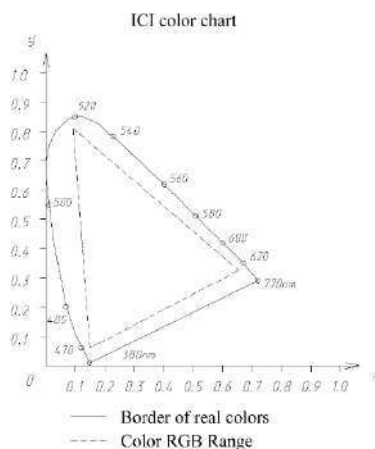


Fig. 1. ICI color chart.

Although color acquisition using the RGB system is widely used in many areas, and the RGB system can be used display a wide range of colors, but it still can not cover all possible colors in the ICI chart. To expand color range reproduction by means

of diffraction gratings, method of color reconstructing in the ICI colorimetric system (1931) is used. The ICI and RGB colormaps are illustrated on Fig. 1.

Any three different diffraction dots representing three different wavelengths can form any color within the ICI diagram. X, Y, Z are the color coordinates, the vertices of the triangle in which the color is formed. According to the ICI theory, as shown in Fig. 2, the required color is the color having the coordinates  $x_0$   $y_0$  and the intensity  $Y_0$  denoted by  $[(x_0, y_0), Y_0]$ . The three selected points have the following coordinates  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ . The intensity of the diffracted light at these three points of light will be denoted by  $Y_1, Y_2, Y_3$ . Then, knowing the coordinates of these points, through non-complex calculations, one can come to the definition of the required coordinate.

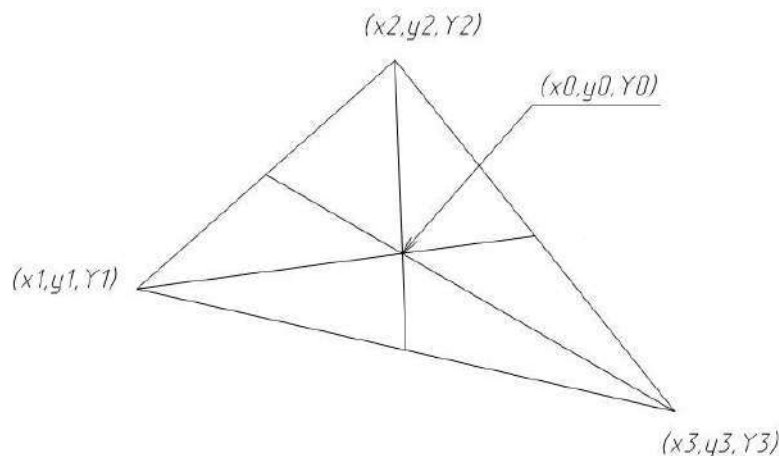


Fig. 2. Forming a color at a point inside the triangle with vertices given by three diffraction points.

The intensity of the diffracted radiation (shown in Fig. 3.) depends on the number of points with diffraction gratings, the more points, the greater the intensity, and, consequently, the brightness of the pixel.

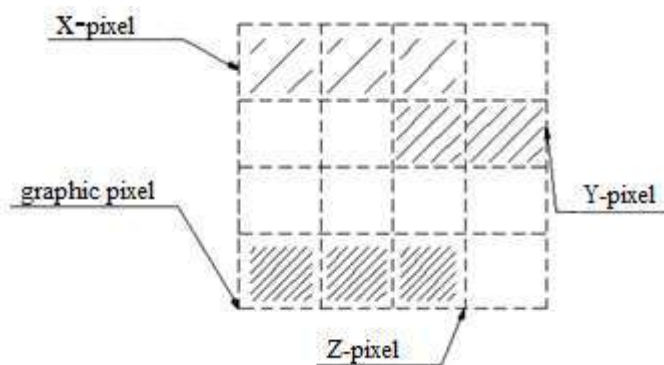


Fig. 3. Schematic representation of a graphic pixel with different intensity of its constituent elements.

The required number of points of each previously defined color depends on the geometric distance in the ICI color chart between the desired color and the three primary colors. In this case, the intensity of the reflected light corresponds to the number of points used to represent the primary color.

### 3. Conclusion

Selecting a sufficient number of diffraction points that create colors close to the border of the ICI color chart, almost all the colors located inside the diagram can be easily display by combining different diffraction pixels. Representation of the image in the ICI colormap system allows to provide a larger color spectrum than RGB.

### References

- [1] Magnusson R, Gaylord TK. Diffraction efficiencies of thin phase gratings with arbitrary grating shape. *J. Opt. Soc. Am.* 1978; 68(6): 87–93.
- [2] Frank SD. Holographic image conversion method for making a controlled holographic grating. U.S. patent 5262879 (16 nov. 1993).
- [3] Kolyuchkin VV, Zlokazov EYu, Odinkov SB, Talalaev VYe, Tsyganov IK. A coherent measurement method for checking the surface microrelief depth in holographic and diffractive optical elements. *Computer Optics* 2015; 39(4): 515–520. DOI: 10.18287/0134-2452-2015-39-4-515-520.
- [4] Khomutov VN, Poleshchuk AG, Cherkashin VV. Measurement of diffraction efficiency of DOE in many diffractive orders. *Computer Optics* 2011; 35(2): 196–201.

# On model of microstructure formation during selective laser melting of metal powder bed

F.Kh. Mirzade<sup>1</sup>, A.V. Dubrov<sup>1</sup>

<sup>1</sup>*Institute on Laser and Information Technologies – Branch of Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 140170, Shatura, Russia*

---

## Abstract

A model of phase field has been developed to investigate the microstructure evolution during selective laser melting (SLM) of metal powder bed. A two-component (degree of order, orientation field) ordering parameter has been used, for which the permissive relationships have been derived reasoning from the principle of entropy production positivity. The application of this principle has made possible obtaining the thermodynamically agreed evolutionary equations for the components of the ordering parameter, conjugate with the fields of temperature, admixture concentration and elastic deformations for the non-isothermal conditions of crystallization of pure metal melts and multicomponent alloys. The model of the microstructure is adjoint with the macroscopic thermodynamic model of SLM that accounts for the processes of heat transfer, thermo-capillary convection and evolution of the melt free surface.

*Keywords:* additive manufacturing; selectiv laser melting; powder bed; phase field method; multiscale model; microstructure; elastic stresses

---

## 1. Introduction

Powder-bed selective laser melting (SLM) is a promising metal additive manufacturing (AM) technique for producing complex structures out of powders (or their mixtures) in a layer by layer fashion. The process creates 3D solid objects by bonding powdered materials using laser beam energy [1-4]. This technology is practically non-waste and universal, as it makes use of a wide enough range of initial powders with the sizes of particles from 10 nm to 100  $\mu\text{m}$  [1, 2]. The SLM process is determined by a large number of factors, such as energy source power, scanning speed, physical and chemical properties of the initial material, etc. Noteworthy also are the interrelation of factors affecting the process and the presence of many interacting processes: absorption and scattering of laser radiation energy by the substrate matter and powder particles, heat conduction and convection, evolution of the melt free surface at the cost of capillary (thermo-capillary as well) forces, evaporation, shrinkage, crystallization, formation of the microstructure and stressed state of the synthesized object [4].

It is known from the experiments that various microstructures (cellular, dendritic, cellular-dendritic structures) emerging at the stage of melt crystallization define to a large extent the physical and mechanical characteristics of the product to be synthesized, so it is essential that the process of microstructure formation should be controlled. The best applicable basis to exert this control is mathematical simulation that allows for establishing a linkage between the SLM process parameters and the quality of the build parts.

Unified simulation of crystallization in SLM involves serious difficulties. The main problems are related to the description of the complex interaction of the nonlinear processes taking place at different scale levels, from the level of interaction between a single growing crystallite and the metastable melt to the macro-level (the description of heat-mass transfer at the level of the whole system). The existing models of crystallization process developed in the framework of one scale level are capable of describing rather complex phenomena (formation of dendrites, growth of crystalline grains, porosity, etc.) at their levels. Despite this, nevertheless, the recent new lines of investigations deal with joint models of micro- and macro-levels [5].

Our previous work [6] has suggested a multiscale model of the processes of crystallization and microstructure evolution in laser surfacing with coaxial injection of metal and alloy powders. In [7], consideration is given to the macromodel of the processes that is a part of the multilevel model of crystallization in laser surfacing. Work [8] presents a numerical multiscale model of melting of the metal powder layer for the conditions of permanent heat flows.

The present work is aimed at the development of a model of solid-phase microstructure formation in SLM of a powder bed, applying two-scale approximation. The essence of the model is that in the physico-mathematical description of the crystallization problem the physical processes are presented as a group of related processes progressing at different spatial scales and exhibiting mutual influence. At each level, a model of crystallization process is developed that allows for the features of melt behavior at this level. This fact defines the range of problems for submodels as well.

The microstructure evolution is described by the equation for the two-component (degree of order, orientation field) ordering parameter, conjugate with the equations of heat conduction and admixture diffusion, as well as with elastic stress/deformation that accompanies the phase transformation (PT). The microstructure model is adjoint with the macroscopic thermodynamic model of SLM that accounts for the processes of heat transfer, thermo-capillary convection and evolution of the melt free surface (liquid-gas interfaces). The macromodel gives self-consistent consideration to the distribution of temperature and melt velocities depending on the SLM process parameters (beam power, scanning speed, powder layer). Modeling of the free surface evolution is performed by the method of volume of fluid (VOF).

## 2. Phase field model

The investigation of microstructures in melt crystallization using the classical (Stefan-type) model presents a rather difficult task, as it calls for the development of special algorithms for an explicit definition of the shape of the interface. Best suited for this purpose is the employment of the continuum model of phase field (MPF) relying on Landau-Ginzburg principles of weakly non-equilibrium thermodynamics and formalism of PT.

In contrast to the classical model that uses the notion of a sharp boundary, the continuum model follows the concept of a diffusion interface between the liquid and solid phases. With this approach, the shape and relative position of the phases making up the microstructure are described by the variables of the phase field (or of the ordering parameter), which are governed by a set of nonlinear differential equations conjugate with the equations of heat conduction and concentration. The ordering parameter is smoothly varied over the width of the narrow transition area, describing the inner structure of PT. Away from the interface it has a constant value corresponding to the structure, orientation and their composition. Therefore, the MPF is a convenient instrument for the numerical investigation of crystallization that does not require explicit tracking of the interface in the course of the microstructure evolution; the phase boundary position is here determined as the phase field isoline.

The MPF was applied to a wide range of problems, including the growth of dendrites in pure metals; dendritic, eutectic and peritectic growth in alloys; microsegregation of the admixture on fast solidification, etc. The thermodynamically agreed MPF were considered in a number of papers [9-13]. The derivation of evolutionary equations in these works is based on the main principles of irreversible thermodynamics. To describe the melt crystallization (micro-level problems) in SLM, we represent the derivation of the dynamic equations for the MPF with regard to elastic (thermal, concentration, and phase) stresses accompanying the process of non-isothermal PT.

To obtain the governing equation of the MPF, make use of the formalism suggested in [9]. Consider an arbitrary region having volume  $V$ , where the binary metal ( $a$ - $b$ ) undergoes the liquid-solid (L-S) PT. Restrict ourselves to the 2D variant of the problem and introduce the characteristic of the material phase state – the ordering parameter consisting of two variables  $\{\varphi(\mathbf{r}, t), \phi(\mathbf{r}, t)\}$  [14]. The variable  $\varphi(\mathbf{r}, t)$  can be interpreted as the degree of the material ordering in the microvolume with the radius-vector  $\mathbf{r}$  at the moment of time  $t$ ;  $\varphi = 0$  corresponds to the liquid state, and  $\varphi = 1$  – to the crystalline state. The narrow region, where  $0 < \varphi < 1$ , corresponds to the phase interface. The  $\phi$  variable describes the crystalline phase orientation (crystallization orientation field). It is determined as  $\phi = N_0 \phi$ , where  $\phi$  is angle between one of the main crystallographic directions and the X-axis in the chosen coordinate system,  $N_0$  is the order of the symmetry axis of the grating type under study. It is apparent that  $\phi \in (0, 2\pi/N_0)$ .

For an arbitrary subvolume of the region  $\Omega \in V$  under consideration the functional of total entropy will be written as

$$E(\Omega) = \int_{\Omega} \left[ \eta(\varphi, c, u) - \frac{1}{2} \varepsilon^2 |\nabla \varphi|^2 - \frac{1}{2} \nu^2 |\nabla \phi|^2 \right] dV \quad (1)$$

where  $\eta(\varphi, c, u)$  is the entropy density;  $u(\mathbf{r}, t)$  is the internal energy density;  $c$  is the concentration of the dissolved matter (admixture);  $\varepsilon$  and  $\nu$  are the positive parameters, which can be the functions of the ordering parameter. The gradient terms in (1) allow for the contributions to entropy at the cost of the interphase boundaries. In the employed functional, the gradients of ordering parameters are only taken into account; the gradients of temperature and concentration in the explicit form are not included (they are supposed to be smallish).

The anisotropy of the surface energy can be taken into consideration, supposing that the coefficient  $\varepsilon$  is a function of  $\theta$ ,  $\varepsilon = \varepsilon_0 \varepsilon(\theta)$  where  $\varepsilon_0 > 0$ ,  $\theta = \theta_0 - \phi / N_0$ . The variable  $\theta_0$  is the angle between the X-axis and the normal vector  $\mathbf{n} = \nabla \varphi$  at the interface. Thus,  $\theta$  characterizes the orientation of the interface normal vector in relation to the neighboring growing crystallite. The third term in (1) takes account of the influence of misorientation of the neighboring crystals. It is supposed that this influence depends on the material ordering, so for  $\nu$  we have  $\nu = \nu_0 \nu(\varphi)$ ,  $\nu_0 > 0$ .

Applying the local conservation laws for concentration and energy, as well as the second law of thermodynamics, obtain

$$u + \nabla \cdot \mathbf{q} - \sigma_{ij} e_{ij} = 0, \quad (2)$$

$$c + \nabla \cdot \mathbf{j} = 0, \quad (3)$$

where  $\mathbf{j} = M_c \nabla (\delta E / \delta c)$  and  $\mathbf{q} = M_u \nabla (\delta E / \delta u)$  are the fluxes of concentration and energy, respectively;  $M_{c,u} > 0$  are the constants describing the dissolved substance diffusion and heat conduction;  $\sigma_{ij}$  and  $e_{ij}$  are the tensors of stresses and deformations, respectively. The third term in (2) characterizes the variation of inner energy due to elastic deformation.

Entropy production in the volume  $\Omega \in V$  can be calculated by subtracting the entropy flow through the surface from the rate of entropy variation  $E$  in  $\Omega \in V$ :

$$G_{prod} = E + \int_A \left( \frac{\mathbf{q}}{T} + \mathbf{p} \right) \cdot \mathbf{n} da, \quad (4)$$

where  $A$  is the surface of  $\Omega$  with the outer normal  $\mathbf{n}$ ,

$$\mathbf{p} = \varphi \left[ \varepsilon^2 \nabla \varphi + \varepsilon \varepsilon' I \cdot \nabla \varphi \right] + \phi \nu \nabla \phi$$

( $I$  is the unit tensor,  $\varepsilon' = d\varepsilon(\theta)/d\theta$ ). In the integrand in (4),  $\mathbf{q}/T$  is the entropy flow due to heat conduction,  $\mathbf{p}$  is the entropy flux related to changing of phase variables (degree of order, orientation of the growing crystal) at the boundary of  $\Omega$  volume.

By substituting expression (1) into equation (4) and applying the divergence theorem, find for the second law of thermodynamics ( $G_{prod} \geq 0$ )

$$\int_{\Omega} \left[ \eta + \nabla \cdot \left( \frac{\mathbf{q}}{T} \right) + h_{\varphi} \varphi + h_{\phi} \phi \right] dv \geq 0, \quad (5)$$

where the following designations are introduced:

$$h_{\varphi} = \nabla \cdot \left[ \varepsilon (I + \varepsilon' J) \nabla \varphi \right] - \nu \nu' |\nabla \phi|^2,$$

$$h_{\phi} = \varepsilon \varepsilon' |\nabla \phi|^2 + \nabla \cdot (\nu \nabla \phi)$$

( $J = \mathbf{i}\mathbf{i} + \mathbf{j}\mathbf{j}$  is the tensor with the orthonormal basis ( $\mathbf{i}, \mathbf{j}$ ) in the Cartesian coordinate system,  $\nu' = d\nu(\varphi)/d\varphi$ ). Hence, we have the local expression

$$\eta + \nabla \cdot \left( \frac{\mathbf{q}}{T} \right) + h_{\varphi} \varphi + h_{\phi} \phi \geq 0. \quad (6)$$

Further, applying the law of conservation of energy (2) and using Gibbs equation for free energy  $g = u - T\eta - \sigma_{ij} e_{ij}$ , write inequality (6) in the form:

$$-g - \eta T - \left( \frac{\mathbf{q}}{T} \cdot \nabla T \right) + h_{\varphi} \varphi + h_{\phi} \phi - e_{ij} \sigma_{ij} \geq 0. \quad (7)$$

The time derivative of free energy ( $g$ ) is represented as

$$g = \frac{\partial g}{\partial T} T + \frac{\partial g}{\partial \sigma_{ij}} \sigma_{ij} + \frac{\partial g}{\partial \varphi} \varphi$$

Then inequality (7) takes the form:

$$\left( h_{\varphi} T - \frac{\partial g}{\partial \varphi} \right) \varphi + h_{\phi} \phi - \left( \eta + \frac{\partial g}{\partial T} \right) T - \left( \frac{\mathbf{q}}{T} \cdot \nabla T \right) + \left( e_{ij} + \frac{\partial g}{\partial \sigma_{ij}} \right) \sigma_{ij} \geq 0. \quad (7)$$

The positivity of entropy production can be locally assured having chosen the following relationships for the thermal flow and the time derivatives of variables of the ordering parameter:

$$\mathbf{q} = M_e \nabla \frac{1}{T}, \quad (8)$$

$$\tau \varphi = \nabla \cdot \left[ \varepsilon (I + \varepsilon' J) \nabla \varphi \right] - \frac{\nu'}{2} |\nabla \phi|^2 - \frac{1}{T} \frac{\partial g}{\partial \varphi}, \quad (9)$$

$$\tau \phi = \varepsilon \varepsilon' |\nabla \phi|^2 + \nabla \cdot (\nu \nabla \phi), \quad (10)$$

where  $\eta = -\partial g / \partial T$ ,  $e_{ij} = -\partial g / \partial \sigma_{ij}$ ,  $\tau = \tau_0 \tau(\theta)$  is the function describing mobility.

Equations (2), (3) and (8)-(10) represent the set of master equations for the phase field, admixture concentration and energy.

For the densities of Gibbs free energy ( $g(\varphi, c, T)$ ) and internal energy ( $u(\varphi, c, T)$ ) in case of regular binary alloys we have the expressions [10]

$$g(\varphi, c, T) = (1-c)g_a(\varphi, T, e_{ij}) + cg_b(\varphi, T, e_{ij}) + \lambda(\varphi)c(1-c) + \frac{RT}{v_m} [c \ln c + (1-c) \ln(1-c)], \quad (11)$$

$$u(\varphi, c, T) = (1-c)u_a(\varphi, T) + cu_b(\varphi, T), \quad (12)$$

where  $g_{a,b}$  and  $u_{a,b}$  are the classical densities of free energy and inner energy of the substances  $a$  and  $b$ , respectively;  $R$  is the gas constant;  $v_m$  is the molar volume;  $\lambda(\varphi)$  is the alloy imperfection parameter. The densities of inner energies of the substance  $a$  and  $b$  are written as

$$u_{a,b}(\varphi, T) = p(\varphi)u_{a,b}^S(T) + u_{a,b}^L(T)(1-p(\varphi)),$$

where  $u_{a,b}^{S,L}(T)$  are the internal energies of the solid and liquid phases of the substances  $a$  and  $b$  at the temperature  $T$ ; the interpolation function  $p(\varphi)$  determines the dependence of the internal energy on the medium order. It is chosen in such a way as to offer a description of the interface  $L-S$  of a finite width, where  $0 < \varphi < 1$  (the free energy potential has its minima at  $\varphi = 0$  and  $\varphi = 1$ ). In accordance with [9]:  $p(\varphi) = \varphi^3(10 - 15\varphi + 6\varphi^2)$ ,  $p(0) = 0$  and  $p(1) = 1$ .

Further, representing  $u_{a,b}^{S,L}(T)$  as the linear dependences on the temperature:  $u_{a,b}^{S,L}(T) = u_{a,b}^{S,L}(T_m^{a,b}) + C_{a,b}^{S,L}(T - T_m^{a,b})$ , where  $u_{a,b}^S(T_m^{a,b})$  are the inner energies of the solid and liquid phases at the melting temperature  $T = T_m^{a,b}$ ,  $C_{a,b}^{S,L}$  are their heat capacities, we have

$$u_{a,b}(\varphi, T) = u_{a,b}^S(T_m^{a,b}) + C_{a,b}(T - T_m^{a,b}) + p(\varphi)L_{a,b}$$

( $L_{a,b} = u_{a,b}^S(T_m^{a,b}) - u_{a,b}^L(T_m^{a,b}) = T_m^{a,b} \eta_{a,b}^S - \eta_{a,b}^L$  is the latent heat of the components,  $C_{a,b}^S = C_{a,b}^L = C_{a,b}$ ). Thereafter, using the thermodynamic relation  $dg = -sdT + \sigma_{ij} de_{ij}$  for the densities of free energies of the components  $g_{a,b}$  after integrating obtain

$$g_{a,b} = \omega_g^{a,b} T d(\varphi) + \left[ L_{a,b} (1 - T/T_m^{a,b}) + \mathcal{G}_{a,b}(e_{ij}) \right] p(\varphi). \quad (13)$$

Here,  $\omega_g^{a,b} = 3\bar{\sigma}_{a,b}/\sqrt{2} T_m^{a,b} \delta_{a,b}$  ( $\bar{\sigma}_{a,b}$  is the surface energy of the  $L$ - $S$  boundary;  $\delta_{a,b}$  is the PT front width (the typical scale of the phase field length) is the height of the energy barrier related to the interphase boundary ( $L$ - $S$ );  $d(\varphi)$  is the double-wall potential. The term  $\mathcal{G}_{a,b}(e_{ij})$  in the right side of (13) describes the effect of the elastic fields of deformations on the potential, which is due to PT:

$$\mathcal{G}_{a,b}(e_{ij}) = \int_0^{e_{ij}} (\sigma_{ij}^S - \sigma_{ij}^L) de_{ij}.$$

Then, making use of the thermodynamic relation  $(\partial\eta/\partial\varphi)_{u,c} = -T^{-1}(\partial g/\partial\varphi)_{T,c}$  find that the gradient of entropy density is

$$\partial\eta/\partial\varphi = -(1-c)\Gamma_a - c\Gamma_b - T^{-1}\lambda'c(1-c), \quad (14)$$

where  $\Gamma_{a,b} = \omega_g^{a,b} d'(\varphi) + 30d(\varphi)T^{-1} \left[ T_m^{a,b} (1 - T/T_m^{a,b}) + \mathcal{G}_{a,b}(e_{ij}) \right]$ ,  $\lambda' = d\lambda(\varphi)/d\varphi = 30(\lambda_L - \lambda_S)d(\varphi)$ . In deriving (14), we have taken into account that  $\lambda(\varphi) = \lambda_S + (\lambda_L - \lambda_S)p(\varphi)$  and  $p'(\varphi) = 30d(\varphi)$ .

We have the following expressions for the energy flows ( $\mathbf{q}$ ) and impurity concentration ( $\mathbf{j}$ ), respectively:

$$\mathbf{q} = -\frac{M_e}{T^2} \nabla T, \quad (15)$$

$$\mathbf{j} = D(\varphi) \frac{c(1-c)v_m}{R} \left[ (\Gamma_a - \Gamma_b - T^{-1}\lambda'(1-2c)) \nabla\varphi - p(\varphi)(\mathcal{G}'_a - \mathcal{G}'_b) \nabla e_{ij} \right] - D(\varphi) \left[ 1 - \frac{2c(1-c)v_m}{RT} \lambda(\varphi) \right] \nabla c, \quad (16)$$

where  $\mathcal{G}'_{a,b} = d\mathcal{G}_{a,b}/de_{ij}$ ;  $D(\varphi) = D_S + p(\varphi)(D_L - D_S)$  is the diffusion coefficient that is linked to the  $M_c$  parameter by the relationship:  $M_c = D(\varphi)v_m c(1-c)R^{-1}$ .

Substituting (11), (12) and (13)-(16) into (2), (3) and (9), obtain the final forms of the governing equations for the phase field

$$\tau\varphi = \varepsilon_0^2 \nabla \cdot \left[ \varepsilon(\theta) (\varepsilon(\theta) I + \varepsilon'(\theta) J) \cdot \nabla\varphi \right] - \nu_0 \nu(\varphi) \nu'(\varphi) |\nabla\varphi|^2 - (1-c)\Gamma_a(\varphi, T, e_{ij}) - c\Gamma_b(\varphi, T, e_{ij}) - T^{-1}\lambda'c(1-c), \quad (17)$$

orientation field

$$\tau\phi = \varepsilon_0^2 \varepsilon(\theta) \varepsilon'(\theta) |\nabla\varphi|^2 + \nu_0 \nabla \cdot (\nu(\varphi) \nabla\phi), \quad (18)$$

energy

$$CT = \chi \nabla^2 T - \left[ p'(\varphi) L + \omega_e d'(\varphi) \right] + 3\kappa T \alpha_T \delta_{ij} e_{ij}, \quad (19)$$

and concentration

$$c = \nabla \cdot D \left[ 1 - \frac{2c(1-c)v_m}{RT} \lambda(\varphi) \right] \nabla c + \nabla \cdot \left[ M_c (\Gamma_b - \Gamma_a + T^{-1}\lambda'(1-2c)) \nabla\varphi + p(\varphi)(\mathcal{G}'_a - \mathcal{G}'_b) \nabla e_{ij} \right], \quad (20)$$

where  $\chi = M_e T^{-2} = (1-c)\chi_a + c\chi_b$  is the heat conduction,  $C = (1-c)C_a + cC_b$ ,  $L = (1-c)L_a + cL_b$ .

If the thicknesses of the interphase boundaries of the binary alloy components are  $\delta_a = \delta_b = \delta$ , the phase field equation (17) is much simplified and takes the form:

$$\tau\varphi = \varepsilon_0^2 \nabla \cdot \left[ \varepsilon(\theta) (\varepsilon(\theta) I + \varepsilon'(\theta) J) \cdot \nabla\varphi \right] - \nu_0^2 \nu(\varphi) \nu'(\varphi) |\nabla\varphi|^2 - \varepsilon_0^2 \left[ \frac{1}{\delta^2} \xi_1(c) p'(\varphi) + \xi_2(c, T, e_{ij}) d'(\varphi) \right] - T^{-1}\lambda'c(1-c),$$

where

$$\varepsilon_0^2 = 6\sqrt{2}\delta(\bar{\sigma}_a + \bar{\sigma}_b)(T_m^a + T_m^b)^{-1}, \quad \xi_1(c) = (1-c)\xi_1^a + c\xi_1^b, \quad \xi_2(c, T) = (1-c)\xi_2^a(T) + c\xi_2^b(T),$$

$$\xi_1^{a,b} = \frac{2\bar{\sigma}_{a,b}}{\bar{\sigma}_a + \bar{\sigma}_b} (1 + T_m^{b,a}/T_m^{a,b}), \quad \xi_2^{a,b}(T, e_{ij}) = \frac{1}{\varepsilon_0^2 T} \left[ L_{a,b} (1 - T/T_m^{a,b}) + \mathcal{G}_{a,b}(e_{ij}) \right].$$

In this case, equation (20) for the concentration is reduced as follows

$$c = \nabla \cdot D \left[ 1 - \frac{2c(1-c)v_m}{RT} \lambda(\varphi) \right] \nabla c + \nabla \cdot \left[ m_0 \left( \frac{1}{\delta} \xi_1'(c) p'(\varphi) + 30\delta \xi_2'(c, T, e_{ij}) p(\varphi) + T^{-1}\lambda'(1-2c) \right) \nabla\varphi + p(\varphi)(\mathcal{G}'_a - \mathcal{G}'_b) \nabla e_{ij} \right],$$

where

$$m_0 = \frac{6\sqrt{2}v_m(\bar{\sigma}_a + \bar{\sigma}_b)}{R(T_m^a + T_m^b)}, \quad \xi_1'(c) = d\xi_1(c)/dc, \quad \xi_2'(c) = d\xi_2(c)/dc.$$

The fields of elastic strains can be expressed in terms of the phase field relying on the condition of mechanical equilibrium:

$$\nabla_j \sigma_{ij} = 0, \quad \sigma_{ij} = \sigma_{ij}^S p(\varphi) + \sigma_{ij}^L (1 - p(\varphi)). \quad (21)$$

The stress tensors of mono phases are represented as

$$\sigma_{ij}^{S,L} = \lambda^{S,L} \delta_{ij} \nabla \cdot \mathbf{u} + 2\mu^{S,L} e_{ij} - 3\kappa \delta_{ij} [\alpha_c^{S,L} (c - c_0) + \alpha_T^{S,L} (T - T_0) + \gamma_v], \quad (22)$$

where  $\lambda^{S,L}$  and  $\mu^{S,L}$  are Lamé moduli of elasticity,  $\kappa^{S,L} = \lambda^{S,L} + 2\mu^{S,L}/3$  is the isothermal compression modulus,  $\alpha_c^{S,L}$  and  $\alpha_T^{S,L}$  are the coefficients of volumetric concentration and thermal expansions, respectively. In (22), the last term including  $\gamma_{vol}$  allows for the stresses generated because of the difference in the volumes of  $L$  and  $S$  phases.  $\nabla \cdot \mathbf{u} = e_{ii} = \partial u_i / \partial x_i$  ( $\mathbf{u}$  is the vector of elastic displacement).

$$\begin{aligned} \sigma_{ij} = & \kappa(\varphi)(\nabla \cdot \mathbf{u}) \delta_{ij} + 2\mu(\varphi)(u_{ij} - \frac{1}{d} \delta_{ij} \nabla \cdot \mathbf{u}) - 3\kappa^l \delta_{ij} [\alpha_c(\varphi)(c - c_0) \\ & + \alpha_T(\varphi)(T - T_0) + \gamma_v(\varphi)\kappa_0], \end{aligned} \quad (22a)$$

where

$$\kappa(\varphi) = \kappa^l + p(\varphi)\Delta\kappa, \quad \mu(\varphi) = \mu^l + p(\varphi)\Delta\mu, \quad \alpha_{c,T}(\varphi) = \alpha_{c,T}^L + p(\varphi)\Delta\alpha_{c,T}, \quad \gamma_v(\varphi) = \gamma_0 p(\varphi),$$

$$\Delta\kappa = \kappa^S - \kappa^L, \quad \Delta\mu = \mu^S - \mu^L, \quad \Delta\alpha_{c,T} = \kappa_0 \alpha_{c,T}^S - \alpha_{c,T}^L, \quad \kappa_0 = \kappa^S / \kappa^L.$$

Under certain assumptions, several existing models of the phase field can be produced from the obtained micromodel. For instance, the removal of the equations for orientation field, diffusion and stresses from the model (17)-(21), reasoning that  $\phi = 0$ ,  $c = 0$  and  $\sigma_{ij} = 0$ , will result in producing the model of pure substance crystallization (Wang and coauthors [9]). If  $T = const$ ,  $\phi = 0$ ,  $\sigma_{ij} = 0$ , as well as  $\lambda(\varphi) = 0$  (an ideal alloy), the Warren and Boettinger [10] model for isothermal crystallization of binary alloys will be obtained.

In the case that the elastic properties of the liquid and solid phases on PT are identical ( $\Delta\kappa = \Delta\mu = 0$ ), and the variations of the temperature and admixture concentration are insignificant ( $T = T_0$ ,  $c = c_0$ ), the phase field model is considerably simplified, and for 2D systems, it takes the form:

$$\begin{aligned} \tau\varphi = & \varepsilon_0^2 \nabla \cdot [\varepsilon(\theta)(\varepsilon(\theta)I + \varepsilon'(\theta)J) \cdot \nabla \varphi] - \nu_0 \nu(\varphi) \nu'(\varphi) |\nabla \varphi|^2 \\ & - \omega_g d'(\varphi) + p'(\varphi) T_0^{-1} \kappa \gamma_0 (\nabla \cdot \mathbf{u}), \end{aligned} \quad (23)$$

$$\kappa \nabla_i (\nabla \cdot \mathbf{u} - \gamma_0 p(\varphi)) + \mu \nabla_j \nabla_j u_i = 0. \quad (24)$$

The equation for the orientation field ( $\theta$ ) remains unchanged. Fourier transform permits rewriting equation (24) as

$$\kappa k_i k_j \hat{u}_j - i \gamma_0 k_i \hat{p} + \mu k_j k_j \hat{u}_i = 0.$$

Then we sum ( $\sum_i$ ) both the parts of this equation after prior multiplying by  $k_i$ . As a result, find the following expression for the Fourier components of elastic displacement:

$$-i k_j \hat{u}_j = \frac{\gamma_0 \sum_i (k_i^2 \hat{p} / \mu k_i k_i)}{\kappa \sum_i (k_i^2 / \mu k_i k_i) + 1}.$$

This solution makes possible the exclusion of the displacement field from the phase field equation (23).

Consider now the isotropic 1D definition of the problem (23) and (24) under the conditions:  $\varphi = 0$ ,  $c = c_0$ . Suppose also that the orientation field of the whole volume is uniform ( $\phi = \phi_0$ ) and constant, and the interface orientation is  $\theta = \theta_0$ . In this case the phase field  $\varphi = \varphi(z)$  describes the 1D two-phase region where the melt (with  $\varphi \approx 0$ ) corresponds to  $z \rightarrow \infty$ , and the crystal (with  $\varphi \approx 1$ ) agrees to  $z \rightarrow -\infty$ . It is anticipated that the interphase region where  $\varphi$  is varied between 0 and 1 is located near  $z = 0$ .

For the conditions of the imposed two-axis deformation  $e_{xx} = e_{yy} = e_0$  we have from (22) for the stress tensor components ( $\sigma_{xx}$ ,  $\sigma_{yy}$ ):

$$\sigma_{xx} = \sigma_{yy} = \frac{18\kappa\mu}{3\kappa + 4\mu} (e_0 - \gamma_0).$$

Accordingly, the following expression can be written for the elastic energy potential:

$$f_{el} = \frac{18\kappa\mu}{3\kappa + 4\mu} (e_0 - \gamma_0)^2.$$

Hence, the following modification of MPF can be obtained from (17)-(20):

$$\varepsilon_0^2(\theta_0) \frac{d^2 \varphi}{dx^2} = \frac{df}{d\varphi}, \quad (25)$$

where



$$f = \omega T d(\varphi) + \left[ L(1 - T/T_m) + \frac{18\kappa\mu}{3\kappa + 4\mu} (e_0 - \gamma_0)^2 \right] p(\varphi).$$

After that multiplying both the sides of (25) by  $d\varphi/dx$  and performing one integration, derive

$$\frac{1}{2} \varepsilon_0^2 \left( \frac{d\varphi}{dx} \right)^2 - f(\varphi, T) = f(0, T) = f(1, T). \quad (26)$$

From (26) follows the expression for the equilibrium temperature

$$\frac{L(T - T_0)}{T_0} = - \frac{2\mu^S (3\lambda^S + 2\mu^L)}{\lambda^S + 2\mu^L} (e_0 - \gamma_0)^2. \quad (27)$$

Equation (27) characterizes the influence of elastic fields on the equilibrium temperature. The solution of (26) with the boundary condition  $f(0, T) = 0$  has the form:

$$x - x_0 = \sqrt{\varepsilon_0} \int_{1/2}^{\varphi} \frac{d\varphi'}{\sqrt{2f(\varphi', T)}}.$$

### 3. Macro-scale model

The macro-scale model of SLM describes the dynamics of variation of the macroscopic fields of the temperature, velocities, pressure, as well as the evolution of the melt free surface. The knowledge of these fields is necessary in solving the microproblem to define the phase fields during crystallization. Fig. 1 shows the schematic diagram of the SLM. In the formulation of the macro-scale model the following assumptions have been made: the Gaussian and “top-hat” distributions of laser beam intensity are considered; laser radiation, when absorbed in the powder layer on the substrate, generates a microscopic region of melt having a certain depth and width; it also induces the emergence of surface forces causing the melt motion owing to the thermocapillary effect at the cost of the temperature gradient; consideration is given to the radiation intensities ( $J_0$ ) whereby evaporation of the powder particles is practically absent. Since metals are intensively evaporated at the temperatures  $T > T_v$ , where  $T_v$  is the temperature of metal evaporation (at atmospheric pressure), evaporation-free regimes are obtainable over a wide range of temperatures  $T_m < T < T_v$ . Newtonian liquids are considered; all physical properties of the liquid except surface tension do not depend on the temperature.

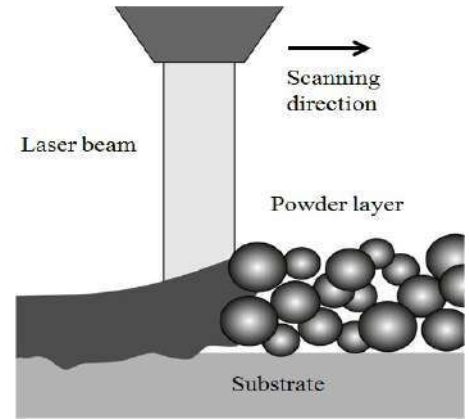


Fig. 1. SLM of a powder bed

The macro-scale model involves the coupled equations of: continuity

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \quad (28)$$

Navier-Stokes

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \nabla (\rho \mathbf{v} \mathbf{v}) - \text{div}(\mu \nabla \mathbf{v}) = -\nabla P - \frac{\mu}{K} \mathbf{v}, \quad (29)$$

energy transfer

$$\frac{\partial (\rho h)}{\partial t} + \nabla \cdot (\rho \mathbf{v} h) = \nabla \cdot (k \nabla T) - \nabla \cdot (\rho f_s (h_L - h_s) \mathbf{v}) + Q_{las}, \quad (30)$$

where  $\mathbf{v} = (u, v, w)$  is the liquid velocity vector;  $P$  the hydrodynamic pressure;  $\mu = \mu_L \rho / \rho_L$  the viscosity;  $h$ ,  $k$  and  $\rho$  are the medium enthalpy, heat conduction and density, respectively;  $Q_{las} = \beta_{las} J$  is the intensity of the volume heat source associated with laser action at different depths of the powder layer ( $J$  and  $\beta_{las}$  are the density of laser radiation energy flow and the radiation absorption coefficient in the local volume of the powder layer, respectively).  $f_s(\mathbf{r}, t) = 1 - f_L(\mathbf{r}, t)$ ,  $f_L(\mathbf{r}, t) = M_L / M_0$  is the mass fraction of the liquid phase formed at the point  $\mathbf{r} = (x, y, z)$  by the moment of time  $t$  ( $M_0$  and  $M_L$  are the total masses of metal and liquid phase, respectively).  $f_L = 0$  for the solid phase,  $f_L = 1$  for the totally transformed phase and for the two-phase region  $0 < f_L < 1$ . In the two-phase region (mushy zone) under study, the mass fraction of the liquid phase is defined by the formula  $f_L = (1 + \rho_s g_s / \rho_L g_L)^{-1}$ , here  $g_L, \rho_L$  and  $g_s, \rho_s$  are the volume fractions and densities of the liquid and solid phases.

The second term in the right side of (29) allows for the variation of angular momentums at the cost of liquid filtration through the porous medium (Darcy law), where  $K$  is the medium permeability,  $K = K_0 (f_L^3 + 10^{-10}) (1 - f_L)^{-2}$ , ( $K_0$  is the empiric constant defined by the interface morphology).  $K \rightarrow 0$  corresponds to the purely solid phase ( $f_L = 0$ ). In this case the

Darcy term becomes large, and the velocity of liquid decays to zero.  $K \rightarrow \infty$  corresponds to the purely liquid phase  $f_L = 1$ , when the Darcy term disappears.

The second term in the left side of (30) deals with convective heat transfer. The first term in the right side of (30) characterizes the transfer of heat at the cost of heat conduction; the second term describes the energy flow connected to relative movement of the  $L$  and  $S$  phases. Taking account of the heat flow released by laser radiation is introduced in the model as a volume source  $Q_{las}$  proportional to the volume absorption coefficient. The absorption accompanying laser radiation penetration into the powder layer is described by the law similar to Bouguer law for the optically uniform media:  $J = J_0 \exp(-\beta_{Las}z)$ , where  $J_0$  is the density of the energy flow on the layer surface.

Considering that the enthalpies of the solid and liquid phases are  $h_s = c_s T$  and  $h_L = c_L T + (c_s - c_L)T_s + L_m$ , respectively, and the mass fraction of the liquid phase is linearly dependent on the temperature  $f_L = (T - T_s)(T_L - T_s)^{-1}$ ,  $T_s \leq T \leq T_L$  ( $T_s$  and  $T_L$  are the solidus and liquidus temperatures, respectively), the energy equation is written as

$$\frac{\partial(\rho c_L T)}{\partial t} + \mathbf{v} \nabla \cdot (\rho c_L T) - \nabla \cdot (k \nabla T) = -\frac{\partial(\rho f_L L)}{\partial t} + \frac{\partial(\rho f_s \Delta c_p T)}{\partial t}, \quad (30a)$$

where the source term in the right side represents a variation in the enthalpy related to PT.

In the two-phase zone the density, velocity vector, enthalpy and heat conduction are found by the values of the volume and mass fractions [15]

$$\rho_m = \rho_s g_s + \rho_L g_L, \quad \mathbf{v} = \mathbf{v}_s f_s + \mathbf{v}_L f_L, \quad k_m = g_s k_s + g_L k_L, \quad h_m = h_s f_s + h_L f_L.$$

The boundary conditions on the free surface (liquid-gas interface) allow for convective ( $q_H$ ) and radiation loss ( $q_T$ ):

$$-k \nabla T = q_H + q_T, \quad q_H = -h_c (T - T_0), \quad q_T = -\sigma_E \sigma_{SB} (T^4 - T_0^4),$$

where  $h_c$  is the coefficient of convective heat transfer;  $T_0$  is the air temperature;  $\sigma_E$  is the surface emission coefficient;  $\sigma_{SB}$  is Stefan-Boltzmann constant.

At the interface of the liquid and solid phases  $\mathbf{v} = 0$ , which agrees with the conditions of non-leakage (the velocity component, normal to the surface, is zero) and adherence (the tangential component of velocity is zero). On the free surface of the liquid the capillary ( $\mathbf{F}_c$ ) and Marangoni ( $\mathbf{F}_M$ ) forces act, which are due to the surface tension gradient associated with the temperature field inhomogeneity along the interface (thermo-capillary forces):

$$\mathbf{F}_{S/L} = \mathbf{F}_c + \mathbf{F}_M = \gamma \kappa \mathbf{n} + \nabla_\tau \gamma,$$

where  $\kappa = -(\nabla \cdot \mathbf{n})$  is the free surface curvature;  $\mathbf{n}$  is the vector of free surface normal,  $\nabla_\tau$  is the surface gradient operator.

For most condensed media:  $\gamma(T) = \gamma_0(T_m) - \gamma_T(T - T_m)$ ,  $\gamma_T = |\partial \gamma_0 / \partial T_m|$ . Accordingly, for the total surface force we have

$$\mathbf{F}_{S/L} = \gamma \kappa \mathbf{n} + \gamma_T (\nabla T - \mathbf{n}(\mathbf{n} \cdot \nabla T)).$$

In the course of SLM the shape of the free surface (the interphase boundary gas-liquid/solid) is changed because of convective flows of the liquid on the surface. To define the evolution of the free surface, make use of the transport equation for the liquid volume fraction ( $\alpha$ ) in the micro-region (VOF model) that has the form [16]

$$\frac{\partial \alpha}{\partial t} + \mathbf{v} \cdot \nabla \alpha = 0. \quad (31)$$

If  $\alpha = 1$ , the microregion is completely filled with metal, if  $\alpha = 0$ , it is filled with gas. If  $0 < \alpha < 1$ , the micro-region contains a free surface. It is evident that  $\alpha + \beta = 1$ , where  $\beta$  is the volume fraction of the gas phase.

The physical properties of the medium in the transition zone (gas-liquid/solid) is found by the VOF function

$$\rho = \rho_g + \alpha(\rho_m - \rho_g), \quad c_p = c_{pg} + \alpha(c_{pm} - c_{pg}), \quad \lambda = \lambda_g + \alpha(\lambda_m - \lambda_g),$$

$$h = h_g + \alpha(h_m - h_g), \quad \mu = \mu_g + \alpha(\mu_m - \mu_g),$$

the values with  $m$  index refer to the metal and the values with  $g$  index belong to the gas.

The surface forces ( $\mathbf{F}_{S/L}$ ), acting per unit of free surface area can be transformed to the volume forces by Dirac delta function  $\delta(\alpha)$ , i.e.

$$\mathbf{F}_{S/L}^{vol} = \mathbf{F}_{S/L} \delta(\alpha) = \gamma \kappa \nabla \alpha + \gamma_T (\nabla T - \mathbf{n}(\mathbf{n} \cdot \nabla T)) |\nabla \alpha|.$$

Accordingly, for the volume sources in the energy equation we have:

$$q^{vol} = (q_H + q_L) \delta(\alpha) + Q_{las}.$$

The normal to the free surface is found by the gradient of VOF function:  $\mathbf{n} = \nabla \alpha / |\nabla \alpha|$ , and its location is defined with the help of VOF function itself.

The set of equations (28)-(31) represents a macroscopic thermodynamic model of SLM. In combination with the appropriate boundary conditions it permits defining the temperature distribution, the velocity fields of thermo-capillary flows and the profile of the melt free surface depending on the regimes (beam power, scanning velocity) of the SLM process.

For the purpose of numerical calculation using the outlined model, the program software has been developed. Its realization involved the C++ class library of numerical modeling OpenFOAM2.4. The finite volume method was applied on the unstructured hexahedral mesh [17].

The initial powder layer was applied by specifying the nonuniform original structure of the field of metal phase volume fraction ( $\alpha$ ). The powder particles were given as the distributed spherical regions. The PISO algorithm was used to solve the continuity and Navier-Stokes equations for the liquid dynamics. The transport equation for the volume fraction is calculated by the MULES method [18]. The energy equation is solved by the “enthalpy – porosity” method using the implicit scheme [19].

The 3D distribution of the volume fraction of the metal phase in the operating region has been obtained. Fig. 2 presents the examples of calculating the evolution of the metal phase field structure under the scanning action of laser radiation of 200 W power on the pre-poured layer of the particles of Inconel 718 powder (the particle size  $40\mu\text{m}$ , the scanning velocity  $1.7\text{ m/s}$ ).

Fig. 2 displays the specific effects accompanying the SLM process – melting of the metal particles, wetting of the solid metal with the melt, coalescence of the liquid droplets. The calculated distributions also demonstrate the development of widespread defects in SLM technology, e.g. residual porosity inside the solidified metal in the form of gas bubbles, incomplete penetration and bonding of the substrate metal and the particles. The obtained results suggest that the capillary effects make a decisive contribution to the dynamics of the liquid phase and, correspondingly, to the final profile and structure of the deposited layer.

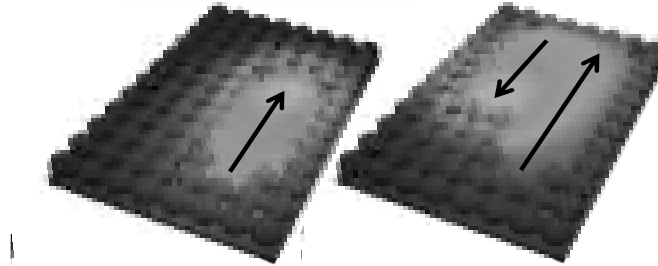


Fig 2. The distribution of the metal phase during scanning two consecutive tracks.

#### 4. Conclusion

The mathematical statement of the problem of crystallization and evolution of solid phase microstructure in the course of SLM of a powder compact has been formulated and substantiated by the use of two-scale approximation. The microstructure formation is described by the equation for the two-component ordering parameter (degree of order, orientation field), conjugate with the equations of energy transfer and impurity diffusion, as well as by the elastic stresses accompanying PT. In formulating the elasticity equation the constitutive relations were used which relate the elastic stresses to the fields of strains, temperature, concentration, as well as to the ordering parameter. The model under study was constructed on basis of the unified entropy functional and the law of its increment (entropy production positivity) that is also valid for non-isothermal conditions, which agrees with the principles of thermodynamics of irreversible processes. The particular cases of the derived evolution equations have been discussed.

The model of microstructure is adjoint with the macroscopic thermodynamic model of SLM. At the macrolevel, taking account of the heat transfer processes, thermocapillary convection and free surface evolution receives primary attention. The macromodel can find application in forecasting thermal flows and melt velocity fields depending on the technological parameters (beam power, scanning velocity) of the SLM process. Modeling of the free surface evolution involved the application of the VOF function. The data obtained from the macrolevel (e.g., heat removal rate) can be used as the input parameters in formulating the boundary conditions for solving the microproblem as well. The macroproblem has been numerically realized, and the test calculation of the metal phase distribution has been conducted.

The developed model can provide the basis for predictive investigation of the formation of microstructure and stress-strain states which is requisite for the control and optimization of the additive SLM technologies of synthesis of polycrystalline materials.

#### Acknowledgments

The work has been performed with the support of the Russian Federation for Basic Research (grant no: 16-29-11743 ofi-m).

#### References

- [1] Gladush GG, Smurov I. Physics of Laser Materials Processing: Theory and Experiment. Berlin: Springer-Verlag, 2011; 534 p.
- [2] Xiao B, Zhang Yu. Laser sintering of metal powders on top of sintered layers under multiple-line laser scanning. J Phys. D: Appl. Phys. 2007; 40: 6725–6732.
- [3] Gusarov AV, Yadroitsev I, Bertrand Ph, Smurov I. Heat transfer modelling and stability analysis of selective laser melting. Appl. Surf. Sci. 2007; 254: 975–983.
- [4] Modern laser and information technologies. Edited by acad V. Ya. Panchenko and prof F.V. Lebedev. M.: Intercontact Nauka, 2014; 959 p.
- [5] Markl M., Korner C. Multi-Scale modeling of powder-bed-based additive manufacturing. Annual Review of Materials Research 2016; 46: 1–34.
- [6] Mirzade FKh. Phase field approach to solidification including stress effects at laser sintering of metal powders. J. Applied Spectroscopy 2017; 84(8) (accepted).

- [7] Dubrov AV, Dubrov VD, Mirzade FK, Panchenko VYa. Heat transfer and thermocapillary convection in laser additive manufacturing process by injection of metal powders. *Poverchnost* 2017 (accepted).
- [8] Wang J, Yang M, Zhang Yu. A multiscale nonequilibrium model for melting of metal powder bed subjected to constant heat flux. *Int. J. Heat and Mass Transfer* 2015; 80: 309–318.
- [9] Wang S-L, Sekerka RF, Wheeler AA, et al. Thermodynamically-consistent phase-field models for solidification. *Physica D* 1993; 69: 189–200.
- [10] Warren JA, Boettinger WJ. Prediction of dendritic growth and microsegregation patterns in a binary alloy using the phase-field method. *Acta Metall. Mater.* 1995; 43: 689–703.
- [11] Bi Z, Sekerka RF. Phase-field model of solidification of a binary alloy. *Physica A* 1998; 261: 95–106.
- [12] Penrose O, Fife PC. Thermodynamically consistent models of phase-field type for the kinetics of phase transitions. *Physica D* 1990; 43: 44–62.
- [13] Karma A, Rappel W-J. Quantitative phase-field modeling of dendritic growth in two and three dimensions. *Phys. Rev. E* 1998; 57: 4323–49.
- [14] Kobayashi R, Warren JA, Carter WC. Vector-valued phase field for crystallization and grain boundary formation. *Phys. D* 1998; 119: 415–423.
- [15] Bennon WD, Incropera FP. A continuum model for momentum, heat and species transport in binary solid-liquid phase change systems. I. Model formulation. *Int. J. Heat Mass Transfer* 1987; 30: 2161–2169.
- [16] Osher S, Sethian JA. Fronts propagation with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.* 1988; 79: 12–49.
- [17] Weller HG, Tabor G, Jasak H, Fureby C. A tensorial approach to computational continuum mechanics using object orientated techniques. *Comput. Phys.* 1998; 12: 620–631.
- [18] Marquez Damian S. An Extended Mixture Model for the Simultaneous Treatment of Short and Long Scale Interfaces: Doctor Thesis. Santa Fe: Universidad Nacional Del Litoral, 2013; 231 p.
- [19] Voller VR, Prakash C. A fixed grid numerical modelling methodology for convection-diffusion mushy region phase-change problems. *Int. J. of Heat and Mass Transfer* 1987; 30(8): 1709–1719.

# Research of the effect of aberrations on image quality in optical systems

A.V. Kozhevnikov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

This work presents a study of the influence and compensation of aberrations in optical imaging systems by superimposing surfaces described by Zernike polynomials. The aim is to investigate how to compensate for aberrations and their effectiveness when dealing with aberrations of different types. The study is done in the Zemax packet by modeling the optical imaging system and passing test images through it.

*Keywords:* wave aberrations; Zernike polynomials; imaging optical system

## 1. Introduction

The aberration of the optical system is the deformation of images that occur at the output of the optical system. The name comes from the lat. Aberratio - evasion, removal. Deformation consist in the fact that the optical images do not completely correspond to the object. This is manifested in the blurriness of the image and is called monochromatic geometric aberration or color image - chromatic aberration of the optical system. Most often, both types of aberration appear together.

In the paraxial region, the optical system works almost perfectly, the dot is represented by a point, and the straight line is a straight line, etc. However, as the point moves away from the optical axis, the rays from it intersect in the image plane not at one point. Thus, a circle of dispersion arises, i.e. there are aberrations. The dimension of the aberration can be determined by calculating the geometric and optical formulas through a comparison of the coordinates of the rays, and also approximately using the formulas of the theory of aberrations [1].

There is a description of the phenomenon of aberration in both the ray theory (deviation from identity is described through geometric aberrations and ray scattering patterns) and in the representations of wave optics (the deformation of a spherical light wave along the path through an optical system is estimated). Usually, ray theory is used to characterize optical systems with large aberrations, otherwise the principles of wave optics are applied.

As a rule, analysis of wave aberrations is performed on the basis of Zernike polynomials [1-4]. In [5-8], it was proposed to use a multi diffractive optical elements for an optical wavefront decomposition on the basis of Zernike polynomials. Moreover, the use of optical elements that are consistent not only with Zernike functions, but also their superpositions, makes it possible to perform optical measurements ensuring the restoration of the shape of the wave front [9-11].

In this work, we investigate the effect of aberrations described by Zernike polynomials on the imaging properties of an optical system using the Zemax packet [12]. The aim of the work is to investigate methods of aberration compensation [13-15] and their effectiveness in dealing with aberrations of different types.

## 2. Modeling of wave aberrations by varying the coefficients of a polynomial surface

The first stage will be modeled several surfaces described by Zernike polynomials. For this part, an optical system consisting of two refractive lenses is used, one of which is the surface under investigation. Light with wavelengths of 450, 550 and 650 nm is transmitted through such a system. As an estimation of aberration distortions, an image of the Latin letter "F", passed through the described optical system, is considered. The most interesting are the coefficients - "Zernike standard coefficients".

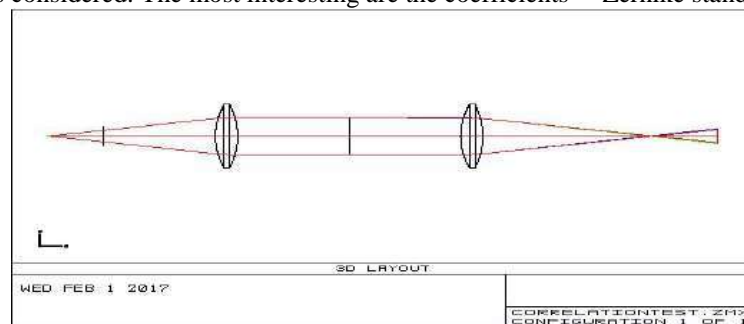


Fig.1. Two-dimensional scheme of the optical system

There are even and odd Zernike polynomials. Even polynomials are defined as:  $Z(\rho, \varphi) = R_n^m(\rho) \cos(m\varphi)$ , and odd as:  $Z_n^{-m}(\rho, \varphi) = R_n^m(\rho) \sin(m\varphi)$ . Where  $m$  and  $n$  are nonnegative integers, such that  $n > m$ ,  $\varphi$  - azimuth angle, and  $\rho$  - radial distance  $0 \leq \rho \leq 1$ . Zernike polynomials are limited in the range from -1 to +1. Radial polynomials are defined as:  $R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k (n-k)!}{k! \left(\frac{n+m-k}{2}\right)! \left(\frac{n-m-k}{2}\right)!} \rho^{n-2k}$

Table.1. The result of simulation of wave aberrations by the Zernike row decomposition.

The value and order of the coefficient to be set	«Zernike standard coefficients»	Diagram of point dispersion «Spot diagram»	Test image after passing through the simulated system.
Zernike 2 = 1.3	<pre> 1 86.94863798 2 8.92983652 3 0.00000000 4 52.88124147 5 0.00000000 6 0.18081260 7 0.00000000 8 3.18866717 9 0.00000000 10 0.22819498 11 1.99700646 12 0.18851551 13 0.00000000 14 0.00994874 15 0.00000000 16 0.00884830 17 0.00000000 18 0.00672670 19 0.00000000 20 0.00088664                     </pre>		
Zernike 3 = 1.0	<pre> 1 86.04342006 2 0.00000000 3 5.59445929 4 52.23096902 5 0.00000000 6 0.01166767 7 1.99456861 8 0.00000000 9 -0.06688107 10 0.00000000 11 1.92430631 12 -0.08080670 13 0.00000000 14 0.00188636 15 0.00000000 16 0.00000000 17 0.00372627 18 0.00000000 19 -0.00191572 20 0.00000000                     </pre>		
Zernike 4 = 0.008	<pre> 1 87.70230255 2 0.00000000 3 0.00000000 4 53.03219018 5 0.00000000 6 0.00000000 7 0.00000000 8 0.00000000 9 0.00000000 10 0.00000000 11 1.80228316 12 0.00000000 13 0.00000000 14 0.00000021 15 0.00000000 16 0.00000000 17 0.00000000 18 0.00000000 19 0.00000000 20 0.00000000                     </pre>		
Zernike 5 = 0.04	<pre> 1 84.91032428 2 0.00000000 3 0.00000000 4 51.42100147 5 8.15314918 6 0.00000000 7 0.00000000 8 0.00000000 9 0.00000000 10 0.00000000 11 1.80368721 12 0.00000000 13 -0.03312496 14 -0.00171468 15 0.00000000 16 0.00000000 17 0.00000000 18 0.00000000 19 0.00000000 20 0.00000000                     </pre>		
Zernike 6 = 0.05	<pre> 1 84.65788774 2 0.00000000 3 0.00000000 4 51.23798754 5 0.00000000 6 10.13078588 7 0.00000000 8 0.00000000 9 0.00000000 10 0.00000000 11 1.77378504 12 -0.04435208 13 0.00000000 14 0.00265525 15 0.00000000 16 0.00000000 17 0.00000000 18 0.00000000 19 0.00000000 20 0.00000000                     </pre>		
Zernike 10 = 0.5	<pre> 1 89.67817589 2 -0.00000803 3 0.00000000 4 55.43783518 5 0.00000000 6 0.00006358 7 0.00000000 8 -0.00001118 9 0.00000000 10 40.86128769 11 2.70599798 12 0.00007826 13 0.00000000 14 -0.00000030 15 0.00000000 16 -0.00001331 17 0.00000000 18 -0.35953301 19 0.00000000 20 -0.00000589                     </pre>		



The orthogonality of the Zernike polynomials gives them great advantages when analyzing aberrations in comparison with the exponentiation basis. The main advantages are:

- 1) The absolute values of the coefficients of the decomposition in the Zernike polynomials decrease with increasing degree of polynomials, that is, the Zernike row, as a rule, always converges, which cannot be said about the exponentiation rows;
- 2) Each coefficient of the row gives the contribution of aberrations of a given type and order to the total wave aberration from the position of mutual balance of all types of aberrations. This means that the individual types of aberrations represented by the Zernike polynomial decomposition affect image quality quite independently of each other.

Table 2. The result of compensating wave aberrations.

Values of the coefficients of the basic and compensating surfaces.	«Zernike standard coefficients»	Wavefront map	Diagram of point dispersion «Spot diagram»
2nd order Zernike 1 = 22 Zernike 2 = 22	Z 1 -635.73327097 Z 2 0.00000010 Z 3 0.00000000 Z 4 -367.05864763 Z 5 0.00000000 Z 6 0.00000000 Z 7 0.00000000 Z 8 -0.00000003 Z 9 0.00000000 Z 10 -0.00000002 Z 11 -0.03709060 Z 12 0.00000000 Z 13 0.00000000 Z 14 0.00000000 Z 15 0.00000000 Z 16 0.00000002 Z 17 0.00000000 Z 18 0.00000001 Z 19 0.00000000 Z 20 0.00000002		
3rd order Zernike 1 = 10 Zernike 2 = 10	Z 1 -643.87679565 Z 2 0.00000000 Z 3 0.00000009 Z 4 -371.79217736 Z 5 0.00000000 Z 6 0.00000000 Z 7 -0.00000003 Z 8 0.00000000 Z 9 0.00000003 Z 10 0.00000000 Z 11 -0.03853468 Z 12 0.00000000 Z 13 0.00000000 Z 14 0.00000000 Z 15 0.00000000 Z 16 0.00000000 Z 17 0.00000000 Z 18 0.00000000 Z 19 0.00000000 Z 20 0.00000000		
4nd order Zernike 1 = 1.3 Zernike 2 = 1.3	Z 1 -537.97787897 Z 2 0.00000000 Z 3 -154.41057438 Z 4 -310.35390764 Z 5 0.00000000 Z 6 177.11633271 Z 7 -54.57324235 Z 8 0.00000000 Z 9 49.67569126 Z 10 0.00000000 Z 11 0.19062547 Z 12 0.23027790 Z 13 0.00000000 Z 14 -0.83171289 Z 15 0.00000000 Z 16 0.00000000 Z 17 0.01070882 Z 18 0.00000000 Z 19 0.02084444 Z 20 0.00000000		
5nd order Zernike 1 = 1.0 Zernike 2 = 1.0	Z 1 -661.32641050 Z 2 -16.08281281 Z 3 -24.1827705 Z 4 -381.96874935 Z 5 -144.52988974 Z 6 0.47724697 Z 7 -8.53856129 Z 8 -5.68637885 Z 9 6.64988930 Z 10 16.47768968 Z 11 -0.11745892 Z 12 0.12289812 Z 13 -0.04481014 Z 14 -0.10619428 Z 15 0.02143110 Z 16 -0.00013610 Z 17 0.00613696 Z 18 -0.00074085 Z 19 -0.00635697 Z 20 0.00488021		
6nd order Zernike 1 = 1.0 Zernike 2 = 1.0	Z 1 -696.95149995 Z 2 0.00000000 Z 3 4.17438084 Z 4 -402.51959586 Z 5 0.00000000 Z 6 -169.39979034 Z 7 1.50470917 Z 8 0.00000000 Z 9 -17.57482110 Z 10 0.00000000 Z 11 -0.10488243 Z 12 0.15135028 Z 13 0.00000000 Z 14 0.00750577 Z 15 0.00000000 Z 16 0.00000000 Z 17 0.01570930 Z 18 0.00000000 Z 19 -0.00688875 Z 20 0.00000000		
10nd order Zernike 1 = 1.0 Zernike 2 = 1.0	Z 1 -693.59874312 Z 2 43.14131352 Z 3 -32.07688033 Z 4 -416.87052236 Z 5 25.29308230 Z 6 23.61726878 Z 7 -17.86472572 Z 8 16.35792099 Z 9 4.94299591 Z 10 -101.59634855 Z 11 -12.58088299 Z 12 8.97828396 Z 13 7.07568058 Z 14 14.01357091 Z 15 -16.84582426 Z 16 0.61519900 Z 17 -3.43159284 Z 18 -1.52872307 Z 19 0.91258667 Z 20 0.62860773		

In each case, the coefficients of different orders are taken, from 1st to 10th, they are assigned the maximum values at which the test image preserves the original contour and sharpness. The parameter "Zernike max term" is set to 16, it is responsible for the number of polynomial surface parameters.

As can be seen from Table 1, the variation of the coefficients of the polynomial surface leads to deformations in the point dispersion diagram of and on the test image passed through the test system.

### 3. Investigation of compensation of aberrations in the optical system

In this section, we study the compensation of aberrations in a test optical imaging system by superimposing surfaces described by Zernike polynomials. The aim is to search for the maximum efficiency of compensating wave aberrations by adding a second polynomial surface. The effect is achieved due to the complex conjugacy of polynomials.

The optical system used consists of two mirrors arranged at a certain angle to each other and an auxiliary paraxial surface used to parallelize the incoming light beam. In both mirrors there are polynomially described surfaces.

As can be seen from Table 2, the addition of the second Zernike surface allows us to compensate for the simulated aberrations in the optical system.

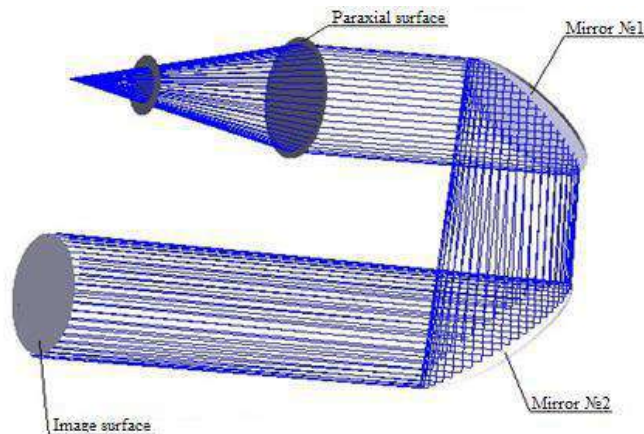


Fig.2. Optical system of two mirrors.

### 4. Conclusion

The decomposition of the wave aberrations in the Zernike row was conducted in this work. Wave aberrations of various types were modeled and their effect on image quality in optical systems was examined. An optical system of two mirrors with two polynomial surfaces is simulated. The first surface was used to model the aberrations themselves, the second was used to compensate them. The principle used is based on the complex conjugacy of polynomials.

The possibility of compensation of wave aberrations by superposition of surfaces described by Zernike polynomials is investigated. The maximum values of aberrations that can be compensated for by relatively effective compensation due to the use of polynomial surfaces were obtained.

The study was carried out in a Zemax packet by modeling test optical imaging systems, and passing test images through them.

### Acknowledgments

The work was supported by the Ministry of Education and Science of the Russian Federation.

### References

- [1] Volf E, Born M. Fundamentals of Optics. Moscow: "Nauka" Publisher, 1973. (in Russian)
- [2] American National Standards Institute, Inc. American National Standards for Ophthalmics – Methods for Reporting Optical Aberrations of Eyes. ANSI Z80.28, 2004.
- [3] Bezdydyko SN. Optimization of optical systems using orthogonal polynomials. Optics and spectroscopy 1980; 48: 222–224. (in Russian)
- [4] Bezdydyko SN. Methodological aspects of application of Zernike polynomials in computational optics. Materials of the International Conference dedicated to the 90th anniversary of the birth of the Nobel Prize winner Academician A.M. Prokhorov. The fundamental foundations of engineering 2006; 4. (in Russian)
- [5] Khonina SN, Kotlyar VV, Soifer VA, Wang Y, Zhao D. Decomposition of a coherent light field using a phase Zernike filter. Proceedings of SPIE 1998; 3573: 550–553.
- [6] Ha Y, Zhao D, Wang Y, Kotlyar VV, Khonina SN, Soifer VA. Diffractive optical element for Zernike decomposition. Proceedings of SPIE 1998; 3557: 191–197.
- [7] Khonina SN, Kotlyar VV, Wang Ya. Diffractive optical element matched with Zernike basis. Pattern Recognition and Image Analysis 2001; 11(2): 442–445.
- [8] Porfirev AP, Khonina SN. Experimental investigation of multi-order diffractive optical elements matched with two types of Zernike functions. Proceedings of SPIE 2016; 9807: 9 p.
- [9] Kotlyar VV, Khonina SN, Soifer VA, Wang Y, Zhao D. Coherent field phase retrieval using a phase Zernike filter. Computer Optics, 1997; 17: 43–48.



- [10] Khonina SN, Kotlyar VV, Kirsh DV. Zernike phase spatial filter for measuring the aberrations of the optical structures of the eye. *Journal of Biomedical Photonics & Engineering* 2015; 1(2): 146–153.
- [11] Khorin PA, Khorina SN, Karsakov AV, Branchevsky SL. Analysis of human eye cornea aberrations. *Computer Optics* 2016; 40(6): 810–817. (in Russian) DOI: 10.18287/0134-2452-2016-40-6-810-817.
- [12] Zemax® User's Guide. Zemax Development Corporation 2005.
- [13] Tokovinin A, Heathcote S. DONUT: measuring optical aberrations from a single extrafocal image. *Publications of the Astronomical Society of the Pacific* 2006; 118(846): 1165–1175.
- [14] Booth MJ. Wavefront sensorless adaptive optics for large aberrations. *Optics Letters* 2007; 32 (1): 5–7.
- [15] Klebanov IM, Karsakov AV, Khonina SN, Davydov AN, Polyakov KA. Wave front aberration compensation of space telescopes with telescope temperature field adjustment. *Computer Optics* 2017; 41(1): 30–36. (in Russian) DOI: 10.18287/0134-2452-2017-41-1-30-36.

# Current problems of development of the journal of Computer Optics

D.V. Kudryashov<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

This paper presents overall results of two years experience in executing tasks defined in article “Quo Vadis” written by the Editor-in-Chief of the journal of Computer Optics, Corresponding Member of RAS V.A. Soifer (Computer Optics 2014; Volume 38, Issue 4). The main bibliometric indicators of the journal in SCOPUS citation database are compared with similar metrics of some most relevant periodicals. The paper declares some important events to be held in 2017. Based on the current progress of the journal, new objectives of its further development are defined, and major events targeted for 2017 are discussed.

*Keywords:* scientific journal; journal promotion; bibliometric indicators; SCOPUS; comparison of metrics; quartile; development plan

---

## 1. Introduction

The scientific journal of Computer Optics has been issued since 1987 in Russian and English languages. During this period totally 40 volumes were published, including over 1.500 scientific articles. From 2007, the journal was issued 4 times a year.

In the mid 2014s, the Editor-in-Chief of the journal, Corresponding Member of RAS V.A. Soifer in his article “Quo Vadis” [1] set new tasks for the authors and the editorial team, the majority of which required at least two years to be implemented. Execution of tasks defined in [1] allowed the journal to significantly improve its bibliometric indicators [2-3].

## 2. Results of two years development

In 2015, we published five issues of the journal in Russian and recommenced publishing its English version entitled Computer Optics Selected Papers (it included articles translated into English and published in previous issues of the journal of Computer Optics). From 2016, the journal has been published 6 times a year, plus one more issue of Computer Optics Selected Papers published additionally. This allows us to reduce deadlines for publication of peer-reviewed papers to two or three months.

The deadline for papers peer-reviewing was significantly reduced in the last two years. The number of reviewers was greatly expanded. Each paper is now evaluated at least by two experts [2], usually being Doctors of Sciences and working in different institutions of RAS and some leading Russian universities.

The Editorial Board of the journal of Computer Optics was also expanded and it now includes, along with leading scientists from Germany, India, China and Finland [4], also some famous scientists from the USA, Great Britain and Ireland.

According to international standards, the journal has been added with information on article citations. Each article has been supplied with its unique Digital Object Identifier (DOI). Requirements for submitting the References in English have been changed, and they are now drew up not according to GOST (Russian State Standard) requirements, but in accordance with requirements of the largest international citation database SCOPUS. In Computer Optics, the number of foreign sources in the References has been significantly increased too. This drastically raises the prospect of an article to be widely cited in other periodicals and thus a citation rate of Computer Optics to be improved [4]. In subsequent issues of the journal, expatriate members of the Editorial Board will focus on the number of citations in foreign sources given in the article. Beginning from 2009, archive materials of Computer Optics are added to SCOPUS database that has an impact on improving the citation rate of the journal [5]. We are currently working on adding its issues dated from 2005 to this database.

Changes have also taken place in the journal content – new publication categories have appeared: Earth’s Remote Sensing Technology, including development of hyperspectral equipment [6-9]; photonic-crystal sensors [10-11]; LED technology [12-13]; video signal stream processing and some other new methods of image subject processing [14-18]; new types of laser beams [19-20]; and academic reviews of up-to-date trends have been published [21-22]. In 2016, for the first time in a long period, a full-text English version of the journal of Computer Optics was issued with the articles originally written in English (Volume 40, Issue 5), which had not previously been published elsewhere.

The journal of Computer Optics is an Open Access periodical: coordinated pdf-versions of its articles are available in open access on the journal website: [www.computeroptics.smr.ru](http://www.computeroptics.smr.ru); moreover, their publication is free for the authors. We may also review or download the articles in Russian and foreign databases, repositories and e-libraries.

Since 2012, the journal already represented in the Russian Science Citation Index (RSCI) has been peer-reviewed and indexed in international scientific citation databases SCOPUS and Compendex that became impactful advances for the regional journal with no full-text version in English [5]. Within 2015, the journal began to be represented in CyberLeninka e-library, in MathNet, Applied Science & Technology Source Ultimate (EBSCO Publishing), Inspec databases. At the end of 2015, the journal of Computer Optics was included into the Russian periodical scientometric database – Russian Science Citation Index (RSCI) – within the Web of Science network from Thomson Reuters. The journal RSCI records will allow us to improve its quality by conforming to international standard requirements and to increase its bibliometric indicators in Web of Science and

total integral indicators of Russia due to increasing its accessibility and citation level worldwide [4]. The next step for the journal should be its entry to Web of Science Core Collection database. For this purpose, we will proceed to attract new high-level scientific papers and to expand the authors range, thus providing actual opportunities for fast and open access publication.

### 3. Facts and figures

Measures taken by the Editor-in-Chief of the journal to fulfill its major objectives have resulted in significant improvement of key scientometric indicators of Computer Optics in the most influential Russian and foreign databases.

Upon its current indicators, the journal of Computer Optics has been almost equated with the Journal of Modern Optics and has exceeded some impactful and influential journals such as Optik (Jena), Optical Engineering and Technical Physics Journal. In accordance with SCImago Journal & Country Rank indicators, the journal of Computer Optics has entered into the second quartile in all relevant subject domains. Its Hirsch index comes up to 10.

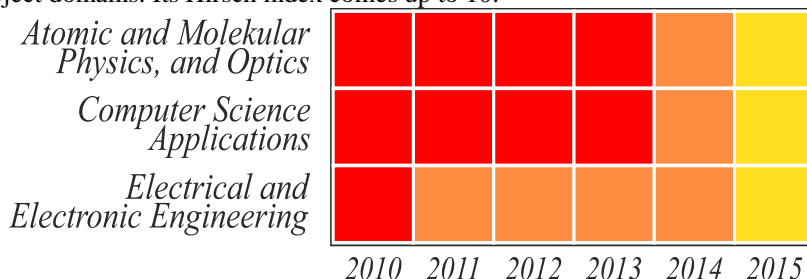


Fig. 1. The journal of Computer Optics entered into quartiles in three main areas: Physics and Optics; IT; Electronics (red - the 4<sup>th</sup> quartile, fawn-colored - the 3<sup>rd</sup> quartile, yellow - the 2<sup>nd</sup> quartile).

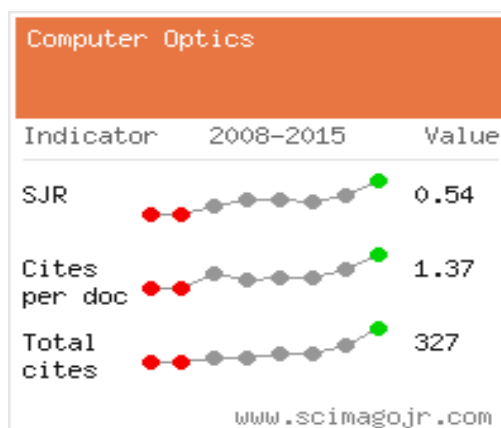


Fig. 2. Key scientometric indicators of the journal of Computer Optics according to SCImago Journal & Country Rank.

Key scientometric indicators of the journal of Computer Optics in RSCI (data valid as of 2015):

- 2-Years Impact Factor – 1.182
- 5-Years Impact Factor– 0.902
- article citations in previous 2 years – 506
- article citations in previous 5 tears – 368
- the Herfindahl Index (last 5 years) according to citing journals – 1390
- the Herfindahl Index according to authors’ institutions – 2470
- the H-Index (last 10 years) – 13
- total journal citations in RSCI – about 6000.

Key scientometric indicators of the journal of Computer Optics in SCOPUS database (data valid as of 2015):

- SJR indicator (SCImago Journal Rank) – 0.535
- IPP (Impact per Publication) – 1.185
- SNIP (Source Normalized Impact per Paper) – 1.284

CiteScore metrics (citations in 2015 Scopus database published in 2012 through 2014 divided by the number of the articles): 1.22 (in 2016 – 1.59).

Table 1. Change in the article citation level in Computer Optics (according to SCOPUS database)

Year	Number of articles	Number of citations
2009	41	11
2010	64	40
2011	70	48
2012	80	84
2013	68	104
2014	124	260
2015	106	452

Key indicators of Computer Optics according to SCOPUS database compared with the most relevant publications are given in details in Figures 3-5.

Table 2. Key scientometric indicators of the journal of Computer Optics as compared to the most relevant journals (according to SCOPUS as of 2015)

No.	Title	CiteScore	SNIP	SJR
1	Optics Express	3.78	1.664	2.186
2	Applied Optics	1.66	1.147	0.898
3	Journal of Optics	1.44	0.631	0.765
4	Computer Optics	1.22	1.284	0.535
5	Quantum Electronics	1.07	1.124	0.631
6	Optical Memory and Neural Networks (Information Optics)	0.76	1.372	0.344

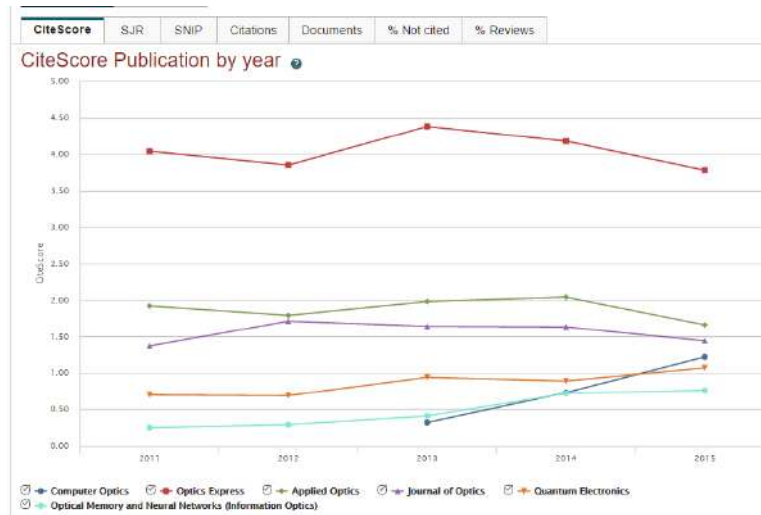


Fig. 3. CiteScore ranking of the journal of Computer Optics as compared to the most relevant journals.

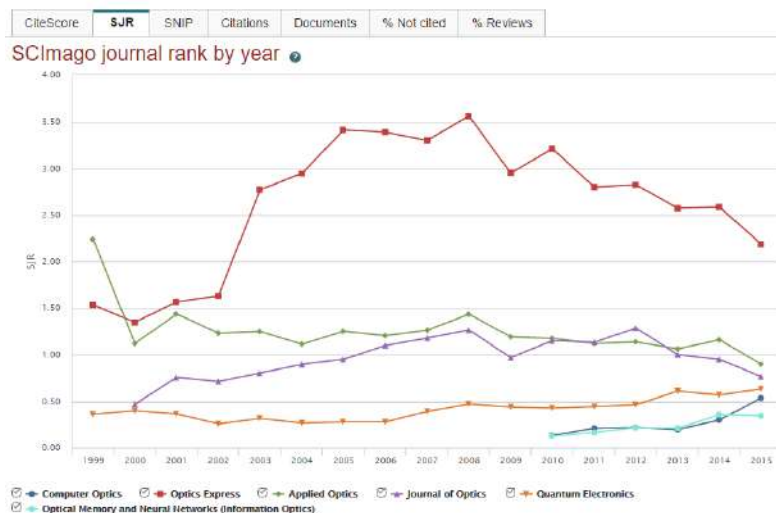


Fig. 4. SJR ranking of the journal of Computer Optics as compared to the most relevant journals.

#### 4. Conclusion

The goal of the present stage of development of the journal is its entry into the Web of Science Core Collection. For this purpose, the Editorial Board intends to continue its work in several ways:

- improvement of publications quality,
- preparation of various reviews on relevant topics,
- strict conformance to peer-reviewing and issuing deadlines.

In 2017, we plan to prepare the third issue of Selected Papers (English versions of the articles published in the journal in 2015-2016) and a fully English-translated issue (Issue 4, 2017), and to expand a list of reference databases wherein the journal is represented.

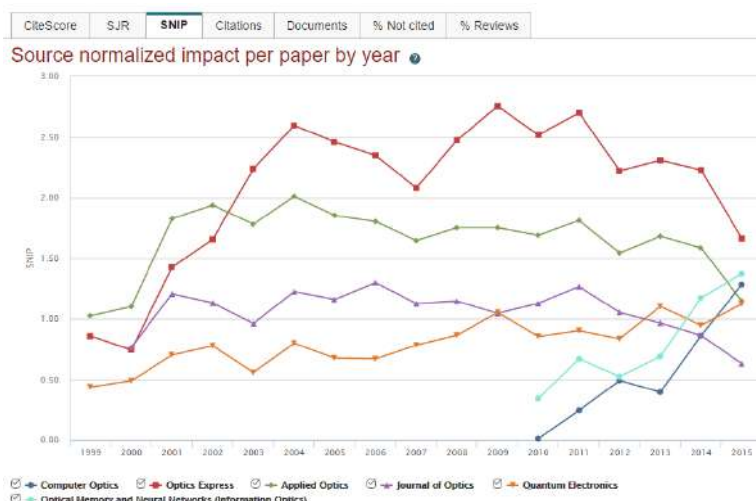


Fig. 5. SNIP ranking of the journal of Computer Optics as compared to the most relevant journals.

## Acknowledgements

The author is deeply indebted to A.A. Bukhanko, Dr. Sci. in Physics and Mathematics, N.L. Kazanskiy, Dr. Sci. in Physics and Mathematics, A.V. Kupriyanov, Dr. Sci. in Physics and Mathematics, and S.S. Stafeev, Cand. Sci. in Physics and Mathematics, for their help and useful discussions.

## References

- [1] Soifer VA. Quo vadis. *Computer Optics* 2014; 38(4): 589.
- [2] Kazanskiy NL. Advances of the journal of *Computer Optics*. *Computer Optics* 2017; 41(1): 139–141. (in Russian) DOI: 10.18287/2412-6179-2017-41-1-139-141.
- [3] Sokolov VO. Contribution of Samara scientists into *Computer Optics* journal development. *CEUR Workshop Proceedings* 2016; 1638: 194–206. DOI: 10.18287/1613-0073-2016-1638-194-206.
- [4] Kolomiets EI. Analysis of activity of the scientific journal *Computer Optics*. *Proceedings of Information Technology and Nanotechnology (ITNT-2015)*. *CEUR Workshop Proceedings* 2015; 1490: 138–150. DOI: 10.18287/1613-0073-2015-1490-138-150.
- [5] Kudryashov DV. The scientific advancement and promotion of the journal "Computer Optics" in 2014–2015. *CEUR Workshop Proceedings*, 2016; 1638: 185–193. DOI: 10.18287/1613-0073-2016-1638-185-193.
- [6] Kazanskiy NL, Kharitonov SI, Khonina SN. Simulation of a hyperspectrometer based on linear spectral filters using vector Bessel beams. *Computer Optics* 2014; 38(4): 770–776. (in Russian)
- [7] Golovin AD, Demin AV. Simulation model of a multichannel Offner hyperspectrometer. *Computer Optics* 2015; 39(4): 521–528. DOI: 10.18287/0134-2452-2015-39-4-521-528.
- [8] Kazanskiy NL, Kharitonov SI, Doskolovich LL, Pavelev AV. Modeling the performance of a spaceborne hyperspectrometer based on the Offner scheme. *Computer Optics* 2015; 39(1): 70–76. DOI: 10.18287/0134-2452-2015-39-1-70-76.
- [9] Karpeev SV, Khonina SN, Kharitonov SI. Study of the Diffraction Grating on a Convex Surface as a Dispersive Element. *Computer Optics* 2015; 39(2): 211–217. DOI: 10.18287/0134-2452-2015-39-2-211-217.
- [10] Egorov AV, Kazanskiy NL, Serafimovich PG. Using Coupled Photonic Crystal Cavities for Increasing of Sensor Sensitivity. *Computer Optics* 2015; 39(2): 158–162. DOI: 10.18287/0134-2452-2015-39-2-158-162.
- [11] Kadomina EA, Bezus EA, Doskolovich LL. Resonant photonic-crystal structures with a diffraction grating for refractive index sensing. *Computer Optics* 2016; 40(2): 164–172. DOI: 10.18287/2412-6179-2016-40-2-164-172.
- [12] Kazanskiy NL, Stepanenko IS, Khaimovich AI, Kravchenko SV, Byzov EV, Moiseev MA. Injectional multilens molding parameters optimization. *Computer Optics* 2016; 40(2): 203–214. DOI: 10.18287/2412-6179-2016-40-2-203-214.
- [13] Doskolovich LL, Andreev ES, Byzov EV. Analytical design of mirrors generating prescribed two-dimensional intensity distributions. *Computer Optics* 2016; 40(3): 346–352. DOI: 10.18287/2412-617-2016-40-3-346-352.
- [14] Kazanskiy NL, Protsenko VI, Serafimovich PG. Comparison of system performance for streaming data analysis in image processing tasks by sliding window. *Computer Optics* 2014; 38(4): 804–810.
- [15] Ilyasova NY, Kupriyanov AV, Paringer RA. Formation of features for improving the quality of medical diagnosis based on discriminant analysis methods. *Computer Optics* 2014; 38(4): 851–855.
- [16] Kotov AP, Fursov VA, Goshin YV. Technology for fast 3D-scene reconstruction from stereo images. *Computer Optics* 2015; 39(4): 600–605. DOI: 10.18287/0134-2452-2015-39-4-600-605.
- [17] Boori MS, Kuznetsov AV, Choudhary KK, Kupriyanov AV. Satellite image analysis to evaluate the urban growth and land use changes in the city of Samara from 1975 to 2015. *Computer Optics* 2015; 39(5): 818–822. DOI: 10.18287/0134-2452-2015-39-5-818-822.
- [18] Protsenko VI, Kazanskiy NL, Serafimovich PG. Real-time analysis of parameters of multiple object detection systems. *Computer Optics* 2015; 39(4): 582–591. DOI: 10.18287/0134-2452-2015-39-4-582-591.
- [19] Porfirev AP, Kovalev AA, Kotlyar VV. Optical trapping and moving of microparticles using asymmetrical Bessel-Gaussian beams. *Computer Optics* 2016; 40(2): 152–157. DOI: 10.18287/2412-6179-2016-40-2-152-157.
- [20] Stafeev SS, Kotlyar MV, O'Faolain L, Nalimov AG, Kotlyar VV. A four-zone transmission azimuthal micropolarizer with phase shift. *Computer Optics* 2016; 40(1): 12–18. DOI: 10.18287/2412-6179-2016-40-1-12-18.
- [21] Soifer VA, Korotkova O, Khonina SN, Shchepakina EA. Vortex beams in turbulent media: review. *Computer Optics* 2016; 40(5): 605–624. DOI: 10.18287/2412-6179-2016-40-5-605-624.
- [22] Gashnikov MV, Glumov NI, Kuznetsov AV, Mitekin VA, Myasnikov VV, Sergeev VV. Image processing, pattern recognition: Hyperspectral remote sensing data compression and protection. *Computer Optics* 2016; 40(5): 689–712. DOI: 10.18287/2412-6179-2016-40-5-689-712.

# Single mode ZnO/Al<sub>2</sub>O<sub>3</sub> Strip loaded waveguide at 633 nm visible wavelength

M.A. Butt<sup>1</sup>, E.S. Kozlova<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

In this work, we proposed a technique in which two-dimensional (2-D) confinement of light is achieved in one-dimensional (1-D) planar waveguide by loading it with a low refractive index material. A waveguide design is based on ZnO/Al<sub>2</sub>O<sub>3</sub> dielectric materials for integrated photonics. These waveguides are capable of propagating TE and TM polarization at 0.633 μm visible light. Based on these waveguide structures, many optical elements such as S-bend, Y-splitter can be realized which will provide a compact platform for integrated optics.

*Keywords:* Strip-loaded waveguide; Beam propagation method; visible light; ZnO; Al<sub>2</sub>O<sub>3</sub>; Y-splitter

---

## 1. Introduction

Optical communications through fibre optics have long been the technology of choice for high-speed long distance data links [1]. Gradually, as the capacity requirements have increased, the optical links have developed into shorter distance applications, such as fiber-to-the-home, local area network, and even into fibre optic interconnects between boards and cabinets. Ultimately, optics would be used to interconnect integrated circuits on a board or even to be used in intra-chip interconnects [2, 3]. That is, electrical interconnects, which have dominated since the infancy of electronics, are likely to be substituted with optical interconnects in some cases [4]. The basic component of any optical circuit is the optical waveguide, which is able to connect different optical devices. In order to replace the microelectronic circuits, there is a need to develop integrated optical circuits that contain optical waveguides. These waveguides should be capable of confining the light in a size of the order of the wavelength. Optical waveguides can be categorized conferring to their geometry, the number of modes, refractive index distribution and material. They are designed as energy flow only along the propagation direction of the light but not perpendicular to it, therefore radiation losses can be avoided. Generally, optical integrated waveguides depend on the principle of total internal reflection, using materials with low absorption loss [5-12]. The waveguide cross section should be small to permit a high-density integration, functionally linking devices or systems or implementation of complex functionalities, such as splitters/combiners, couplers, AWGs, and modulators. A wide variety of materials can be used with their corresponding benefits and shortcomings. The improvement of optical interconnects devices including waveguides will continue through a harmonious collaboration among materials and processing technologies, design and fabrication of integrated optoelectronics, and optoelectronic packaging technology [4].

In this paper, we proposed a technique in which two-dimensional (2-D) confinement of light is achieved in one-dimensional (1-D) planar waveguide by loading it with a low refractive index material [13-15]. These waveguides are established on the effective index modification caused by a planar waveguide loaded with a material having a different refractive index, which causes a lateral confinement of light. Eventually, a lateral variation of the effective index is induced, which depends on the dimensions and the refractive index of the top structured region (cladding). As a consequence, a 2D effective index distribution is attained, which is capable of lateral confinement of light. Fig. 1, shows the schematic of the strip loaded waveguide.

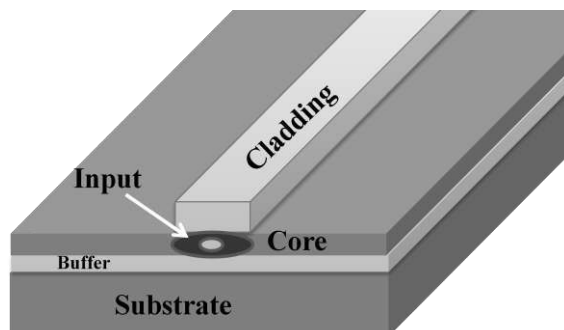


Fig.1. Schematic of strip loaded waveguide, where light propagating in the planar waveguide is confined in 2-D by loading the waveguide with low refractive index material.

## 2. Modeling of the waveguide

Nowadays, an increasing number of optical modulators, filter and other functions relevant to telecommunication networks have been proposed as integrated or embedded in waveguides [16, 17]. Many of them share the widespread feature of being

based on the propagation of the light beam inside a waveguide which has been designed to sustain only its fundamental mode of propagation to allow lower insertion losses when coupled to optical fibres. For the modeling of such waveguides, we propose ZnO ( $n=1.989$  @633 nm) [18] and  $\text{Al}_2\text{O}_3$  ( $n=1.766$  @633nm) [19] with refractive index contrast ( $\Delta n$ ) of 0.223. The materials in which the guided light propagates must avoid scattering and absorption losses in the wavelength range of interest [13]. Three layers are deposited on top of the quartz substrate with Low-High-Low refractive indices. At first thin layer of  $\text{Al}_2\text{O}_3$  is deposited which acts as a buffer while ZnO functions as a core for the propagation of light. As a final point, a cladding layer of  $\text{Al}_2\text{O}_3$  is deposited on the top of the core layer. The structuration of the cladding layer can be achieved by means of the method of [11, 20]. This structured cladding layer provides the lateral confinement of the light by a local escalation in the effective refractive index of the planar waveguide. The propagating mode is confined in the region far from the lateral walls of the ridge, which can help to avoid the losses that could arise from the roughness of the etched walls [21].

### 2.1. Dependence of the propagation power on the thickness of planar waveguide layer

Strip loaded waveguide structures were simulated by using Rsoft Beam Prop software [22] which is based on the beam propagation method (BPM) in order to obtain the optimized geometrical parameters for the propagation of a fundamental guided optical mode at  $0.633 \mu\text{m}$ . The confinement of light under the cladding structure highly depends on the thickness of the planar waveguide. We used different thicknesses of the core layer to monitor the power propagation in the area under the cladding structure. The parameters used in this analysis are shown in table 1.

Table 1. Optimization of core height of strip loaded waveguide at  $0.633 \mu\text{m}$ .

Design no.	ZnO	Al <sub>2</sub> O <sub>3</sub>	
	Core height, $\mu\text{m}$	Cladding height, $\mu\text{m}$	Cladding width, $\mu\text{m}$
1	0.5	1	3
2	0.7	1	3
3	0.9	1	3

In order to simulate the propagation of light along the z-axis, Implicit Crank-Nicolson scheme was used with a grid size of  $0.04 \mu\text{m}$  in X, Y and Z and by applying the Simple Transparent Boundary Condition (TBC). The power versus propagation distance plot for table 1 is shown in fig. 2, where CW and CH are the width and height of the cladding layer, respectively. The propagation length of the straight strip loaded waveguide is 3 mm.

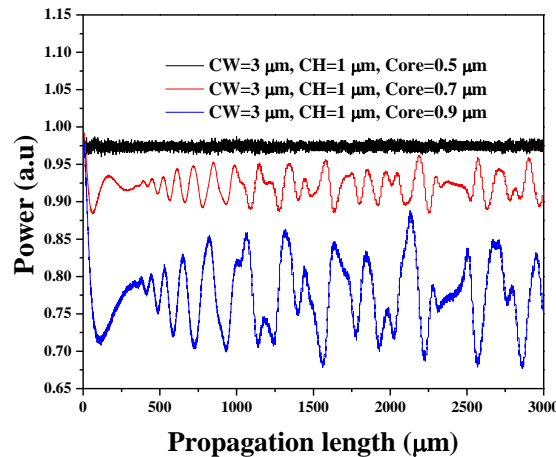


Fig. 2. Power versus propagation distance plot for the designs of straight strip loaded waveguide.

It is well noting that the planar waveguide with a small thickness is able to confine the light better than the layer with a greater thickness, having constant dimensions of cladding layer on top of it. After optimizing the core thickness, next thing is to verify the dependence of cladding width and height on the output.

### 2.2. Dependence of the propagation power on the width and height of the cladding layer

Cladding dimensions play an important role to confine the light in the planar waveguide by introducing an effective index distribution. In this section, we verify the effect of the cladding width and height on the propagation power in the core. A power monitor equals to the width of the cladding layer is placed in the planar waveguide just under the cladding region. The parameters used in this analysis are shown in table 2, where cladding width is varied by keeping the core height and cladding height constant.

Table 2. Optimization of the cladding width of strip loaded waveguide at  $0.633 \mu\text{m}$ .

Design no.	ZnO	Al <sub>2</sub> O <sub>3</sub>	
	Core height, $\mu\text{m}$	Cladding height, $\mu\text{m}$	Cladding width, $\mu\text{m}$
1	0.5	1	1
2	0.5	1	3
3	0.5	1	5



The power versus propagation distance plot for the designs (table 2) is shown in fig. 3, where CW and CH are the width and height of the cladding layer, respectively. The propagation length of the straight strip loaded waveguide is 3 mm. From fig. 3, it can be seen that, at 1  $\mu\text{m}$  of cladding width, the light spreads in the full area of the planar waveguide and cladding is unable to provide the confinement under it. That is why a drop of power is observed by the power monitor, whereas, at 3  $\mu\text{m}$ , maximum confinement of light is obtained. However, multimode appears when cladding width goes beyond 5  $\mu\text{m}$ . The width of the launch field was fixing at the width of the cladding layer (in each design) and height of the launch field was equal to the core. From the simulation, it was observed that the height of the cladding layer (thickness= 0.2, 0.5, 1, 2  $\mu\text{m}$ ) has no such influence on the confinement of the light in the planar waveguide.

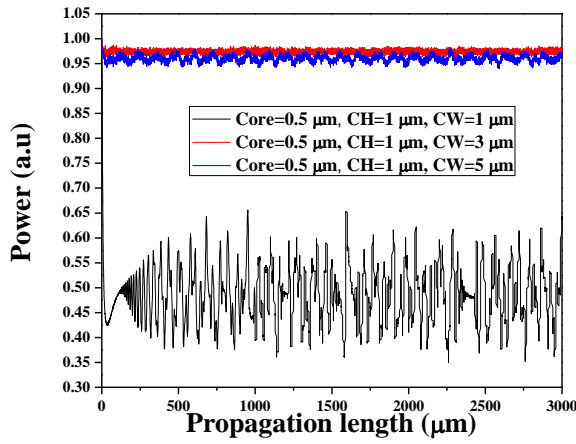


Fig. 3. Power versus propagation distance plot for the designs of straight strip loaded waveguide.

The propagation of light in a waveguide with a design number 1 (table 1) and the mode profile at the output of the waveguide is shown in fig. 4. Figure 4(a) shows the top view of the waveguide; it can be observed that the mode is travelling in the core layer confined under the structured cladding. Figure 4(b) shows the mode profile at the output of the 3 mm long waveguide.

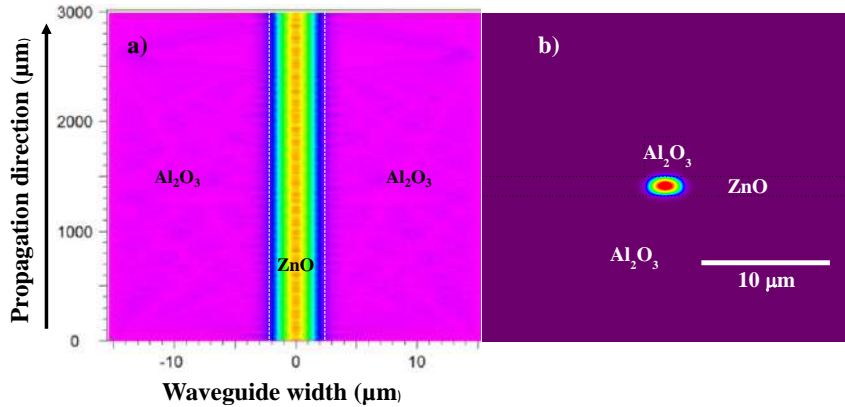


Fig. 4. Straight strip loaded waveguide (Design 1-table 1) with a fundamental mode at 0.633 microns (a) Propagation of light, (b) Mode profile at the output of the waveguide.

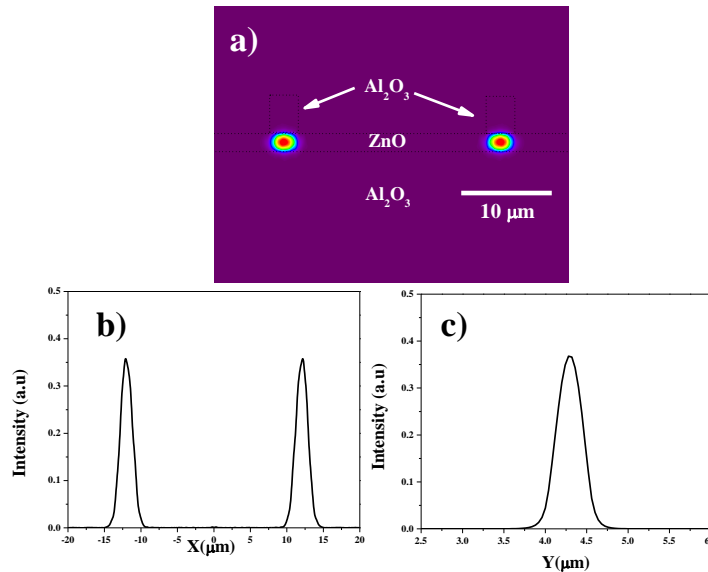


Fig. 5. Simulated results of a strip loaded Y-splitter at 0.633  $\mu\text{m}$ , core height of 1  $\mu\text{m}$ , cladding width and height of 3 and 1  $\mu\text{m}$ , respectively: (a) Mode profile at the output of Y-splitter (b) Horizontal cross-section profile of the mode (c) Vertical cross-section of the mode.



### 3. Design of a power splitter structure

The behaviour of light in S-bend was investigated by designing a Y-splitter of 10 mm in length. The separation between two arms was kept at a safe distance of 24  $\mu\text{m}$  to avoid any evanescent field coupling between two arms. Fig. 5, shows the simulation results for the Y-splitter structure at 0.633  $\mu\text{m}$  by using Beam Prop software. The S-bend is made up of two matched bends and Y-branch is very long to avoid any field distortion and loss. As is evident from the figure, the optical field at the output of the matched S-bends and in the subsequent straight waveguide section is undistorted. The intensity of 0.36 a.u/arm was obtained in case of Y-splitter with 24  $\mu\text{m}$  of separation between its arms. As a consequence, the fields at the Y-branch output are balanced. The high electromagnetic field confinement in strip loaded waveguide permits the realization of bends waveguides with a small radius of curvature.

### 4. Conclusion

We proposed a 2-D optical confinement of light in ZnO planar waveguide by loaded with a structured layer of  $\text{Al}_2\text{O}_3$ . Single mode waveguides are modeled by optimizing the core and cladding dimensions of the waveguide with the help of Beam Prop software. These waveguides are polarization independent and are able to guide light both in TE and TM polarization. Based on these waveguide structures, many optical elements such as S-bend and Y-splitter can be realized which can be used for diverse optical applications. A power splitter is also modeled for the visible wavelength of 0.633  $\mu\text{m}$  by using the optimal parameters derived from the straight strip loaded waveguide. The estimation of total losses in power splitter is nearly 27 % in terms of intensity which makes these designs suitable for integrated optics.

### Acknowledgements

This work was partly funded by RF Ministry of Education and Science ## SP-4375.2016.5, RF President's grants for leading scientific schools ## NSH-9498.2016.9 and RFBR grant ##17-47-630420.

### References

- [1] Agrell E, Karlsson M, Chraplyvy AR, Richardson DJ, Krummrich PM, Winzer P, Roberts K, Fischer JK, Savory SJ, Eggleton BJ. Roadmap of optical communications. *J. Opt.* 2016; 18: 063002 (40 p).
- [2] Yang P, Nakamura S, Yashiki K, Wang Z, Duong LHK., Wang Z, Chen X, Nakamura Y, Xu J. Inter/intra-chip optical interconnection network: opportunities, challenges and implementations, 2016, 10<sup>th</sup> IEEE/ACM International Symposium on Networks-on-Chip (NOCS), Nara, Japan, 31 Aug-2 Sept., 2016.
- [3] Kotlyar MI, Triandafilov YR, Kovalev AA, Soifer VA, Kotlyar MV, O' Faolain L. Photonic crystal lens for coupling two waveguides, *Appl. Opt.* 2009; 48: 3722–3730.
- [4] Tong XC. *Advanced materials for integrated optical waveguides.* Springer International Publishing Switzerland, 2014.
- [5] Butt MA, Pujol MC, Sole R, Rodenas A, Lifante G, Wilkinson JS, Aguilo M, Diaz F. Channel waveguides and Mach-Zehnder structures on  $\text{RbTiOPO}_4$  by  $\text{Cs}^+$  ion exchange. *Optical Material Exp.* 2015; 5: 1183–1194.
- [6] Degtyarev SA, Butt MA, Khonina SN, Skidanov RV. Modeling of  $\text{TiO}_2$  based slot waveguides with optical confinement in sharp bends. *Proceedings ICE Cube 2016*; 7495222: 10–13
- [7] Kazanskiy NL, Serafimovich PG, Khonina SN. Optical nanoresonator in the ridge of photonic crystal waveguides crossing. *Computer Optics* 2011; 35: 426–431.
- [8] Strilets TS, Kotlyar VV, Nalimov AG. Simulation of waveguide modes in multilayer structures. *Computer Optics* 2010; 34: 487–493.
- [9] Moiseeva NM. The calculation of eigenvalues modes of the planar anisotropic waveguides for various angles the optical axis. *Computer Optics* 2013; 37: 13–18.
- [10] Butt MA, Nguyen HD, Rodenas A, Romero C, Moreno P, Vazquez de Aldana JR, Aguilo M, Sole RM, Pujol MC, Diaz F. Low- repetition rate femtosecond laser writing of optical waveguides in KTP crystals: analysis of anisotropic refractive index changes. *Optics Express* 2015; 23: 15343–15355.
- [11] Butt MA, Sole R, Pujol MC, Rodenas A, Lifante G, Choudhary A, Murugan GS, Shepherd DP, Wilkinson JS, Aguilo M, Diaz F. Fabrication of Y-splitters and Mach-Zehnder structures on  $(\text{Yb}, \text{Nb})\text{:RTiOPO}_4/\text{RbTiOPO}_4$  Epitaxial layers by Reactive Ion Etching. *J. Lightw Technol.* 2015; 33: 1863–1871.
- [12] Butt MA, Degtyarev SA, Khonina SN, Kazanskiy NL. An evanescent field absorption gas sensor at mid-IR 3.39  $\mu\text{m}$  wavelength. *J. Mod. Opt.* 2017. DOI: 10.1080/09500340.2017.1325947.
- [13] Suzuki K, Ogusu K. Single-mode  $\text{Ag-As}_2\text{Se}_3$  strip-loaded waveguides for applications to all-optical devices. *Opt. Exp.* 2005; 13: 8634–8641.
- [14] Yeatman EM, Pita K, Ahmad MM. Strip-loaded high confinement waveguides for photonic applications. *J. of Sol-Gel Science and Tech.* 1998; 13: 517–521.
- [15] Martínez de Mendivil J, Hoyo J, Solís J, Pujol MC, Aguilo M, Diaz F, Lifante G. Channel waveguide fabrication technique in  $\text{KY}(\text{WO}_4)_2$  combining liquid-phase-epitaxy and beam-multiplexed fs-laser writing. *Opt. Mater.* 2015; 47: 304–309.
- [16] Kotlyar VV, Kovalev AA, Triandafilov YaR, Nalimov AG. Simulation of propagation of modes in planar gradient-index hyperbolic secant waveguide, 2011. 11<sup>th</sup> International conference on laser and fiber-optical networks modeling (LFNM). Kharkov, Ukraine, 5-9 Sept. 2011.
- [17] Kozlova ES, Kotlyar VV. Simulation of Ultrafast 2<sup>nd</sup> light Pulse. *Computer Optics* 2012; 36: 158–164.
- [18] Bond WL. Measurement of the refractive indices of several crystals. *J. App. Phys.* 1965; 36: 1674–1677.
- [19] Dodge MJ. Refractive Index" in *Handbook of Laser Science and Technology, Volume IV, Optical Materials: Part 2*, CRC Press, Boca Raton, 1986.
- [20] Sun J, Chen C, Gao L, Sun X, Gao W, Ma C, Zhang D. Polarization-insensitive strip loaded waveguide for electro-optic modulators and switches. *Optics Comm.* 2009; 282: 2255–2258.
- [21] Ding R, Jones TB, Kim WJ, Xiong X, Bojko R, Fedeli JM, Fournier M, Hochberg M. Low loss strip loaded slot waveguides in Silicon-on-insulator. *Opt. Exp.* 2010; 18: 25061–25067.
- [22] RsoftBeamPROP software. Synopsys Optical solutions 2013; 12.

# Efficient generation of a perfect optical vortex by using a phase optical element

V.V. Kotlyar<sup>1,2</sup>, A.A. Kovalev<sup>1,2</sup>, A.P. Porfirev<sup>1,2</sup>

<sup>1</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

<sup>2</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We consider generation of a perfect optical vortex by three elements: (i) amplitude-phase element with its transmission being proportional to the Bessel function, (ii) optimal element with the transmission proportional to the sign of the Bessel function, and (3) helical axicon. Maximal intensity of light on the ring is shown to be achieved by using the optimal element. Thickness of the light ring generated by the axicon is approximately two times higher than that for other elements. For perfect optical vortices with the radius of several wavelengths we detect the range of topological charges, within which the ring radius is almost independent on them.

*Keywords:* optical vortex; perfect optical vortex; topological charge; axicon; Bessel function; Fourier optics, Fraunhofer diffraction

---

## 1. Introduction

In [1], a "perfect" optical vortex (POV) has been considered. Its radius is independent on the topological charge. In [1], POVs are generated by using a phase optical element consisting of a set of concentric rings, the thickness of each of which approximates the delta function. In [2], POVs are generated by using a conical axicon and a spiral phase plate (SPP). In both works, the POV is generated approximately and its quality turned out to be low. In [3], a narrow ring is imaged by using a 4f-setup, but this way does not allow generation of a POV in the focus of a high-aperture objective for optical manipulation. In [4], instead of the axicon, an amplitude-phase optical element approximating the Bessel mode is proposed. The element in [4] is closest to the optimal phase filter proposed in our work. Generation of POV is compared by using 1) an amplitude-phase light field described by a Bessel mode of limited radius, 2) by an optimal phase light field, described in [5] and 3) by the phase field generated by using a conical axicon and SPP [2].

## 2. Generation of the "perfect" optical vortex by using an amplitude-phase optical element

"Perfect" optical vortex [1] has the following complex amplitude:

$$E_0(\rho, \theta) = \delta(\rho - \rho_0) \exp(in\theta), \quad (1)$$

where  $\delta(x)$  is the Dirac delta-function,  $(\rho, \theta)$  are polar coordinates in the Fourier-plane of a spherical lens,  $n$  is the vortex topological charge. It is seen in Eq. (1) that the radius of an infinitely thin ring  $\rho_0$  does not depend on the topological charge. The POV (1) can be generated by using an ideal Bessel mode in the focal plane of the lens:

$$F_0(r, \varphi) = J_n(\alpha r) \exp(in\varphi), \quad (2)$$

where dimensional parameter  $\alpha = k\rho_0/f$  determines the scale of the  $n$ -th order Bessel function of the first kind  $J_n(x)$ . radius of the ring with maximal intensity is  $\rho_0 = \alpha f/k$ , where  $k$  is the wavenumber of a monochromatic coherent light,  $f$  is the focal length of the lens. For generation of the field (2) and amplitude mask is needed. In addition, the Bessel function in Eq. (2) is in practice bounded by a circular aperture or illuminated by a Gaussian beam. Both these reasons distort the POV and lead to low energy efficiency, weak dependence of the on-ring maximal intensity on the topological charge, and to a wider ring.

## 3. Generation of the "perfect" optical vortex by using an optimal phase optical element

Here under the optimal element we mean such optical element that directs the major part of the light into the ring of a specified radius. Transmission of such element is [5]:

$$F_2(r, \varphi) = \text{circ}\left(\frac{r}{R}\right) \text{sgn} J_n(\alpha r) \exp(in\varphi). \quad (3)$$

In the focus of the lens, the field from the complex amplitude (3) reads as

$$E_2(\rho, \theta) = (-i)^{n+1} \left( \frac{k}{f} \right) e^{in\theta} \sum_{m=0}^{N-1} (-1)^m \int_{r_m}^{r_{m+1}} J_n \left( \frac{k\rho r}{f} \right) r dr, \quad (4)$$

where  $r_0=0$ ,  $r_m = \gamma_{n,m}/\alpha$ ,  $m=1,2,\dots,N$ ,  $r_N = R$ . Putting  $\rho = \alpha f/k$  in Eq. (4), the argument of the Bessel function becomes independent on the parameters  $f$  and  $k$  and equals  $\alpha r$ . It can be shown that at  $\rho_0 = \gamma_{n,N} f/(kR)$  all terms in the sum are positive and their contribution into the light field on the ring is maximal. For the radius of the POV ring to be independent of the topological charge, close roots of the Bessel function need to be chosen:  $\gamma_{n,N} = \gamma_{m,M}$ .

#### 4. Generation of a "perfect" optical vortex by using a axicon

The POV is often generated by using a conical axicon and a SPP [2]. When light passes through such setup, it is equivalent to passing an element with the following transmittance function:

$$E_3(r, \varphi) = \text{circ} \left( \frac{r}{R} \right) \exp(i\alpha r + in\varphi). \quad (5)$$

For the first time, optical element (5) was considered in [6] for generation of light pipes. It was also studied in [7-9]. Instead of scaling factor of the Bessel function, here the parameter  $\alpha$  determines an axicon parameter, related with the vertex-angle of the generated conical wave. It is supposed in [2] that in the focus of a spherical lens the light field (5) generates a POV with its amplitude distribution described by the function  $E_3(\rho, \theta) \exp[-(\rho - \rho_0)^2 / \Delta\rho^2] \exp(in\theta)$ . It is a simplification and the complex amplitude of the POV is defined by a Fourier transform of the function (5), taking a much more complex form [6]. The intensity on the ring is lower than that for the optimal element (3), but higher than for the amplitude element. The intensity on the ring depends on the topological charge, while the ring thickness is approximately two times larger than in previous cases.

#### 5. Simulation results

In this section, we describe the simulation results of generating the POV by using the considered above optical elements. The simulation parameters are as follows: wavelength of light  $\lambda = 532$  nm, circular aperture radius is  $R = 20$  mm, and the focal length of an ideal spherical lens  $f = 100$  mm, while the Bessel function's scale factor  $\alpha$  is chosen so that  $\alpha R = \gamma_{1,20} = 63,6114$ , where  $\gamma_{1,20}$  is 20th zero of the first-order Bessel function ( $\nu=20$ ,  $n=1$ ). The POV was simulated for two topological charges:  $n = 1$  and  $n = 14$ , while the other parameters were kept unchanged. Note that for the Bessel function of order  $n = 14$  we chose the 14th root ( $\nu = 14$ ) because  $\gamma_{14,14} \approx \gamma_{1,20} = 63.6114$ . Figure 1 shows the absolute values of two Bessel functions,  $|J_1(\gamma_{1,20}x/R)|$  and  $|J_{14}(\gamma_{14,14}x/R)|$ . It is seen in Fig. 1 that both functions are seen to take a zero value at  $x = R$ .

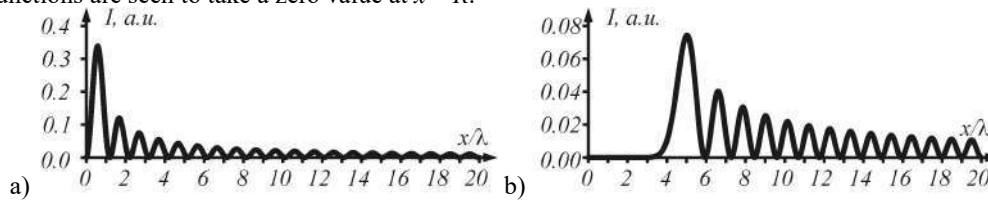


Fig. 1. Absolute values of the Bessel functions  $|J_1(\gamma_{1,20}x/R)|$  (a) and  $|J_{14}(\gamma_{14,14}x/R)|$  (b), bounded by the radius  $R$ .

Fig. 2 shows intensity distributions of the POV in the Fourier plane of a spherical lens, obtained with an initial light field in the form of a bounded Bessel mode. Parameters of the calculated POVs are given in Table 1. It is seen in Table 1 that the POV radius was not changed with changing of the topological charge. At the chosen parameters, the POV radius is  $\rho_0 \approx 50,62\lambda$ . This value is different from the value in Table 1 by only 3%. Maximal intensity of the POV decreased by only 5% with an increase in the vortex topological charge by almost an order of magnitude. Note that, according to the theory, with the chosen simulation parameters the maximum intensity in Fig. 2 should be equal to  $I_1(\rho_0) = [kR/(\pi\alpha f)]^2 \approx 0,015816$ . It is consistent with the value in Table 1. Since the radius of the ring has not changed and the radius of the aperture  $R$  of the optical element has not changed either, the width of the ring should not change. Table 1 shows that, indeed, the ring thickness does not change when the topological charge of the optical vortex changes. According to the theory, the ring thickness at the selected simulation parameters should be equal to  $\text{FWHM} = 5/2\lambda$ . This value differs by 11% from the value of the ring thickness in Table 1.

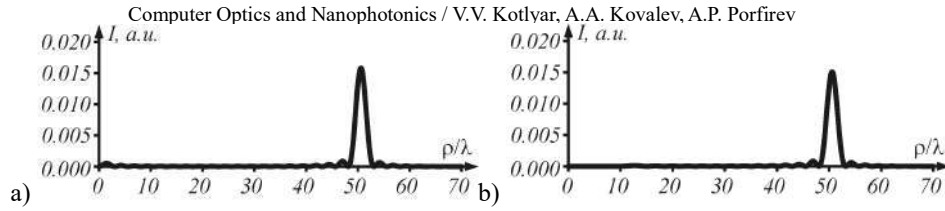


Fig. 2. Intensity distributions of the POV with  $n=1$  (a) and  $n=14$  (b) for the initial light field in a form of the Bessel mode.

Table 1. Comparison of the parameters of POV, generated with the initial optical field in Eq. (6) (bounded Bessel function) at different topological charges  $n$ .

Topological charge	$n=1$	$n=14$
Radius of maximum intensity ring, $\rho_0, \lambda$	50.781563	50.781563
Maximum intensity, $I_{\max}$ (a.u.)	0.0157968	0.0150522
Ring thickness, FWHM, $\lambda$	2.244489	2.244489

Now we consider generation of the POV by using the optimal phase element (3). Fig. 3 shows the intensity distributions for the initial light field (3). Table 2 shows the computed parameters of the POV from Fig. 3. From Table 2, the POV radius became slightly smaller than that in Fig. 2 (less by just 0.3%). The radius remained almost unchanged when the topological charge was increased by a factor of 14. The intensity at the ring is almost 100 times greater than the intensity for the POV in Fig. 2. We note that with an increase of the topological charge by a factor of 14, the intensity on the ring decreased by only 2%. The ring thickness became smaller by approximately 14% compared to the thickness of the ring in Fig. 2. The thickness of the ring is preserved when the topological charge of the optical vortex changes. Fig. 3 shows that side lobes have increased.

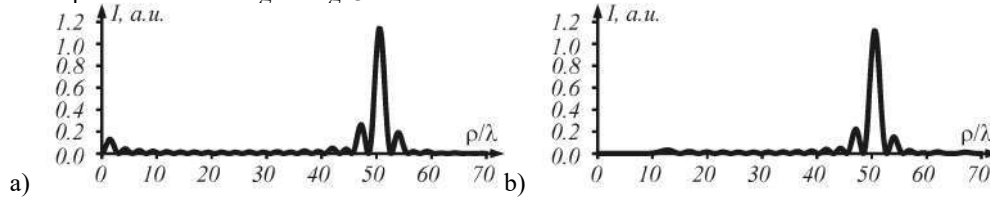


Fig. 3. Intensity distributions of the POV at  $n=1$  (a) and  $n=14$  (b) for the initial light field (3).

Table 2. Comparison of parameters of the POV, generated with the optimal phase element [Eq. (3)] at different topological charges  $n$ .

Topological charge	$n=1$	$n=14$
Radius of maximum intensity ring, $\rho_0, \lambda$	50,641283	50,641283
Maximum intensity, $I_{\max}$ (a.u.)	1,140685	1,1181689
Ring thickness, FWHM, $\lambda$	1,9639279	1,9639279

Next, we consider generation of the POV by using a helical axicon (5). Fig. 4 shows intensity distributions for the initial light field (5), while the Table 3 shows the computed parameters of this POV. From Fig. 4 and Table 3, the POV ring thickness is approximately 2.5 times larger than the thickness of the ring in Fig. 2.

In addition, with increasing topological charge of the vortex, the ring thickness increases by a factor of 1.3.

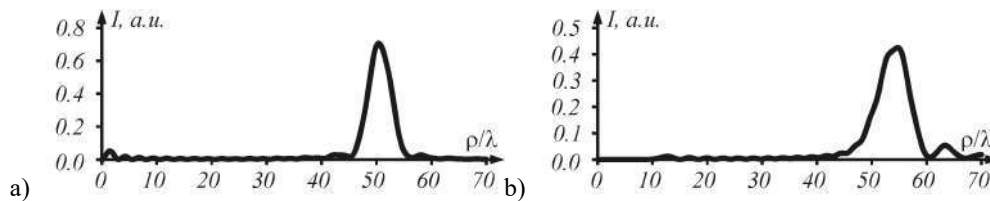


Fig. 4. Intensity distributions of the POV at  $n=1$  (a) and  $n=14$  (b) for the initial light field (5).

Table 3. Comparison of Parameters of the POV generated by using a spiral axicon [Eq. (25)] at different topological charges  $n$ .

Topological charge	$n=1$	$n=14$
Radius of maximum intensity ring, $\rho_0, \lambda$	50,501002	54,849699
Maximum intensity, $I_{\max}$ (a.u.)	0,7070332	0,4249419
Ring thickness, FWHM, $\lambda$	4,9098196	6,5931864

An increase of the ring thickness (Fig. 4) with increasing  $n$  leads to the decreasing intensity on this ring. From Table 3, it is seen that the maximal intensity on the ROV ring (Fig. 3) decreases 1.7 times as the number  $n$  increases by a factor of 14. And even the radius of the maximal intensity ring increases by 8%. In this case, the thickness of the ring and the maximum intensity almost do not change with increasing topological charge of the optical vortex from 1 to 14. The only drawback of the POV generated by the element (3) is an increased level of side lobes, which constitute about 20% of the maximum intensity.

So, the simulation has shown that among the three optical elements for generating the POV, the optimal phase element in Eq. (3) is the best one, since the narrowest ring ( $\text{FWHM} = 1,96\lambda = 0.39\lambda f/R$ ) is generated in this case with the maximum intensity being 1.6 times higher than that from the spiral axicon in Eq. (5).

Above, the topological charge  $n$  and the scaling factor of the axicon  $\alpha$  were chosen so that the product  $\alpha R$  remained approximately the same and was equal to the root of the Bessel function. Now we consider the case when this is not so. Let this product be arbitrary and let the optimal element or axicon with a diameter of  $2R = 40\lambda$  to generate a POV by using a lens with a focal length  $f = 100\lambda$ . Fig. 5 shows the dependence of the light ring thickness (at half-maximum of the intensity) on the radius of the ring.

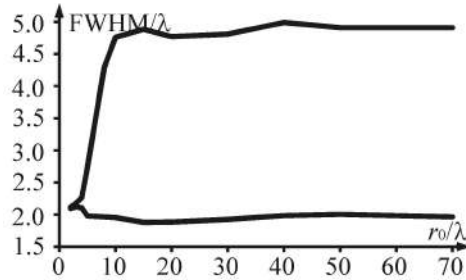


Fig. 5. Thickness of light ring vs. its radius. Upper curve – axicon (5), lower curve – optimal element (3).

For the POV radius  $r_0 = 2\lambda$ , both elements work identically. Both of them generate a light ring with a radius of about  $r_0$  and with about same thickness. At the same time, for the optimal element the maximal energy is approximately 30% higher. This is explained by the fact that, despite the same width at the level of half-maximum, the thickness at a lower intensity level is larger for the axicon. When  $r_0 = 3\lambda$ , both elements form a ring of wrong radius of about  $2\lambda$ . However, the maximum energy for the optimal element is 86% higher. When  $r_0 = 4\lambda$ , the optimal element generates a ring of radius  $4.1\lambda$ , while the axicon generates two light rings, one of which has a radius  $1.9\lambda$ , and the second -  $4.7\lambda$ . When  $r_0 = 5\lambda$ , both elements generate two light rings with radii  $5\lambda$  and  $1.5\lambda$ . Further, for  $r_0 > 5\lambda$ , both elements generate a light ring of radius  $r_0$ . With increasing  $r_0$  up to  $r_0 = 70\lambda$ , the ring thickness remains practically unchanged, but in all cases the ring generated by the axicon is about 2 to 2.5 times wider.

Now we consider the dependence of the radius of the generated light ring on the topological charge. Let the optimal element (3) or the axicon (5) of the diameter  $2R$  to generate a POV by using a lens with a focal length  $f = 100\lambda$ . Fig. 6 shows the dependence of the radius of the generated light ring on the topological charge for elements of radius  $R = 10\lambda$  (Fig. 6a),  $R = 20\lambda$  (Fig. 6b), and  $R = 30\lambda$  (Fig. 6c).

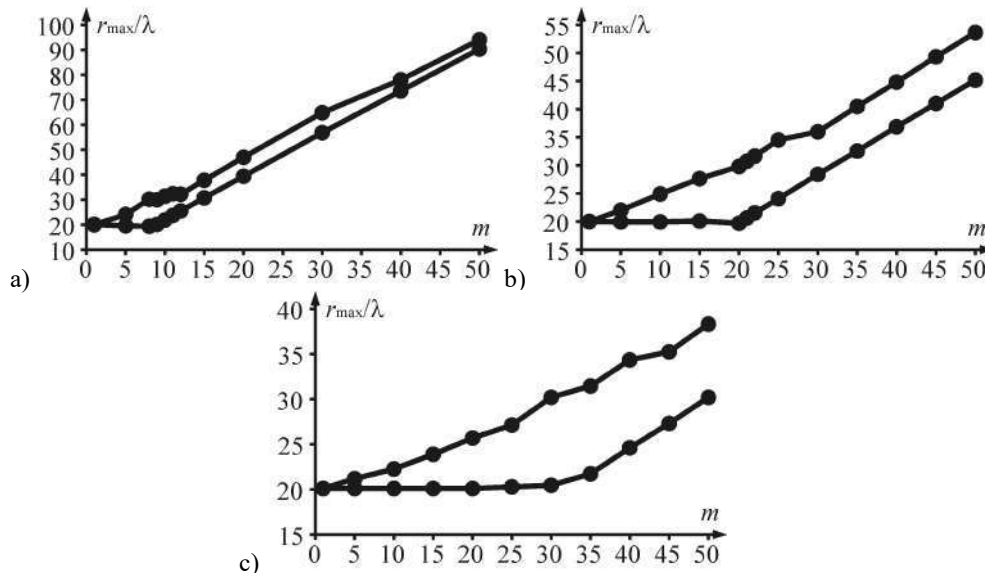


Fig. 6. Radius of the generated light ring vs. the topological charge for different element radii:  $R = 10\lambda$  (a),  $R = 20\lambda$  (b), and  $R = 30\lambda$  (c). Upper curve – axicon (5), lower curve – optimal element (3).

It is seen in Fig. 6 that with using the axicon (5) the light ring radius rises with the topological charge nearly linearly. When using the optimal element (3), the radius is almost constant for the topological charges up to  $R/\lambda$ . With larger topological charges, the radius begins to increase linearly at about the same rate as for the axicon.

## 6. Experiment

For experimental study of the optical elements for generating the POVs we used an optical setup shown in Fig. 7. The fundamental Gaussian beam was a light source generated by a solid-state laser  $L$  ( $\lambda=532$  nm). The laser beam was expanded and collimated using a system composed of a 40- $\mu\text{m}$  pinhole  $PH$  and a lens  $L_1$  ( $f_1 = 250$  mm). Then the beam illuminated the display of a spatial light modulator SLM (PLUTO VIS,  $1920 \times 1080$  resolution, with 8  $\mu\text{m}$  pixels). The diaphragm  $D_1$  was used to separate the central bright spot from the surrounding dark and bright rings caused by diffraction by the pinhole. Further, using the lenses  $L_2$  ( $f_2 = 350$  mm) and  $L_3$  ( $f_3 = 150$  mm) and the diaphragm  $D_2$ , spatial filtering of the phase-modulated laser beam reflected at the SLM display was performed. Using a lens  $L_4$  ( $f_4 = 500$  mm), the laser beam was focused on the CCD array of a video camera LOMO TC 1000 (pixel size  $1.67 \times 1.67$   $\mu\text{m}$ ). To generate the POVs, we used the phase masks shown in Fig. 8, which were output to the SLM display. To separate the non-modulated beam reflected at the display and the phase-modulated beam, a linear phase mask was superimposed on the initial phase mask.

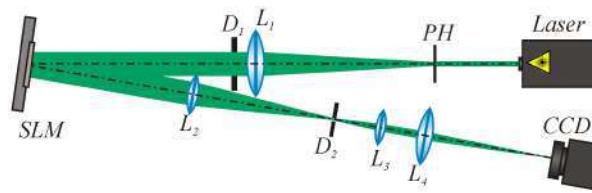


Fig. 7. Experimental setup:  $L$  is a solid-state laser ( $\lambda = 532$  nm);  $PH$  is a 40- $\mu\text{m}$  pinhole;  $L_1, L_2, L_3$ , and  $L_4$  are lenses with focal lengths  $f_1 = 250$  mm,  $f_2 = 350$  mm,  $f_3 = 150$  mm, and  $f_4 = 500$  mm;  $D_1$  and  $D_2$  are diaphragms; SLM is a spatial light modulator PLUTO VIS; and CCD is a video camera LOMO TC-1000.

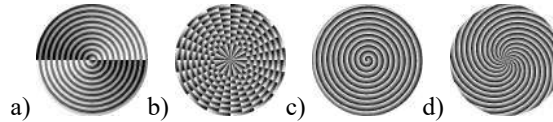


Fig. 6. Phase masks of optical elements to generate a POV with a topological charge (a, c)  $n = 1$  and (b, d)  $n = 14$ . (a) and (b) depict optimal phase elements and (c) and (d) are for spiral axicons.

Fig. 7 shows the intensity distributions in the focus of lens  $L_4$  generated by using phase masks for the optimal phase elements with topological charges  $n = 1$  and  $n = 14$ . The values of the parameters of the resulting POV are given in Table 4.

Figure 8 depicts the intensity distributions in the focus of lens  $L_4$  generated using phase masks corresponding to the spiral axicons with  $n = 1$  and  $n = 14$ . The values of the parameters of the resulting POV are given in Table 5. Analyzing the experimentally measured parameters of the POV, we can conclude that their relative values are in good agreement with the simulation results.

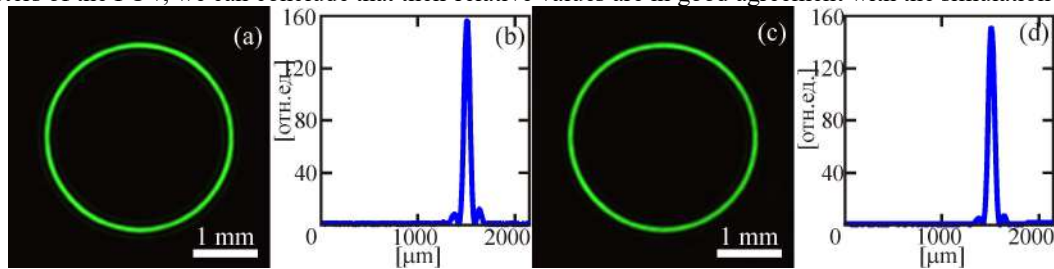


Fig. 9. Intensity distributions of the POV (a, c) and respective profiles depicted from the center to the edge (b, d) obtained by using an optimal phase element with (a, b)  $n = 1$  and (c, d)  $n = 14$ .

Table 4. Comparison of parameters of POV obtained by using an optimal phase element with topological charges  $n = 1$  and  $n = 14$ .

Topological charge	$n = 1$	$n = 14$
Radius of maximum intensity ring, $\mu\text{m}$	$1491.0 \pm 2.0$	$1496.5 \pm 2.0$
Maximum intensity, a.u.	$156.0 \pm 0.5$	$151.0 \pm 0.5$
Ring thickness, FWHM, $\mu\text{m}$	$70.0 \pm 2.0$	$73.0 \pm 2.0$

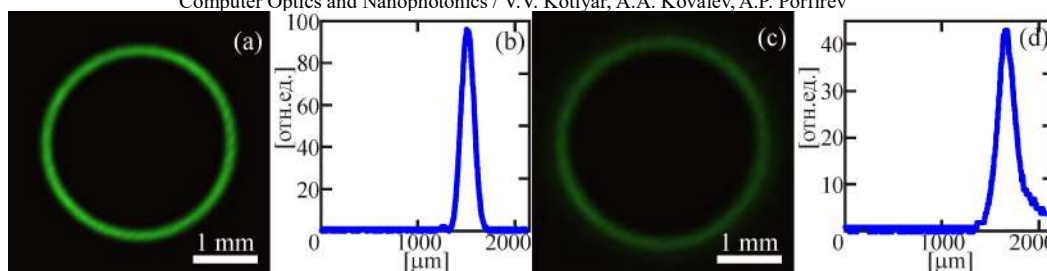


Fig. 10. Intensity pattern of a POV (a, c), with the respective profiles depicted from the center to the edge (right b, d) for a spiral axicon with (a,b)  $n = 1$  and (c,d)  $n = 14$ .

Table 5. Comparison of parameters of POV obtained by using a spiral axicon with topological charges  $n = 1$  and  $n = 14$ .

Topological charge	$n = 1$	$n = 14$
Radius of maximum intensity ring, $\mu\text{m}$	1498.0 $\pm$ 2.0	1655.0 $\pm$ 2.0
Maximum intensity, a.u.	96.0 $\pm$ 0.5	43.0 $\pm$ 0.5
Ring thickness, FWHM, $\mu\text{m}$	158.0 $\pm$ 2.0	206.0 $\pm$ 2.0

## 7. Conclusion

In this work, generation of a perfect optical vortex by three different optical elements is considered: amplitude-phase element with a transmission proportional to the Bessel mode, an optimal phase element and a vortex axicon. It is shown that using any of these three optical elements leads to generation of light rings with the same radius, which depends little on the topological charge of the optical vortex. However, if the POV radius is equal to several wavelengths, then it depends on the topological charge. For the axicon, this dependence is almost linear, while for the optimal element this dependence is almost absent for topological charges that are smaller than the ratio of the element's radius to the wavelength. The intensity of light on the ring is greater (with other conditions being equal) for the optimal phase element. The intensity of all three rings depends little on the topological charge. The thickness of the light ring generated by the vortex axicon is approximately twice larger compared to that of the other two rings. Thus, the optimal filter (3), studied for the first time in [3], is the best candidate for generation of a perfect optical vortex.

## Acknowledgements

The work was funded by the Russian Science Foundation (RSF) grant # 17-19-01186 (the results pertaining to the experimental generation of perfect optical vortices presented in Section 6), as well as by the Russian Foundation for Basic Research (RFBR) grant # 17-47-630420 (the results pertaining to the numerical simulation of the perfect optical vortices generated by different elements presented in Sections 1-5).

## References

- [1] Ostrovsky AS, Rickenstorff-Parrao C, Arrizon V. Generation of the "perfect" optical vortex using a liquid-crystal spatial light modulator. *Optics Letters* 2013; 38(4): 534–536.
- [2] Chen M, Mazilu M, Arita Y, Wright EM, Dholakia K. Dynamics of microparticles trapped in a perfect vortex beam. *Optics Letters* 2013; 38(22): 4919–4922.
- [3] Li P, Zhang Y, Liu S, Ma C, Han L, Cheng H, Zhao J. Generation of perfect vectorial vortex beams. *Opt. lett.* 2016; 41(10): 2205–2208.
- [4] García-García J, Rickenstorff-Parrao C, Ramos-García R, Arrizón V, Ostrovsky A. Simple technique for generating the perfect optical vortex. *Optics Letters* 2014; 39(18): 5305–5308.
- [5] Fedotowsky A, Lehovec K. Optimal filter design for annular imaging. *Applied Optics* 1974; 13(12): 2919–2923.
- [6] Kotlyar VV, Kovalev AA, Skidanov RV, Moiseev OY, Soifer VA. Diffraction of a finite-radius plane wave and a Gaussian beam by a helical axicon and a spiral phase plate. *Journal of the Optical Society of America A.* 2007; 24(7): 1955–1964.
- [7] Kotlyar VV, Kovalev AA, Cojoc D, Garbin V, Ferrari E. Diffraction of a Gaussian beam by a spiral axicon. *Computer Optics* 2006; 30: 30–35.
- [8] Kotlyar VV, Kovalev AA, Soifer VA, Davis JA, Tuvey CS, Cottrell DM. Diffraction of a finite-radius plane wave by a spiral axicon and by a spiral phase plate: comparison. *Computer Optics* 2006; 30: 36–43.
- [9] Degtyarev SA, Khonina SN, Podlipnov VV. Formation of spiral intensity by binary vortical axicon. *Computer Optics* 2014; 38(2): 237–242.

# Symmetric encryption algorithm using “twisted” light

S. A. Burlov<sup>1</sup>, A. V. Gorokhov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoye shosse, Samara, 443086, Russia

---

## Abstract

An algorithm for applying a “twisted” light for constructing an encryption scheme is described. Our approach is founded on famous classical symmetric permutation algorithm based on NP-full task for “Knapsack Problem” with changes taken into account the quantum origin of the information carrier. As a measuring device for selection of pure states from a mixed one, the Mach-Zehnder interferometer cascade is supposed to use, which allows sorting the parity of the mixed state of the orbital angular momentum (OAM) of photons.

*Keywords:* quantum cryptography, encryption algorithm, orbital angular momentum of photons, “twisted” light

---

## 1. Introduction

The modern bit cryptography is developing rapidly due to the active development of information storage and transmission devices. The search for good algorithms among algebraic structures leads to the known problems of discrete logarithm and factorization, which have a large history of applications in cryptanalysis.

The quantum cryptography was appeared, potentially having unlimited information capacity, huge transmission speed and stability, based on the laws of quantum physics. Among the types of algorithms for quantum cryptography are known only a schemes of key distribution and their using is considered mainly as an auxiliary position. The possibilities of quantum encryption are not yet fully disclosed, but steps to this are done every day. At present widely used the quantum two-dimensional systems based on the particles spin states and polarization states of photons. The main goal of this paper consist in the use for encryption a potentially infinite-dimensional quantum systems based on the states of orbital angular momentum of photons [1].

## 2. Orbital angular momentum of photon

The light beams with an azimuthal phase that depends on a complex factor  $\exp(-il\phi)$  carry an orbital angular momentum. The angle  $\phi$  is the azimuthal coordinate in the cross section of the beam, and  $l$  can be any integer number. The value of  $l$  indicates the amount of twist of the spiral phase front. The value OAM is equal to  $L = l \cdot \hbar$  per photon [1].

Many researches of this phenomenon are connected with a certain type of light beam - the Laguerre-Gaussian mode. In works [2, 3] showed the modification scheme of the quantum key distribution (QKD), the transmission of information with superposition of states with non-zero OAM values of photons [4]. Many works are related to the generation of the light beams with OAM [5, 6, 7, 8]. The main difficulty in the practical use of this phenomenon consist in the problem of measurement the OAM value of a photon and in search of the appropriate transmission medium for such beams. Some methods have been developed, which allows to measure OAM value of a photons: the measurement method with generating hologram [9], the sorting method using the Mach-Zehnder interferometers cascade [10, 11, 12], the method of optical geometric transformation [13, 14] and etc.

In this paper we offer to use a method that uses generating holograms and cascade of Mach-Zehnder interferometers. It is proposed to construct a measurement scheme in a such way as to minimize the uncertainty of the receiving beam. The absence of photons at the output of the detector is also an useful unambiguous information for the process.

It is well known that when studying the states of photons with an orbital angular momentum, we get to an infinite-dimensional Hilbert space, which is formed from the set of eigenstates of the operator  $\widehat{L}_z$ :

$$\widehat{L}_z = i \frac{\partial}{\partial \phi} \quad (1)$$

In principle, states with an arbitrary OAM value may be generated in an experiment. In the paper [15] it is shown the possibility of continuous beam generation with various values of OAM using computer-controlled holograms.

## 3. Merkley’s scheme

The basis of algorithm of the Merkley’s scheme is a secret super-growing sequence of the natural numbers



$$A = \{a_1, a_2, \dots, a_k\}, \quad \text{where } a_j \geq \sum_{i=1}^{j-1} a_i, \quad (2)$$

which distributed between the subscribers of the network (Alice, Bob, ...) and pair of numbers  $n$  and  $w$

$$n, w \in \mathbb{N}, n > 2 \cdot a_k, \quad \text{GCD}(n, w) = 1. \quad (3)$$

Here  $\text{GCD}(n, w)$  means the greatest common divisor of the numbers  $n, w$ , and the number  $n$  is greater than the sum of elements of the sequence (2) [16]. Next, the numbers  $n$  and  $w$  create the new sequence according to the rule:

$$G = \{g_1, g_2, \dots, g_k\}, \quad \text{where } g_j = a_j \cdot w \pmod{n}. \quad (4)$$

An original message is divided into blocks of bits of length  $k$

$$M = \{m_1, m_2, \dots, m_n\}, j \in \overline{1..n} \Rightarrow \{M_i\} = \{m_{i1}, m_{i2}, \dots, m_{ik}\}, i \in \overline{1..[\frac{n}{k}]}. \quad (5)$$

After it the corresponding sum is calculated

$$c_i = \sum_{j=1}^k g_j \cdot m_{ij}. \quad (6)$$

This number is a block of encrypted text that is transmitted to another subscriber of a network. In its turn, the receiver calculates the value  $f_i$  from the obtained value  $c_i$  given by expression (6).

$$f_i = c_i \cdot w^{-1} \pmod{n} \quad (7)$$

This number is decomposed on the sequence (2) basis and as result the original message is obtained. These actions are performed for all blocks. The reliability and validity analysis of this scheme can be found, for example, in the article [17].

#### 4. Adapted Merkle's schemes

Let the secret sequence (2) and secret numbers (3) are distributed between subscribers of the network. The permutation sequence  $T$  is formed by the sequence (4)

$$T = \sigma(G) = \{g_{j1}, g_{j2}, \dots, g_{jk}\}, \quad \text{where } g_{j1} < g_{j2} < \dots < g_{jk} \quad (8)$$

using the substitution

$$\sigma = \begin{pmatrix} 1 & 2 & \dots & k \\ j1 & j2 & \dots & jk \end{pmatrix}. \quad (9)$$

The control device for spatial light modulator (SLM) is being configured to generate laser beams with OAM photon projection only for values from the set  $T$ . The generation of target beams can be realized, for example, using computer-controlled holograms of diffraction gratings according to Refs [9].

The opentext is converted to a bit string. Each block is processed separately. The  $i$ -th iteration is performed as follows:

$$B_i = \sigma(M_i) = \{b_{i1}, b_{i2}, \dots, b_{ik}\}. \quad (10)$$

Schematically, the design of the sender and receiver of the encryption process is shown in Fig. 1.

The digital-to-analog converter (DAC) specify SLM to generate the required beam type. Below are two versions of the encrypted text generation that correspond to light rays with OAM of different types. The measurement block is also different for each option, but the result of his work is the same: we get a list of values, which were laid in the ciphertext. This data is transferred to the computer and deciphered by computing of expression (7). The resulting number is decomposed based on the sequence (2).

##### 4.1. Variant I

Here, in order to encrypt the transmitted text, it is suggested to use a mixed state, which corresponds to superposition

$$|\Psi\rangle = \sum_{i=1}^k a_i \cdot b_i \cdot |OAM = g_{ji}\rangle, \quad (11)$$

here factors  $a_i$  are given by the sender and factors  $b_i$  are calculated in accordance to the bit decomposition (10), and

$$f_i = c_i \cdot w^{-1} \pmod{n}. \quad (12)$$

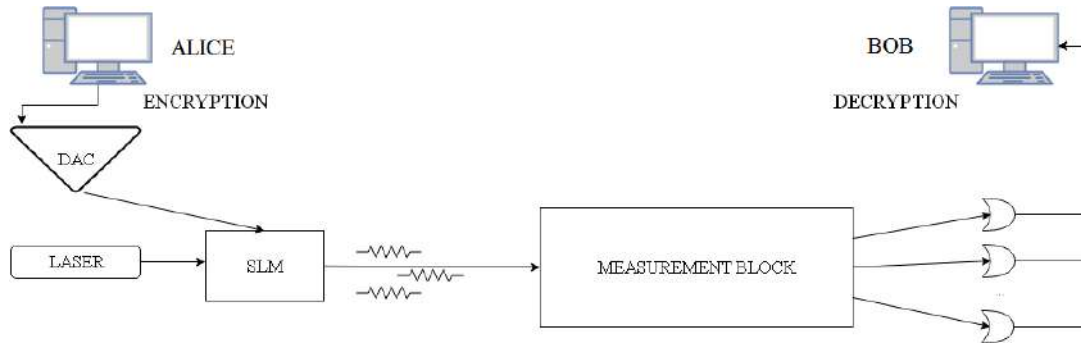


Figure 1: Scheme of the process of formation, transmission and measurement of packages

It is necessary to obtain a mixed state with the density matrix  $\rho$ . In general, the density matrix has  $k^2$  elements, but for a mixed state only the diagonal elements can differ from zero, which correspond to the elements of the sequence (4).

$$\rho = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \alpha_1^2 & \dots & 0 & \dots & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \dots & \alpha_2^2 & \dots & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & \dots & 0 & \dots & \alpha_k^2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}. \quad (13)$$

The SLM control unit generates a mixed state (11) and transmits it during the iteration period. In this case, the input measurement block should detect which states are participating in the generation of the mixed state. According to [10], the cascade of the Mach-Zehnder interferometers can “do this work”. But there is one important feature: to determine  $2^p$  states,  $2^p - 1$  interferometers required.

To optimize the measurement, it is proposed to use a short cascade. Optimization is based on the fact that for sorting out  $k$  values (knowing these values), each photon will pass no more than  $p$  interferometers. In total, we need a maximum of  $k \cdot p$  interferometers. Therefore, when building a cascade, one can block empty paths, thereby greatly reducing the number of constituent elements.

Having received the statistics, one need to select those indicators that satisfy the specified threshold values, find their sum, which corresponds to (6), then calculate the expression (7) and get the source text.

#### 4.2. Variant II

In this variant it is proposed to use a sequence of pure OAM photons states as an encryption text. The SLM control unit, before starting the transmission, gives the beamforming device a control signal for sending the zero Gaussian mode to the receiver. The receiver and the sender should be synchronized during the iteration period -  $\nu$  of the beams sequence transmission. When the sequence  $B'_i$  is obtained during the time interval  $\tau = \frac{\nu}{k}$ , depending on the value of 0 or 1, the SLM sends a pure state corresponding to  $g_{ij}$  or its inversion.

Reception is carried out after receiving the signal state, which can be a zero Gaussian mode, and during a time interval  $\frac{\nu}{k}$  the detector perceives the beam with predetermined OAM value. If it is not detected, 0 is sent. Each iteration needs a time interval equal  $\nu$ . After the successful transfer of one packet the sum of indicators that are assumed to be equals to 1 is calculated.

$$c_j = \sum_{i=1}^k b'_i \cdot g_i. \quad (14)$$

Then, the expression (7) should be calculated and the resulting number is decomposed using a basis of the secret sequence (2). As result, the opentext block is obtained. Having received all the blocks and deciphering them, the recipient decrypts the transmitted message.

## 5. Conclusion

Described encryption scheme is symmetric scheme due to restrictions imposed earlier, so that for effective measurement it is necessary to minimize the uncertainty of the received signal for legal subscribers. This can be done primarily due to the fact that the legal subscriber knows what sequence and what physical signals should be received and the messages themselves are unknown a priori.

The persistence of the presented variant I is determined by the durability of the classical Merkle's scheme. The reliability of the variant II schema is determined by a stability of the permutational interrelations, which are used to calculate the transmitted

sequence: the probability of determining key is equal  $\frac{1}{k!}$ . Therefore the length of the original sequence should be optimal. Optimum in this case is understood as a weighting between the length of the cipher sequence (2) and the maximum index of the OAM of the beam, which will be detected with a minimum error. Based on the maximum "well" detectable value  $f$  of the beam orbital angular momentum, the length of the bit sequence can not exceed the value of  $\log_2(f)$ , whereas the maximum length is reached for the "bad" superincreasing sequence  $\{1, 2, 4, 8, 16, 32, \dots\}$ .

For an eavesdropper, obtaining a stream without accurate detection does not provide any information about the signal being transmitted, because the zeros of the sequence are sent also by a non-zero OAM value. Negative sign of the projection of the orbital angular momentum also needs to be revealed, for this the eavesdropper will be given a very short time interval (therefore it is important that the carrier can not be uniquely stored).

## References

- [1] L. Allen, M. W. Beijersbergen, R. J. C. Spreeuw, J. P. Woerdman, Allen, L. Orbital angular momentum of light and the transformation of laguerre-gaussian laser modes, *Phys. Rev.* 45 (1992) 8185–8189.
- [2] R. W. Boyd, A. Jha, M. Malik, C. O'Sullivan, B. Rodenburg, D. J. Gauthier, Boyd, R. W. Quantum key distribution in a high-dimensional state space: exploiting the transverse degree of freedom of the photon, *Advances in Photonics of Quantum Computing, Memory, and Communication IV. Proc. of SPIE* Vol. 7948. (2011) 79480L–1 79480L–6.
- [3] M. Mirhosseini, O. S. Magana-Loaiza, M. N. O'Sullivan, B. Rodenburg, M. Malik, M. P. J. Lavery, M. J. Padgett, D. J. Gauthier, B. R. W., Mirhosseini, M. High-dimensional quantum cryptography with twisted light, *New J. Phys.* 17 (2015) 1–11.
- [4] M. Krenn, R. Fickler, M. Fink, J. Handsteiner, M. Malik, T. Scheidl, R. Ursin, Z. A., Krenn, M. Communication with spatially modulated light through turbulent air across vienna, *New Journal of Physics.* 16 (2014).
- [5] E. Abramochkin, V. V., Beam transformations and nontransformed beams, *Opt. Commun.* 83 (1991) 123–135.
- [6] M. Beijersbergen, R. Coerwinkel, M. Kristensen, J. Woerdman, Beijersbergen, M. Helical-wavefront laser beams produced with a spiral phaseplate, *Optics Communications.* 112(5-6) (1994) 321327.
- [7] N. Yoshida, H. Toyoda, Y. Igasaki, N. Mukohzaka, Y. Kobayashi, T. Hara, Yoshida, N. Nonpixelated electrically addressed spatial light modulator (slm) combining an optically addressed slm with a crt, *Holographic Optical Elements and Displays. Proc. SPIE* 2885 (1996) 132–136.
- [8] L. Marrucci, C. Manzo, D. Paparo, Pancharatnam-berry phase optical elements for wave front shaping in the visible domain: switchable helical mode generation, *Applied Physics Letters.* 88 (2006).
- [9] M. Padgett, J. Courtial, L. Allen, Light's orbital angular momentum, *Phys. Today.* 57 (2004) 35–40.
- [10] J. Leach, M. J. Padgett, S. M. Barnett, J. Franke-Arnold, S. Courtial, Leach, J. Measuring the orbital angular momentum of a single photon, *Phys. Rev. Lett.* 88 (2002) 257901–1–257901–4.
- [11] G. C. C. Berkhout, M. P. J. Lavery, J. Courtial, M. W. Beijersbergen, M. J. Padgett, Berkhout, G. C. C. Efficient sorting of orbital angular momentum states of light, *Phys. Rev. Lett.* 105 (2010) 153601–1–153601–4.
- [12] J. Leach, J. Courtial, K. Skeldon, S. M. Barnett, S. Franke-Arnold, M. J. Padgett, Leach, J. Interferometric methods to measure orbital and spin, or the total angular momentum of a single photon, *Phys. Rev. Lett.* 92 (2004) 013601–1–013601–4.
- [13] M. P. J. Lavery, G. C. C. Berkhout, J. Courtial, M. J. Padgett, Lavery, M. P. J. Measurement of the light orbital angular momentum spectrum using an optical geometric transformation, *J. Opt.* 13 (2011) 1–4.
- [14] M. P. J. Lavery, D. Roberston, M. Malik, B. Rodenburg, J. Courtial, R. W. Boyd, P. M. J., Lavery, M. P. J. The efficient sorting of light's orbital angular momentum for optical communications, *Electro-Optical Remote Sensing, Photonic Technologies, and Applications VI. Proc. SPIE* 8542 (2012) 85421R–1–85421R–7.
- [15] J. Arlt, K. Dholakia, L. Allen, M. J. Padgett, Arlt, J. The production of multiringed laguerre-gaussian modes by computer-generated holograms, *Mod. Opt.* 45 (1998) 1231–1237.
- [16] B. Schneier, *Applied cryptography. Protocols, Algorithms and Source code in C.*, 2th ed., Triumf, M., 2002.
- [17] A. Shamir, A polynomial-time algorithm for breaking the basic merkley-hellman cryptosystem, *IEEE Transactions on informations theory.* IT-30 (1984).

## Table of Contents

### Image Processing, Geoinformation Technology and Information Security

1. Acceleration of the reliable shortest path algorithm in a time-dependent stochastic transport network I. Abdulganiev, A. Agafonov.....	1-5
DOI: 10.18287/1613-0073-2017-1901-1-5	
2. Attacking the problem of continuous speech segmentation into basic units I.A. Andreev, A.I. Armer, N.A. Krasheninnikova, V.S. Moshkin.....	6-9
DOI: 10.18287/1613-0073-2017-1901-6-9	
3. Anomalies detection on spatially inhomogeneous polyzonal images N.A. Andriyanov, K.K. Vasiliev, V.E. Dementiev.....	10-15
DOI: 10.18287/1613-0073-2017-1901-10-15	
4. Voice command recognition for noisy environments by means of cross-correlation portraits A.I. Armer, E.Yu. Galitskaya, N.A. Krasheninnikova.....	16-22
DOI: 10.18287/1613-0073-2017-1901-16-22	
5. Development the algorithm of positioning industrial wares in-plant based on radio frequency identification for the products tracking systems A.V. Astafiev, A.A. Orlov, D.P. Popov.....	23-27
DOI: 10.18287/1613-0073-2017-1901-23-27	
6. The reliability of pattern-match searching for the fragment on image using set of pseudo-gradient procedures L.Sh. Biktimirov, A.G. Tashlinskii.....	28-31
DOI: 10.18287/1613-0073-2017-1901-28-31	
7. Development of informative neighborhood selection technology for modeling texture images E. Biryukova, R. Paringer, A. Kupriyanov.....	32-36
DOI: 10.18287/1613-0073-2017-1901-32-36	
8. Comparison of classification algorithms in the task of object recognition on radar images of the MSTAR base A.A. Borodinov, V.V. Myasnikov.....	37-41
DOI: 10.18287/1613-0073-2017-1901-37-41	
9. Spatio-temporal analysis through remote sensing and GIS in Moscow region, Russia Komal Choudhary, M.S. Boori, A. Kupriyanov.....	42-46
DOI: 10.18287/1613-0073-2017-1901-42-46	
10. Program-algorithm complex for image imposition in aircraft vision systems A.I. Efimov, A.I. Novikov.....	47-54
DOI: 10.18287/1613-0073-2017-1901-47-54	
11. The algorithm of the high-capacity information embedding into the digital images DCT domain using differential evolution O.O. Evsutin, A.O. Osipov.....	55-64
DOI: 10.18287/1613-0073-2017-1901-55-64	
12. A model for data hiding system description V. Fedoseev.....	65-71
DOI: 10.18287/1613-0073-2017-1901-65-71	
13. DPCM with an adaptive extrapolator for image compression M.V. Gashnikov.....	72-77
DOI: 10.18287/1613-0073-2017-1901-72-77	
14. Intelligent geographic information platform for transport process analysis O. Golovnin, A. Fedoseev, T. Mikheeva.....	78-85
DOI: 10.18287/1613-0073-2017-1901-78-85	
15. Feature Selection Methods for Remote Sensing Images Classification E. Goncharova, A. Gaidel.....	86-91
DOI: 10.18287/1613-0073-2017-1901-86-91	

16. Development and study of methods for estimating retinal vessel parameters using a modified local fan transform N.Yu. Ilyasova, A.S. Baisova, A.V. Kupriyanov.....	92-98
DOI: 10.18287/1613-0073-2017-1901-92-98	
17. Concerning the possibilities of successional changes revealing in anthropogenically transformed ecosystems on the base of remote sensing and ground-based survey data integration L.M. Kavelenova, N.V. Prokhorova, E.S. Korchikov, A.Yu. Denisova, D.A. Terentyeva.....	99-103
DOI: 10.18287/1613-0073-2017-1901-99-103	
18. Method of automated epileptiform seizures and sleep spindles detection in the wavelet spectrogram of rats' EEG I.A. Kershner, Yu.V. Obukhov, I.G. Komoltsev.....	104-109
DOI: 10.18287/1613-0073-2017-1901-104-109	
19. The application of OpenCL to accelerate the lossless image compression algorithm based on cascading fragmentation and pixels sequence ordering A. Khokhlachev, V. Smirnov, A. Korobeynikov.....	110-117
DOI: 10.18287/1613-0073-2017-1901-110-117	
20. Methods of IPD normalization to counteract IP timing covert channels K. Kogos, A. Sokolov.....	118-126
DOI: 10.18287/1613-0073-2017-1901-118-126	
21. Automatic adjustment of image processing pipeline D.A. Kolchaev, E.R. Muratov, M.B. Nikiforov.....	127-131
DOI: 10.18287/1613-0073-2017-1901-127-131	
22. Heuristic Malware Detection Mechanism Based on Executable Files Static Analysis A.V. Kozachok, M.V. Bochkov, E.V. Kochetkov.....	132-139
DOI: 10.18287/1613-0073-2017-1901-132-139	
23. Retinamorphic bichromatic Schrödinger metamedia V. Labunets, I. Artemov, V. Chasovskikh, E. Ostheimer.....	140-148
DOI: 10.18287/1613-0073-2017-1901-140-148	
24. Retinamorphic color Schrödinger metamedia V. Labunets, I. Artemov, V. Chasovskikh, E. Ostheimer.....	149-158
DOI: 10.18287/1613-0073-2017-1901-149-158	
25. Automated pathological growths parametrization based on computed tomography layer segmentation N.I. Limanova, S.G. Ataev.....	159-162
DOI: 10.18287/1613-0073-2017-1901-159-162	
26. Color discrimination thresholds and the Einstein's field equations L.D. Lozhkin, O.V. Osipov.....	163-168
DOI: 10.18287/1613-0073-2017-1901-163-168	
27. Method for identification of perlite-class steel microstructure parameters using metallographic images R.G. Magdeev, A.G. Tashlinskiy.....	169-175
DOI: 10.18287/1613-0073-2017-1901-169-175	
28. A fast one dimensional total variation regularization algorithm A. Makovetskii, S. Voronin, V. Kober.....	176-179
DOI: 10.18287/1613-0073-2017-1901-176-179	
29. Method of analysis of geomagnetic data based on wavelet transform and threshold functions O. Mandrikova, I. Solovev, S. Khomutov, K. Arora, L. Manjula, P. Chandrasekhar.....	180-186
DOI: 10.18287/1613-0073-2017-1901-180-186	
30. Large scale networks security strategy Ya. Mostovoy, V. Berdnikov.....	187-193
DOI: 10.18287/1613-0073-2017-1901-187-193	
31. Computationally efficient methods of clustering ensemble construction for satellite image segmentation I.A. Pestunov, S.A. Rylov, Yu.N. Sinyavskiy, V.B. Berikov.....	194-200
DOI: 10.18287/1613-0073-2017-1901-194-200	

32. Edge Detection in Remote Sensing Images Based on Fuzzy Image Representation E.V. Pugin, A.L. Zhiznyakov.....	201-206
DOI: 10.18287/1613-0073-2017-1901-201-206	
33. Compressing deep convolutional neural networks in visual emotion recognition A.G. Rassadin, A.V. Savchenko.....	207-213
DOI: 10.18287/1613-0073-2017-1901-207-213	
34. Real-time tracking of multiple objects with locally adaptive correlation filters A.N. Ruchay, V.I. Kober, I.E. Chernoskulov.....	214-218
DOI: 10.18287/1613-0073-2017-1901-214-218	
35. Methods for automated vectorization of point objects on cartographic images S. Rychazhkov, V. Fedoseev, R. Yuzkiv.....	219-225
DOI: 10.18287/1613-0073-2017-1901-219-225	
36. EEG Beta Wave Trains Are Not the Second Harmonic of Mu Wave Trains in Parkinson's Disease Patients O.S. Sushkova, A.A. Morozov, A.V. Gabova.....	226-234
DOI: 10.18287/1613-0073-2017-1901-226-234	
37. Effectiveness of correlation and information measures for synthesis of recurrent algorithms for estimating spatial deformations of video sequences A.G. Tashlinskiy, A.V. Zhukova.....	235-239
DOI: 10.18287/1613-0073-2017-1901-235-239	
38. Optimal bandwidth selection in geographically weighted factor analysis for education monitoring problems A. Timofeeva, K. Tesselkina.....	240-246
DOI: 10.18287/1613-0073-2017-1901-240-246	
39. A learning based feature point detector A. Verichev.....	247-252
DOI: 10.18287/1613-0073-2017-1901-247-252	
40. Combined method for calculating the disparity value on stereo images in problems of stereo-range metering A.N. Volkovich.....	253-258
DOI: 10.18287/1613-0073-2017-1901-253-258	
41. Increasing the energy efficiency of OFDM systems using differential signal conversion G.S. Voronkov, I.V. Kuznetsov, A.Kh. Sultanov.....	259-263
DOI: 10.18287/1613-0073-2017-1901-259-263	
42. Complex Matrix Model for Data and Knowledge Representation for Road-Climatic Zoning of the Territories and the Results of Its Approbation A. Yankovskaya, A. Sukhorukov.....	264-270
DOI: 10.18287/1613-0073-2017-1901-264-270	

# Preface

V.V. Myasnikov<sup>1</sup>, V.V. Sergeev<sup>1</sup>, V.A. Fedoseev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

This volume contains the papers presented at the two sessions, “Image Processing and Geoinformation Technology, and Information Security” within the 3rd International Conference on Information Technology and Nanotechnology - 2017 (ITNT-2017). The conference took place in Samara, Russia, April 25–27, 2017 (<http://ru.itnt-conf.org/itnt17ru/>). In addition to the two sessions mentioned above, the conference also included the following tracks: “Computer Optics and Nanophotonics”, “Mathematical Modeling”, “High-Performance Computing”, “Data Science”, “Information Technology and Education”.

The topics of the papers in this volume include signal and image processing and analysis, computer vision, pattern recognition, geographic information technology, remote sensing data processing and analysis, digital watermarking and steganography, digital forensics, malware detection, cryptography and cryptanalysis, data authentication, and reverse engineering.

This year we have received 118 submissions addressed to the sessions “Image Processing and Geoinformation Technology”, and “Information Security”. Each submission was carefully reviewed by the program committee members and the reviewers, which are well-known experts in computer vision and information security from Russia, Germany, Italy, Poland, The Netherlands, Brazil, Azerbaijan, and Switzerland. Based on the reviews, 72 papers were accepted for presentation at the conference and publication. Some of the accepted papers were published in Elsevier Procedia Engineering Series, in the volume covering all topics of the conference. The current volume contains 42 papers not included in the Procedia Engineering volume.

We are grateful to the authors, and to the program committee members for their time and efforts and we look forward to meeting you again at future events.

## Guest Editors

- Vladislav Myasnikov, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics", Russian Academy of Sciences, Samara, Russia
- Vladislav Sergeev, Samara National Research University, Samara, Russia
- Victor Fedoseev, Samara National Research University, Samara, Russia
- Yulia Vybornova, Samara National Research University, Samara, Russia
- Denis Kudryashov, Samara National Research University, Samara, Russia

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov, Samara National Research University, Samara, Russia

## Vice-chairs

- Vladimir Bogatyrev, Samara National Research University, Samara, Russia
- Nikolay Kazanskiy, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Samara, Russia
- Eduard Kolomiets, Samara National Research University, Samara, Russia
- Alexander Kupriyanov, Samara National Research University, Samara, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Samara, Russia

# Acceleration of the reliable shortest path algorithm in a time-dependent stochastic transport network

I. Abdulganiev<sup>1</sup>, A. Agafonov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

In this paper, we propose a modification of the reliable routing algorithm in a stochastic time-dependent network. We consider a stochastic on-time arrival problem. Reliability means maximization of the probability of arrival at a destination within a given period of time. Modification of the shortest-path algorithm is aimed to decrease the computation time of the algorithm. The base idea of the proposed modification is to select a certain subset of nodes and links of the graph which can be used for calculating the shortest path. We propose two methods for selecting subset of nodes: based on a bounding box and based on the k-shortest path algorithm. Experimental studies of the base and modified algorithms are carried out on the transport network of Samara, Russia.

*Keywords:* reliable shortest path; adaptive route; time-dependent network; k shortest path algorithm

---

## 1. Introduction

Road congestion is a serious problem of modern society, which has several significant consequences. For traffic participants congestion reduces quality of life, consuming their free time. For organizations, congestion reduces employee productivity and increases freight transportation costs. For the society as a whole, traffic jams lead to the disruption of emergency services, and negatively affects the quality of the environment, causing a large amount of exhaust emissions. Traffic jams threaten traffic safety, raising the level of stress and fatigue among drivers. Thus, it becomes increasingly important to solve the optimal routing navigation problems.

In recent decades, a large number of papers have been devoted to the shortest path problem in transport networks. However, most of them focused their attention on finding the least expected travel time. Several types of models are distinguished depending on the weight type of the road segment. In classical models [1,2,3], the travel time of a road segment is considered to be constant or time-dependent. At the same time, the real situation shows that the travel time is continuously changing and depends on many factors such as time of day, weather conditions, traffic situation, etc. In models with stochastic travel time [4,5,6], the time is represented by a random variable with a time-dependent distribution function. In either case, the optimality condition for routing can be determined in different ways depending on the used objective function. The following types of objective functions can be used:

- 1) minimization of the least expected travel time [4 - 7];
- 2) maximization of the probability of arriving at the destination at a predetermined time interval [8, 9];
- 3) maximization of the probability that travel time is less than a given threshold [10, 11];
- 4) minimization of the worst possible travel time [12];
- 5) minimization of the travel time to guarantee a given likelihood of arriving on-time in a stochastic network [13].

These types of objective functions can be divided into two groups: the least expected time (LET) (1) and the reliable shortest path (RSP) problem (2-5). The LET problem has been well studied, there are many effective algorithms for different types of the problem [6,7]. Nevertheless, in a number of practical tasks, the path with the least expected time may not be suitable, since it does not take into account the dispersion of the travel time and does not give any guarantees of reliability. In many cases, road users decided to increase travel time and to choose a more reliable route [8]. The problem of finding a reliable shortest path was investigated in [10, 11]. In [14, 15], a shortest path algorithm was proposed, taking into account current and predictive information about the transport flows in the network, which is a modification of the algorithm from [8]. However, the algorithms described in the works are computationally complex and cannot be used to determine the shortest path in large-scale networks in real time.

In this paper, we investigate the method for reliable routing in a time-dependent stochastic transport network. We consider the following optimality criterion: maximizing the probability of arrival at a destination within a predetermined time interval. The aim of this work is to decrease the computation time of the existing algorithm. Acceleration of the algorithm is achieved by choosing a subset of the graph nodes used to find the shortest path. The algorithm proposed by the group of authors in [8] is used as a basic algorithm.

The paper is organized as follows. The second section introduces the basic notation, the problem statement, and describes the base algorithm. In the third section, we propose modifications of the reliable routing algorithm. In the fourth section, experimental studies of the proposed algorithms are presented. Finally, we provide our conclusions.

## 2. Stochastic on-time arrival problem

The transport network is defined as an oriented, time-dependent stochastic graph:

$$G = (N, A, P)$$



where  $N$  is a nonempty set of nodes that correspond to road intersections,  $|N|$  is the number of nodes,  $A$  is a set of links that correspond to road segments,  $|A|$  is the number of links,  $P$  is the probabilistic description of the links travel time. We assume that the graph has a spatial reference, i.e. each node of the graph  $i \in N$  has coordinates  $(x, y)_i$  that are determined by the physical location of the corresponding intersection in the real road network.

Let  $T_{ij}(\tau)$  be the travel time of the link  $(i, j) \in A$ . Travel time  $T_{ij}(\tau)$  is represented as a random variable with a probability density function  $p_{ij}^\tau(t)$  that depends on the time at which the vehicle enters this link.

Let  $r \in N$  be the origin node,  $s \in N$  be the destination node,  $T$  be the maximum amount of time allowed to reach the destination, i.e. the time budget.

The optimal routing policy is defined as the policy of maximizing the probability of arrival at the destination  $s$  in a time less than  $T$ . This problem is abbreviated as SOTA (Stochastic On Time Arrival) [8, 14, 16].

In papers [8, 19] for the definition of the optimal routing policy, an additional notation is introduced. Let  $u_i^\tau(t)$  be the probability of reaching the destination node  $s$  from the node  $i$  at the time  $\tau$  with a time budget  $t$ .  $u_i^\tau(t)$  is called reliability of the path.

**Definition 1** [8]. The optimal routing policy for the SOTA problem can be formulated as follows

$$u_i^\tau(t) = \max_j \int_0^t p_{ij}^\tau(\omega) u_j^{\tau+\omega}(t-\omega) d\omega$$

$$\forall_i \in N, i \neq s, (i, j) \in A, t \in [0, T], \tau \geq 0; \quad (1)$$

$$u_s^\tau(t) = 1 \quad t \in [0, T], \tau \geq 0.$$

The discrete algorithm for solving the problem (1) was described in [8]. We will consider it as the base algorithm. The algorithm can be formulated as follows:

**Step 0. Initialization.**

$$k = 0,$$

$$u_i^k(x) = 0, \forall i \in N, i \neq s, x \in N, x \in [0, T / \Delta t],$$

$$u_s^k(t) = 1, x \in N, x \in [0, T / \Delta t].$$

**Step 1. Update.**

FOR  $k = 1, 2, \dots, L = \lceil T / \Delta t \rceil$

$$\tau^k = k\delta,$$

$$u_s^k(t) = 1, x \in N, x \in [0, T / \Delta t],$$

$$u_i^k(x) = u_i^{k-1}(x), \forall i \in N, i \neq s, (i, j) \in A, x \in N, x \in [0, (\tau^k - \delta) / \Delta t],$$

$$u_i^k(x) = \max_j \sum_{h=0}^x p_{ij}(\tau^k) u_j^{k-1}(x-h), \forall i \in N, i \neq s, (i, j) \in A, x \in N, x \in [((\tau^k - \delta) / \Delta t + 1), (\tau^k / \Delta t)]$$

END FOR

where  $\Delta t$  is a sampling interval,  $\delta$  is the minimum travel time across the transport network.

The decision rule for selecting the next node  $j$  in the transport network graph for a given time budget  $t$  and calculated probabilities  $u_i(x)$  is as follows:

$$j = \arg \max_{i \in N} u_i(t) \quad (2)$$

In the described algorithm, all graph nodes are used to construct the shortest route that makes this algorithm computationally complex and not applicable for finding the shortest path in large-scale networks in real time. In the next section we propose modifications of the algorithm, consisting in selecting a certain subset of the graph nodes that will be used to find the shortest route.

### 3. Modified algorithm

The main idea of the modification is to achieve the acceleration of the base algorithm by reducing the number of nodes that are considered as candidates in the shortest path. We propose two methods for selecting a subset of nodes and links of the graph used to construct a reliable shortest path: on the basis of the bounding box and on the basis of the k-shortest path algorithm.

#### 3.1. Subset based on a bounding box

The obvious way to reduce the number of nodes used to find the shortest route is to select only those nodes whose coordinates is inside the bounding box between the origin and destination nodes

Let  $(x, y)_{r \in N}$  be the coordinates of the origin node and  $(x, y)_{s \in N}$  be the coordinates of the destination node.

The bounding box coordinates can be written as follows:

$$\{x_{\min} = \min(x_r, x_s) - \Delta, y_{\min} = \min(y_r, y_s) - \Delta, x_{\max} = \max(x_r, x_s) + \Delta, y_{\max} = \max(y_r, y_s) + \Delta\},$$

where  $\Delta$  is the buffer distance chosen experimentally.

Then the subsets of nodes and links of the graph can be defined as follows:

$$\begin{aligned} \bar{N} &= \{i \in N : x_{\min} \leq x_i \leq x_{\max} \wedge y_{\min} \leq y_i \leq y_{\max}\}, \\ \bar{A} &= \{(i, j) \in A : i \in \bar{N} \wedge j \in \bar{N}\}. \end{aligned}$$

This method of selecting subsets is computationally simple, but the resulting subsets may be redundant or, conversely, insufficient, depending on the network structure and the location of the origin and destination nodes..

### 3.2. Subset based on the k-shortest path algorithm

The k-shortest path algorithm is an extension algorithm of the shortest path routing algorithm in a given network [17]. Depending on the problem statement, the shortest paths may or may not contain the same nodes and links. We suggest that the nodes and links that are included in the shortest path are marked as passed and cannot be used to construct the next shortest paths. To find different shortest paths, various algorithms can be used; in this paper we use the Dijkstra algorithm [1].

For a formal description of the method, we introduce additional notations.

Let  $d_{rs}^k = \{v_1^k, v_2^k, \dots, v_M^k\}$  be the  $k^{\text{th}}$  shortest path between nodes  $r$  and  $s$ , where  $v_m^k \in N, m=0, M^k-1$  is the graph node included in the  $k^{\text{th}}$  shortest path:  
 $M^k$  is the number of nodes in the shortest path;  
 $k = \bar{1}, \bar{K}$  is the shortest path index;  
 $K$  is the number of shortest path chosen experimentally.

Then the subsets of nodes and links of the graph can be defined as follows:

$$\begin{aligned} \bar{N} &= \{i \in \{d_{rs}^k\}_{k=\bar{1}, \bar{K}}\}, \\ \bar{A} &= \{(i, j) \in A : i \in \bar{N} \wedge j \in \bar{N}\}. \end{aligned}$$

The method of selecting a subset of nodes based on the k-shortest path algorithm has a greater computational complexity than the method based on the bounding box, but it does not depend on the network structure and allows to adjust the size of the subset by selecting the appropriate parameter  $K$ .

### 3.3. Modified reliable routing algorithm

The modified reliable routing algorithm uses subsets of nodes  $\bar{N}$  and links  $\bar{A}$  of the graph. In the form of a pseudo code, the algorithm can be written as follows:

#### Step 0. Initialization.

$$k = 0,$$

$$u_i^k(x) = 0, \forall i \in N, i \neq s, x \in \bar{N}, x \in [0, T / \Delta t],$$

$$u_s^k(t) = 1, x \in \bar{N}, x \in [0, T / \Delta t].$$

#### Step 1. Update.

FOR  $k = 1, 2, \dots, L = \lceil T / \Delta t \rceil$

$$\tau^k = k\delta,$$

$$u_s^k(t) = 1, x \in \bar{N}, x \in [0, T / \Delta t],$$

$$u_i^k(x) = u_i^{k-1}(x), \forall i \in N, i \neq s, (i, j) \in \bar{A}, x \in \bar{N}, x \in [0, (\tau^k - \delta) / \Delta t],$$

$$u_i^k(x) = \max_j \sum_{h=0}^x p_{ij}(h) u_j^{k-1}(x-h), \forall i \in N, i \neq s, (i, j) \in \bar{A}, x \in \bar{N}, x \in [((\tau^k - \delta) / \Delta t + 1), (\tau^k / \Delta t)]$$

END FOR

The subsets  $\bar{N}$  and  $\bar{A}$  are selected by one of the methods described above.

#### 4. Results and Discussion

The purposes of the conducted experiments were to compare the base and modified algorithms according to the criteria of the computation time of the algorithm and the reliability  $u_i^r(t)$  of the calculated route.

The experiments were carried out on the transport network of Samara, Russia. The transport network consists of 9721 nodes and 26088 links. As the weight of the road link, we used travel time data averaged over a ten-minute interval. As the probability density function on the link  $(i, j)$  we used the lognormal distribution.

To conduct experiments, 10 initial and 10 final vertices of the transport network have been chosen, located at a considerable distance from each other. For each pair of vertices, the shortest paths have been found using the base and modified algorithms. To study the algorithm based on the bounding box, the following values of the buffer distance were chosen:  $\Delta = \{200; 300; 400; 500; 600\}$ . The study of the algorithm based on k-shortest paths was conducted depending on the number of shortest paths  $K = \{3; 4; 5; 6; 7\}$ .

First of all, we compare the algorithms by the reliability criterion (the probability of reaching a destination node with a given time budget). The results of the comparison are presented in Table 1.

Table 1. Algorithm's comparison by the reliability.

Algorithms	Base algorithm	Bounding box					<i>k</i> -shortest paths				
		200	300	400	500	600	3	4	5	6	7
Reliability	0.99836	0.99568	0.99567	0.99624	0.996328	0.99642	0.99494	0.99512	0.99512	0.99523	0.995239

All algorithms have shown similar results for the chosen criterion for specified time budgets, modified algorithms slightly inferior to the base one.

The next stage of the experimental analysis was the comparison of algorithms by the criterion of the computation time. The computational time of the base algorithm was about 10 minutes in average, depending on the used time budget  $T$ . Figure 1 shows a histogram of the modified algorithms acceleration in comparison with the base algorithm for different values of the buffer distance and the number of shortest paths.

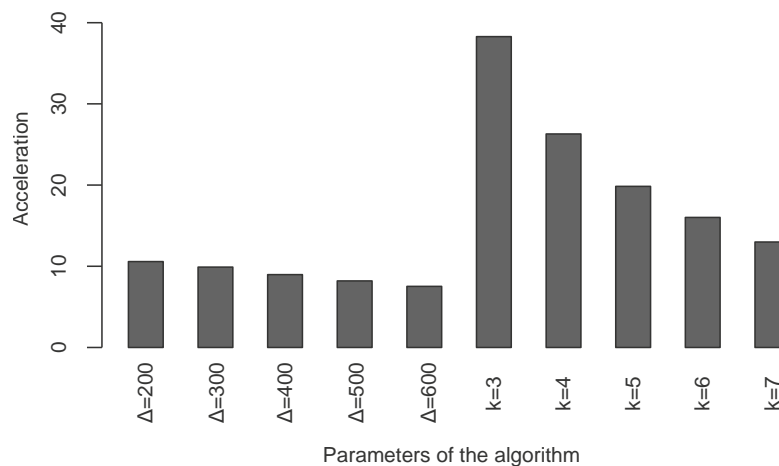


Fig 1. Acceleration of the modified algorithm.

The best results on this criterion have been shown by the algorithm based on k-shortest paths. The computation time in comparison with the base algorithm decreased by 10-35 times depending on the number of shortest paths  $K$ . Modification of the base algorithm based on the bounding box allows to decrease the computation time by 8-10 times. Note that the value of the buffer distance does not have a strong effect on the speed of the algorithm, but too much value can significantly increase the computation time.

The results show, that the proposed modifications allow to significantly decrease the computation time of the base algorithm (in 8-35 times) with a slightly decrease in the reliability of the shortest path. The modified algorithm can be used to find the reliable shortest route in large-scale networks in real time.

#### 5. Conclusion

In this paper we have proposed modifications of the reliable routing algorithm in a time-dependent stochastic transport network. Reliability means maximizing the probability of arrival at a destination within a time budget. Two modifications are proposed for the purpose of decreasing the computation time of the algorithm: based on the bounding box and based on the k-shortest path algorithm. The proposed modified algorithms are compared with the base algorithm on the Samara transport network.

The results of experimental studies have shown that the proposed modifications allow to significantly decrease the computation time (by 8-35 times) with practically identical reliability of the routes. The modified algorithms allow to find the shortest route in large-scale networks in real time.

Further research includes:

- studies related to the route choice depending on the traffic flows;
- studies related to the choice of a travel time budget.

## Acknowledgements

This work was supported by the Russian Foundation for Basic Research (RFBR) grant 16-37-00055-mol-a.

## References

- [1] Dijkstra EW. A Note on Two Problems on Connexion with Graphs. *Numerische Mathematik* 1959; 1: 269–271.
- [2] Bellman RE. On a routing problem. *Quarterly of Applied Mathematics* 1958; 16: 87–90.
- [3] Dreyfus SE. An appraisal of some shortest-path algorithm. *Operations Research* 1969; 17: 395–412.
- [4] Gao S, Chabini I. Optimal routing policy problems in stochastic timedependent networks. *Transportation Research Part B* 2006; 40: 93–122.
- [5] Gao S, Huang H. Real-time traveler information for optimal adaptive routing in stochastic time-dependent networks. *Transportation Research Part C* 2012; 21: 196–213.
- [6] Hall RW. The fastest path through a network with random time-dependent travel times. *Transportation Science* 1986; 20(3): 182–188.
- [7] Fu L, Rilett LR. Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B* 1998; 32(7): 499–516.
- [8] Samaranyake S, Blandin S, Bayen A. A tractable class of algorithms for reliable routing in stochastic networks. *Transportation Research Part C* 2012; 20: 199–217.
- [9] Fan Y, Nie Y. Optimal routing for maximizing the travel time reliability. *Networks and Spatial Economics* 2006; 6(3-4): 333–344.
- [10] Frank H. Shortest paths in probabilistic graphs. *Operations Research*. 1969; 17(4): 583–589.
- [11] Mirchandani PB. Shortest distance and reliability of probabilistic networks. *Computers and Operations Research* 1976; 3(4): 347–355.
- [12] Montemanni R, Gambardella L. An exact algorithm for the robust shortest path problem with interval data. *Computers and Operations Research* 2004; 31(10): 1667–1680.
- [13] Nie Y, Wu X. Shortest path problem considering on-time arrival probability. *Transportation Research Part B* 2009; 43(6): 597–613.
- [14] Agafonov A, Myasnikov V. Reliable routing in stochastic timedependent network with the use of actual and forecast information of the traffic flows. *IEEE Intelligent Vehicles Symposium, Proceedings* 2016; 1168–1172.
- [15] Agafonov A. Method for the reliable shortest path search in time-dependent stochastic networks and its application to GIS-based traffic control. *Computer Optics* 2016; 40 (2): 275–283. DOI: 10.18287/2412-6179-2016-40-2-275-283.
- [16] Nie Y, Fan Y. Arriving-on-time problem: Discrete algorithm that ensures convergence. *Transportation Research Record* 2006; 1964: 193–200.
- [17] Yen JY. Finding the K shortest loopless paths in a network. *Management Science* 1971; 17(11): 712–716.

# Attacking the problem of continuous speech segmentation into basic units

I.A. Andreev<sup>1</sup>, A.I. Armer<sup>1</sup>, N.A. Krashennikova<sup>2</sup>, V.S. Moshkin<sup>1</sup>

<sup>1</sup>*Ulyanovsk State Technical University, Severny Venetz St., 32, 432027, Ulyanovsk, Russia*

<sup>2</sup>*Ulyanovsk State University, Lev Tolstoy St., 42, 432017, Ulyanovsk, Russia*

---

## Abstract

The paper considers the algorithm of continuous speech segmentation into basic units, namely phonemes, certain combination of phonemes and pauses. The algorithm is based on speech signal transformation into a two-dimensional image, i.e. an autocorrelation portrait. To determine the boundaries of speech units the portraits of the analyzed signal are aligned with the model portraits of each speech unit. The authors apply the dynamic programming to find out the optimal distance between portraits.

*Keywords:* speech signal; segmentation; autocorrelation portrait; speech units; discrete dynamic programming

---

## 1. Introduction

At present, the algorithms for continuous speech segmentation into verbal units - phonemes, their combinations and pauses - are quite in demand. For example, this problem arises, while creating systems for research, processing, modeling and automatic speech recognition. To use such systems under different acoustic conditions, they should be subject to strict requirements for acoustic noise impedance and speech signal distortion. The article presents a method for determining the boundaries of speech pauses and speech units, which correspond to SAMPA + for the Russian language [1], [2]. The algorithms of speech signal transformation and processing used in the suggested method correspond to the strict requirements for acoustic noise impedance and speech signal distortion.

## 2. The subject of investigation

The problem of speech signal segmentation into its basic units is extremely complicated and challenging, and at present there is no simple solution for the general case. It is noted [3] that there are certain cases, for which exact segmentation is problematic. Different methods [3],[4],[5] are used for continuous speech signal segmentation. It is possible to distinguish the methods based on spectral analysis, trajectories of signal energy, energy logarithm, number of transitions through zero, and statistical parameters of speech units. The abovementioned methods give good results under favorable acoustic conditions, but the results deteriorate due to the presence of noise. Moreover, the time length of a speech signal varies from one pronunciation to another, which also makes its segmentation into basic units difficult. The authors suggest using the autocorrelation transformation [6],[7] of the speech signal into a two-dimensional image as well as the certain ways of image alignment in order to improve noise stability when determining the speech unit boundaries. The autocorrelation transformation has a number of characteristics, which make it somewhat noise-resistant [8]. Thus, the proposed method of speech signal segmentation is to be assumed to be less dependent on the current acoustic conditions, in which it was pronounced. Using discrete dynamic programming [9], when aligning two-dimensional speech signal images makes it possible to increase the stability of the method under consideration to the changes in the time length of speech units.

## 3. Algorithm for determining speech unit boundaries

### 3.1. General algorithm

The algorithm for determining speech unit boundaries is as follows: a speech signal containing a fragment of continuous speech analyzed for speech unit boundaries is represented in the form of digital readouts. The models of each speech unit are also represented as digital readouts. For benchmarking each example of the speech unit corresponding to SAMPA + is pronounced by the speaker, then the boundaries are defined by ear, and the speech unit becomes a model. By means of the autocorrelation transformation digital readouts of the analyzed continuous speech segment and the readouts of every model speech unit are transformed into particular two-dimensional images, which are called autocorrelation portraits (ACPs). For further alignment portraits of the analyzed speech segment and every model speech unit have the same line length.

Next, the portrait of the analyzed speech segment is aligned with all portraits of model speech units to determine the speech unit boundaries. For this purpose, the distance [10],[11] is calculated in the sliding window. The size of the window is equal to the number of lines in a corresponding speech unit portrait. During the calculation, the distance between the windows is optimized using the discrete dynamic programming. For each speech unit, a distance array along the portrait of the analyzed speech segment is determined. The distances corresponding to the same fragments of the analyzed speech segment portrait are compared with each other. As a result, speech unit portraits, which have the smallest distances, form the desired boundaries. If the smallest distance is obtained from the portraits of identical speech units, which follow one another, they are combined into the boundaries of one speech unit.

### 1.2. Autocorrelation portraits of speech signals

Since autocorrelation links are rather informative, i.e. they reflect speech signal features ACPs are unique for each speech unit. This provides good results in obtaining the speech unit boundaries for continuous speech. In [12] ACPs are modeled in the following way. Let  $s(i)$  be the  $i$ -th readout of a digital speech signal;  $s(i+k)$  is a readout spaced  $k$  readouts apart  $s(i)$ . Dependency factor of these readouts is expressed by a sample correlation coefficient:

$$R_s(k) = R[s(i), s(i+k)] = \frac{\text{cov}[s(i), s(i+k)]}{\sqrt{\frac{1}{N} \sum_{i=1}^N s^2(i) - m_{s(i)}^2} \sqrt{\frac{1}{N} \sum_{i=1}^N s^2(i+k) - m_{s(i+k)}^2}},$$

$$\text{cov}[s(i), s(i+k)] = \frac{1}{N} \sum_{i=1}^N s(i)s(i+k) - \left[ \frac{1}{N} \sum_{i=1}^N s(i) \right] \left[ \frac{1}{N} \sum_{i=1}^N s(i+k) \right], \quad (1)$$

where  $N$  is a number of readouts in the interval, in which the dependency is sought;  $\text{cov}[s(i), s(i+k)]$  is the sample covariance  $s(i)$  and  $s(i+k)$  when  $i = 1..N$ ;  $m_{s(i)}$  is a sample mean  $s(i)$  when  $i = 1..N$ ;  $m_{s(i+k)}$  is a sample mean  $s(i+k)$  when  $i = 1..N$ . Function determined by the sample correlation coefficient using (1) is an autocorrelation function (ACF) of a signal. While calculating ACF we perform the transformation of speech signal (SS) readouts  $s(i)$   $i = 1..M$  ( $M$  is the number of readouts in a speech signal) into a two-dimensional image. For this purpose,  $s(i)$  is divided into intervals including  $N < M$  readouts, then, in each  $j$ -th ( $j = 1, N, 2N, \dots, M - 2N$ ) interval the local signal maximum  $i_m^j = \max|s|$  is sought. Let us assume that  $M$  is divisible by  $N$  evenly, otherwise the remaining final SS readouts are omitted. Then, using equation (1) we calculate the elements of the corresponding ACP line beginning with  $i_m^j$  ( $j = 1, N, 2N, \dots, M - 2N$ ) and generate ACP lines:

$$R[s(i_m^j), s(i_m^j + k)] \quad \begin{matrix} k=1..N \\ j=1, N, 2N, \dots, M-2N \end{matrix}, \quad (2)$$

$$X(j, k) = R.$$

The two-dimensional image  $X(j, k)$  obtained from (2), where  $j$  is the line number, and  $k$  is the column number, is the ACP of a speech signal  $s(i)$  dimensioned  $N \times \left(\frac{M}{N} - 2\right)$ , generated using SS local maxima. Note, that ACPs generated using local maxima are unique for each speech unit, and due to their link with SS local maxima they are less subject to geometrical distortions associated with speech variability. Figure 1 represents ACPs of speech units [“a”, [o], [n`:], [f] (SAMPA+).

### 1.3. Alignment of autocorrelation portraits using discrete dynamic programming

Due to high degree of speech signal variability, autocorrelation portraits of one speech unit pronounced at different times differ from each other. Figure 2 shows ACPs of a speech unit “unstressed [a]”, one of them (a) was obtained from the pronunciation of the word «Вера» / “Vera”, and another (b) from the word «сопутствующие» / “soputstvujushhie”. It is obvious, that the portraits differ in the number of lines. Nevertheless, some lines of portrait a) can correspond to one line of portrait b).

The distance between the corresponding ACP lines is determined for the  $i$ -th line of portrait  $X$  and the  $j$ -th line of portrait  $Y$  using the following formula:

$$\rho_{i,j} = \sum_{k=1}^N (X(i, k) - Y(j, k))^2. \quad (3)$$

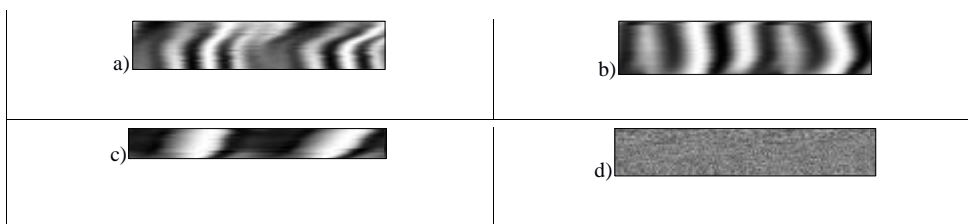


Fig. 1. ACPs of speech units a) [“a], b) [o], c) [n`:], d) [f].

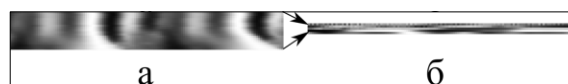


Fig. 2. ACPs of a speech unit “unstressed [a]”: a) model, b) as a part of the word «сопутствующие» / “soputstvujushhie”.

To determine the measure of ACP concordance the discrete dynamic programming [9] is applied. It allows to minimize the functional  $\rho = \min \sqrt{\sum \rho_{i,j}}$ , which characterizes ACPs identity. Set  $\Omega$  predetermines the permitted correspondences of the portrait lines, which are obtained on the basis of the following rules. 1. The number of lines in ACPs can differ. 2. Any line of one particular ACP cannot correspond to the line of another one spaced from the previous corresponding line more than  $c$  lines apart. 3. The order of line correspondence is preserved, i.e. if the  $i$ -th line of one ACP corresponds to the  $j$ -th line of the other one, then the  $(i+1)$ -th line cannot correspond to  $j-l$ ,  $l = 1, 2, \dots$ . 4. The total distance between ACP pronunciations of the same

speech units formed from the distances between the corresponding lines according to the second metrics rule should be minimal according to rules 1)-3).

To determine the measure of speech signal ACP correspondence (in a two-dimensional sliding window) to speech unit ACP the following algorithm is obtained. Matrix  $D$  containing  $m \times m$  elements is created, where  $m$  is the number of CP lines in a sliding window  $X$ ; the number of speech unit  $Y$  ACP lines is the same. For example, let  $c = 3$ . At first, the distances between  $Y(1)$  and  $X(1), X(2), X(3)$  are found, then these distances are stored in  $D$

$$D_{1,i} = \rho(Y(1), X(i)), i = 1..3. \quad (4)$$

Then, distances between  $Y(2)$  and  $X(1), X(2), X(3), X(4), X(5)$  are found. The position of the line  $Y(1)$  is taken into account, i.e. if  $Y(1)$  corresponds to  $X(2)$ , then  $Y(2)$  can be compared only with  $X(2), X(3), X(4)$ . Each time it is necessary to remember portrait  $X$  line number, and fill in the matrix  $D_{2,i} = D_{1,i} + \rho(Y(2), X(j)), j = i..i + 2$ . Besides, each element from  $D$  due to intersection of possible line positions can be filled in several times. In such a case, the minimum value (Figure 3) is preserved:

$$D_{k,i} = \min[D_{k,j}, D_{k-1,j} + \rho(Y(k), X(j))], j = i..i + 2. \quad (5)$$

During the next stages, all the remaining matrix  $D$  elements are found using formula (5), at each stage  $i$  changes from 1 to  $I + 2$ , where  $I$  is the maximum value of  $i$  at the previous stage. For the first stage  $I = 1$ . The algorithm is stopped when matrix  $D$  is completely filled. The minimal element from the  $m$ -th line and the  $m$ -th column of the matrix corresponds to the minimal distance between  $X$  and  $Y$ .

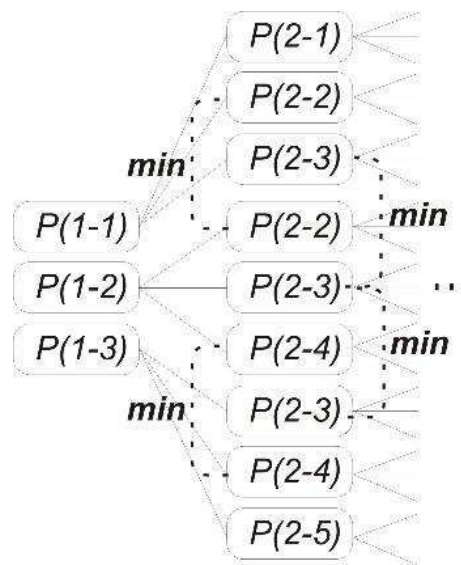


Fig. 3. Distribution of compared ACP lines.  $P(i-j)$  is the distance between the  $i$ -th line of one ACP and the  $j$ -th line of another one. Mark **min** shows that from all possible identical comparisons at different stages of programming the comparison with the minimal **distance** is chosen.

#### 4. Experiments

The suggested algorithm for determining speech unit boundaries in continuous speech was tested experimentally. Figure 4 shows the speech unit boundaries in the utterance containing the pronunciation of the word «основного» / “osnovnogo”. For example, the interval of speech unit [a] pronunciation, which starts the word «основного» / “osnovnogo”, was correctly defined within the range from 800 to 4800 speech signal digital readouts, speech unit [s] – in the range from 2400 to 5600 readouts, speech unit [n] – in the range from 5600 to 9200 readouts, speech unit [a] – in the range from 9600 to 11200 readouts, speech unit [v] – in the range from 11200 to 16000 readouts, speech unit [n] – in the range from 16000 to 17200 readouts, speech unit [“o] in the range from 17200 to 26400 readouts, speech unit [v] – in the range from 26400 to 28000 readouts and the last of the analyzed speech signal unit [a] – in the range from 28000 and up to the end of the signal.

Comparison with expert borders was not made. However, visual comparison of the determined boundaries with the real ones shows their closeness. Experiments show the practical applicability of the algorithm for determining the speech unit boundaries in continuous speech.



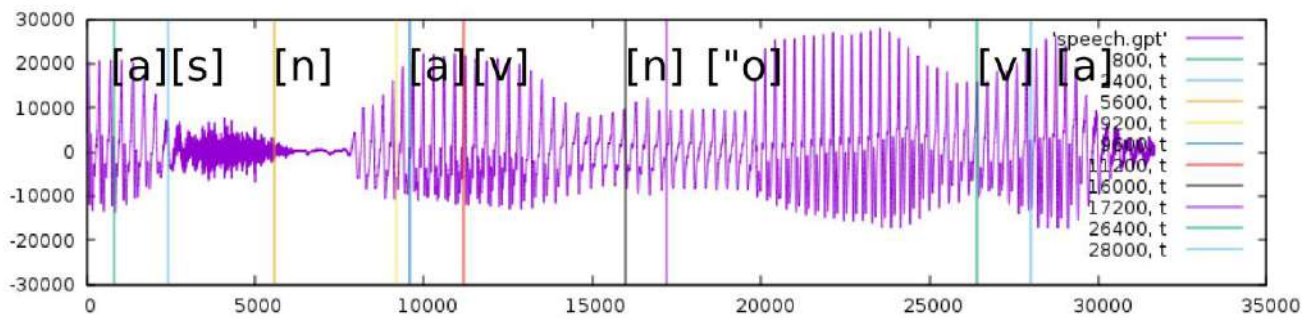


Fig. 4. Speech unit boundaries in continuous speech containing the pronunciation of the word «основного» / “osnovnogo”.

## 5. Conclusion

The determined speech unit boundaries are to be used for a more detailed analysis of the speech signal in order to identify the speech units. In order to solve this problem the authors also want to transform speech signals into ACPs. However, the parameters of transformation into ACPs and the method of portrait alignment will be different.

## Acknowledgements

The work was supported by grants 16-48-732046 and 16-48-730305 from the Russian Foundation for Basic Research.

## References

- [1] Galounov VI, Heuvel H, Kochanina JL, Ostroukhov AV, Tropf H, Vorontsova AV. Speech Database for the Russian Language. Proceedings of international workshop SPEECOM 1998.
- [2] Michael P, Rasanen O, Thiollière R, Dupoux E. Improving Phoneme Segmentation With Recurrent Neural Networks. *Computation and Language*, 2016, preprint:1608.00508.
- [3] Rabiner LR, Schafer RV. Digital processing of speech signals. Edited by M.V. Nazarov and Yu.N. Prokhorov. Moscow: Radio i svyaz', 1981; 496 p. (in Russian)
- [4] Goldenthal W. Statistical Trajectory Models for Phonetic Recognition. PhD thesis. M.I.T., 1994; 170 p.
- [5] Ostendorf M, Roukos SA. A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Transaction on Acoustics, Speech, and Signal Processing* 1989; 37(12): 1857–1869.
- [6] Therrien C, Tummala M. Probability and Random Processes for Electrical and Computer Engineers. CRC Press, 2012; 287 p.
- [7] Amirgaliyev Y, Mussabayev T. The speech signal segmentation algorithm using pitch synchronous analysis. *Open Comput. Sci.* 2017; 7: 1–8.
- [8] Krasheninnikov VR, Armer AI, Krasheninnikova NA, Kuznetsov VV, Khvostov AV. Some problems connected with speech command recognition on the background of intense noise. *Infokommunikatsionnye tekhnologii. Samara* 2008; 1: 72–75. (in Russian)
- [9] Bellman R. Dynamic programming. Moscow: IL, 1960; 400 p. (in Russian)
- [10] Krasheninnikov VR, Armer AI, Kuznetsov VV. Autocorrelated Images and Search for Distance between them in Speech Commands Recognition. *Pattern Recognition and Image Analysis.* 2008; 18(4): 663–666.
- [11] Greibus M. Rule Based Speech Signal Segmentation. *Journal of telecommunications and information technology* 2010; 4: 37–43.
- [12] Krasheninnikov VR, Armer AI, Krasheninnikova NA, Khvostov AV. Speech command recognition on the background of intense noise using autocorrelated portraits. *Naukojomiye tekhnologii* 2007; 8(9): 65–76. (in Russian)



# Anomalies detection on spatially inhomogeneous polyzonal images

N.A. Andriyanov<sup>1</sup>, K.K. Vasiliev<sup>1</sup>, V.E. Dementiev<sup>1</sup>

<sup>1</sup>*Ulyanovsk State Technical University, Severniy Venets street, 32, 432027, Ulyanovsk, Russia*

---

## Abstract

The text deals with the problem of detecting anomalies on a background of multi-dimensional images. We synthesized a detection algorithm based on the use of doubly stochastic models of random fields and which requires pre-filtering the image. We propose to use the modified Kalman filter. We also investigated an efficiency of extended signals detection on real images. It is shown that the resulting algorithm has a higher efficiency than the known algorithms which based on the traditional autoregressive image description. The gain is explained by more adequate description of the real inhomogeneous material using doubly stochastic models.

*Keywords:* doubly stochastic models; random fields; anomalies detection; image filtering; Kalman filter

---

## 1. Introduction

The tasks of detecting and estimating the parameters of anomalies in images are of interest for a number of applications. Among them, we can distinguish radio and sonar systems with spatial antenna arrays, aerospace systems for global Earth monitoring, systems of technical vision, etc. For these systems [1-3] the description of signals and interference is realized by means of random functions of several variables, i.e., by multidimensional random fields (RF). Typical examples of the use of such RF are the tasks of describing and processing the results of multispectral (up to 10 spectral ranges) and hyperspectral (up to 300 ranges) surveys of earth surface areas. It is necessary, on the one hand, to consider aerospace observations as a single multidimensional aggregate, and on the other hand, to take into account a number of characteristic features of satellite images, for example, a pronounced spatial heterogeneity. Among the tasks of such images processing, the problem of detecting anomalies occupies a special place [4-7]. The examples of such anomalies can be foci of fires, the flare of the starting rocket, polynyas on the ice, shoals of fish in the ocean, etc. At the same time, the background for detection are sequences of polyzonal images, i.e. images of the territory at different times in different spectral ranges. In this paper, the results of synthesis and analysis of algorithms for detecting anomalies on polyzonal satellite images are presented.

## 2. A brief overview of detecting anomalies algorithms

Typically, statistical algorithms for detecting signals (Bayes, Neumann-Pearson) are often used in detection problems, but they require a sufficient amount of a priori information. Nevertheless, the development of statistical algorithms is an actual task. First, for such algorithms, it is possible to use various mathematical models of images. Secondly, the analysis of the effectiveness of such algorithms can be studied both theoretically and experimentally. The algorithms [7] differing in their approaches to the detection of "anomalies" and algorithms based on various image models have been proposed relatively recently. There are following algorithms: the algorithm of spatial-spectral mismatch, in which the image is described by the stationary RF model, the adaptive spectral mismatch algorithm, where the "anomaly" value is determined by the authors, as an error in the representation of the pixel through its neighborhood, and the probabilistic algorithm for detecting anomalies using images signatures quantization. It should be noted that the comparison of the work of algorithms in the work was carried out only with the standard RXD-algorithm.

Another option in anomalies detection task is the detection of anomalies on multidimensional grids using wavelet transform [5]. This method refers to methods with pre-processing, so with its use it is possible to increase the performance of anomaly detection. However, it is difficult to use an algorithm with preliminary discrete wavelet transform when solving real-time anomaly detection problems.

Recently, topological tools have been used to process hyperspectral images, along with ideas from network theory. A standard RX (I. S. Reed, X. Yu) algorithm was proposed. It is based on the calculation of standard deviations of pixels from the mean value in the multidimensional sense. However, it works well only on simple images, such as a large forest, but not on complex urban scenes. Usually algorithms with transition to abbreviated description, local algorithms or algorithms with preliminary segmentation [4] are used for complex images processing.

In our investigation, we will consider an anomaly as an a priori defined and observed object on a polyzonal image. We note that within the framework of this work the signal parameters (its values and location) will be considered known. Otherwise, it would be necessary to conduct a preliminary classification of the anomalies to determine the possible signal levels, and also to search for such anomalies not in a specific region, but throughout the entire image.

### 3. Algorithms for filtering and detecting anomalies against a background of doubly stochastic random fields

Let's imagine a polyzonal image as a collection of data sets. Then we have a polyzonal image consisting of  $N$  components,  $\{z_{ijk}^k, k=1..N, i=1..M_1, j=1..M_2\}$ , which are obtained as a result of spatial discretization of signals received from various sensor systems. When the useful signal is absent (hypothesis  $H_0$ ) the model of observations can be represented by an additive mixture:

$$z_{ijk} = x_{ijk} + \theta_{ijk}, (i, j) \in G^k, k=1..N,$$

of RF  $x_{ijk}$  with zero mean and given correlation function (CF)  $B_{(m)}^{kr} = M\{x_{ij}^k, x_{i+m, j+r}^k\}$  and spatial white noise  $\theta_{ijk}$  with zero mathematical expectation and variance  $\sigma_\theta^2$  in an area  $G$ , where the appearance of a signal is considered impossible for all components of the image.

If there is a useful signal (hypothesis  $H_1$ ) the model of observation  $s$  is written in the form:

$$z_{ijk} = x_{ijk} + s_{ijk} + \theta_{ijk}, (i, j) \in G_0^k, k=1..N,$$

$$z_{ijk} = x_{ijk} + \theta_{ijk}, (i, j) \notin G_0^k, k=1..N,$$

where  $G_0^k$  is the area at the  $k$ -th component of the image, in which we can wait the appearance of a useful signal with known levels  $s_{ijk}, (i, j) \in G_0^k$ . To simplify the calculations, we assume that on each of the components this area has the same form:  $G_0^k = G_0$ . And also we shall assume that the region  $G_0$  is known in advance.

The general solution of the detection problem is based on the construction of the modified likelihood ratio [3]:

$$L = \frac{w(\{z_{ijk}\}/H_1)}{w(\{z_{ijk}\}/H_0)},$$

and comparison to the threshold value. A decision is made in favor of the hypothesis of the existence of a useful signal or a hypothesis about its absence. And the decision is based on the results of the comparison.

Proceeding from the central limit theorem, let us approximate the conditional probability distribution densities  $w(\{z_{ijk}\}/H_1)$  and  $w(\{z_{ijk}\}/H_0)$  by Gaussian [2,3,8-10]:

$$w(\{z\}/H_1) = \frac{1}{\sqrt{2\pi}\sigma_{z1}} \exp\left(-\frac{(z - m_{z1})^2}{2\sigma_{z1}^2}\right), \quad w(\{z\}/H_0) = \frac{1}{\sqrt{2\pi}\sigma_{z0}} \exp\left(-\frac{(z - m_{z0})^2}{2\sigma_{z0}^2}\right),$$

where  $m_{z1}$  and  $m_{z0}$  are mathematical expectations of observations  $\{z_{ijk}\}$  in the presence of a useful signal and in its absence, respectively;  $\sigma_{z1}^2$  and  $\sigma_{z0}^2$  are variations of observations  $\{z_{ijk}\}$  in the presence of a useful signal and in its absence, respectively.

So the optimal signal detection rule can be written in the form [3]:

$$L = \bar{s} V_\theta^{-1} (\bar{z} - \hat{x})^T \begin{cases} > L_0 - \text{signal} & \text{presence,} \\ \leq L_0 - \text{signal} & \text{absence,} \end{cases}$$

where  $V_\theta$  is a diagonal matrix with values  $\sigma_\theta^2$ ,  $\bar{s}$  is extended signal with known characteristics,  $L_0$  is a threshold that can be found based on a given false alarm probability.

For the case of the absence of a useful signal, estimates  $\hat{x}$  are optimal linear estimates in the usual sense of the minimum of variance of errors, based on all available observations  $\{z_{ijk}\}$ . If there is a signal, the  $\hat{x}$  values obtained aren't optimal estimates. It should be considered as a pseudo-evaluation containing in its composition a transformed input signal  $\bar{s}$ .

Thus, the best detection procedure involves the optimal filtration of the RF, the calculation of the covariance matrix of the filtering errors, and the execution of the weighted summation in accordance with the indicated formulas. The most complex of these steps is the filtration of the RF. This is due to the fact that real satellite imagery has a pronounced spatial heterogeneity. Using standard optimal linear filters for such images leads to significant errors. The solution to this problem is possible due to the use of special filters that take into account the complex nature of the images. Consider the synthesis of such filters for the case where correlation between individual components of a polyzonal image can be ignored. In this case, the processing of the image component can be carried out independently of one another. The conducted studies [11-13] show that to form such filters it is possible to use doubly stochastic image models, which allow describing inhomogeneous signals [14]. As an example, consider the following model [8]:

$$\begin{aligned}
x_{ijk} = & 2\rho_{xij}x_{i-1,j,k} + 2\rho_{yij}x_{i,j-1,k} - 4\rho_{xij}\rho_{yij}x_{i-1,j-1,k} - \rho_{xij}^2x_{i-2,j,k} - \rho_{yij}^2x_{i,j-2,k} + \\
& + 2\rho_{xij}^2\rho_{yij}x_{i-2,j-1,k} + 2\rho_{yij}^2\rho_{xij}x_{i-1,j-2,k} - \rho_{xij}^2\rho_{yij}^2x_{i-2,j-2,k} + b_{ij}\xi_{ijk}
\end{aligned} \quad (1)$$

where  $x_{ijk}$  is simulated RF with normal distribution  $M\{x_{ijk}\}=0$ ,  $M\{x_{ijk}^2\}=\sigma_x^2$ ;  $\xi_{ijk}$  is RF of independent standard Gaussian random values  $M\{\xi_{ijk}\}=0$ ,  $M\{\xi_{ijk}^2\}=\sigma_\xi^2=1$ ;  $\rho_{xij}$  and  $\rho_{yij}$  are correlation coefficients of the model with multiple roots of the characteristic equations of multiplicity (2,2) [3];  $b_{ij}$  is a scale factor of the modeled RF.

Random values  $\rho_{xij}$  and  $\rho_{yij}$  with a Gaussian probability distribution density can be described by the following autoregressive equations:

$$\begin{aligned}
\tilde{\rho}_{xij} &= r_{1x}\tilde{\rho}_{x(i-1)j} + r_{2x}\tilde{\rho}_{xi(j-1)} - r_{1x}r_{2x}\tilde{\rho}_{x(i-1)(j-1)} + \sigma_{\rho_x}\sqrt{(1-r_{1x}^2)(1-r_{2x}^2)}\zeta_{\rho_{xij}}, \\
\tilde{\rho}_{yij} &= r_{1y}\tilde{\rho}_{y(i-1)j} + r_{2y}\tilde{\rho}_{yi(j-1)} - r_{1y}r_{2y}\tilde{\rho}_{y(i-1)(j-1)} + \sigma_{\rho_y}\sqrt{(1-r_{1y}^2)(1-r_{2y}^2)}\zeta_{\rho_{yij}}, \\
\rho_{xij} &= \tilde{\rho}_{xij} + m_{\rho_x}, \\
\rho_{yij} &= \tilde{\rho}_{yij} + m_{\rho_y},
\end{aligned} \quad (2)$$

where  $r_{1x} = M\{\tilde{\rho}_{xij}\tilde{\rho}_{x(i-1)j}\}$ ,  $r_{2x} = M\{\tilde{\rho}_{xij}\tilde{\rho}_{xi(j-1)}\}$  are correlation coefficients of a random parameter  $\tilde{\rho}_{xij}$ ;  $r_{1y} = M\{\tilde{\rho}_{yij}\tilde{\rho}_{y(i-1)j}\}$ ,  $r_{2y} = M\{\tilde{\rho}_{yij}\tilde{\rho}_{yi(j-1)}\}$  are correlation coefficients of a random parameter  $\tilde{\rho}_{yij}$ ;  $\zeta_{\rho_{xij}}$  and  $\zeta_{\rho_{yij}}$  are Gaussian random values with  $M\{\zeta_{\rho_{xij}}\} = M\{\zeta_{\rho_{yij}}\} = 0$ ,  $M\{\zeta_{\rho_{xij}}^2\} = M\{\zeta_{\rho_{yij}}^2\} = \sigma_\xi^2 = 1$ .

Note that model (1) with parameters (2) imitates inhomogeneous images [14], which allows us to recommend it for describing real satellite images. In this case, we can use vector (row by row) nonlinear Kalman filter to reduce the noise [11, 12]. To do this, we combine the elements of the image line into a vector  $\bar{x}_i = (x_{i1}, x_{i2}, \dots, x_{iN})^T$ . Then the model of a individual component of the image can be written in the form:

$$\bar{x}_i = \text{diag}(\bar{\rho}_{xi})\bar{x}_{i-1} + \nu(\bar{\rho}_{xi}, \bar{\rho}_{yi})\bar{\xi}_i, \quad \bar{\rho}_{xi} = r_{1x}\bar{\rho}_{x(i-1)} + \nu_{\rho_x}\bar{\xi}_{\rho_x}, \quad \bar{\rho}_{yi} = r_{1y}\bar{\rho}_{y(i-1)} + \nu_{\rho_y}\bar{\xi}_{\rho_y},$$

where  $\text{diag}(\bar{\rho}_{xi})$  is diagonal matrix with elements  $\bar{\rho}_{xi}$  on the main diagonal; lower-triangular matrix  $\nu$  is the matrix, which is determined by the decomposition of the covariance matrix:  $V_x = \nu\nu^T$ .

The process of row by row estimation is described by a nonlinear Kalman filter:

$$\hat{x}_{pi} = \hat{x}_{spi} + P_i \frac{\partial \Phi^T}{\partial \bar{x}_{pi}} V_n^{-1} (\bar{x}_i - \hat{x}_{spi}), \quad \bar{x}_{pi} = \begin{pmatrix} \bar{x}_i \\ \bar{\rho}_{xi} \end{pmatrix} = \Phi(\bar{\rho}_{x(i-1)}, \bar{x}_{i-1}) + \nu(\bar{\rho}_{x(i-1)}, \bar{\rho}_{y(i-1)})\bar{\xi}_i,$$

where  $\bar{x}_{spi} = \Phi(\bar{x}_{p(i-1)})$ ,  $\Phi_p(\bar{x}_{p(i-1)}) = \begin{pmatrix} \Phi(\rho, x) \\ r_{1x}\bar{\rho}_{x(i-1)} \\ r_{1y}\bar{\rho}_{y(i-1)} \end{pmatrix}$ ,  $\bar{\xi}_i = \begin{pmatrix} \bar{\xi}_i \\ \bar{\xi}_{xi} \\ \bar{\xi}_{yi} \end{pmatrix}$ ,  $P_i$  is covariance filter error matrix.

The use of this algorithm is possible under the condition of precisely known characteristics of the information RF. So we need to know coefficients  $r_{1x}$ ,  $r_{2x}$ ,  $r_{1y}$ ,  $r_{2y}$ , and also parameters  $\rho_{0x}$ ,  $\rho_{0y}$  and  $\sigma_{\rho_x}^2$ ,  $\sigma_{\rho_y}^2$ ,  $\sigma_x^2$ . Otherwise, a preliminary evaluation of these parameters is necessary. For this, pseudo-gradient procedures [13,15] can be used, as well as expressions for CF of doubly stochastic RF models [14].

#### 4. Results of the investigation of the efficiency of detection of signals on real images

Let's compare two detectors of anomalies constructed on the basis of a doubly stochastic model (**Algorithm 1**) and on the basis of the usual autoregressive model [2] (**Algorithm 2**). In this case, the detection will be performed on real images obtained from the LandSat-8 satellite. Studies are conducted for three images. We choose 4 areas for each image, where an anomaly may occur. It should be noted that the areas are selected based on the structure of the images to be examined, taking into account the greater and smaller heterogeneity, and the detection procedures are performed not for the entire image, but only for these areas. Fig. 1a-1c show examples of images with signals located in different parts of the images, and also reflect the probabilities of correct detection obtained using two algorithms. The sizes of all images are 250x250. The images are distorted by white Gaussian noise with a single dispersion. The size of the square is 4x4, the radius of the circle is 2. The signal-to-noise ratio is 1. The statistics are removed 150 times.

Table 1 shows the gain of **Algorithm 1** in relation to **Algorithm 2** for the magnitude of the threshold signal when the probability of correct detection is 0.5 and the probability of false alarm is 0.001. It corresponds to the threshold  $L_0 = 3,1\sigma_{z_1}^2$ .

Table 1. Gain (in percent) of the proposed detection algorithm based on a doubly stochastic model in comparison with the detection algorithm based on the AR model

Shape/Image	Location 1	Location 2	Location 3	Location 4
Square in Image 1	0	0	0	0
Circle on Image 1	5	2	0	2
Square in Image 2	68	3	13	4
Circle on Image 2	60	4	3	5
Square in Image 3	21	4	4	5
Circle on Image 3	70	5	7	7

Analysis of the results shows that the algorithm based on the doubly stochastic model works better than the algorithm based on the autoregressive model and provides reliable detection of the signal in 90-95% of cases. The small values of the gains in Table 1 are explained by the fact that the signals have small dimensions, and their neighborhoods are on a comparable scale to homogeneous ones. If the signal is "at the junction" of homogeneous regions, an algorithm based on a doubly stochastic model provides a significant (up to 70%) gain in the signal level term. The gains presented in Table 1 are calculated for each case from expression

$$Gain = \frac{Pd_{ds} - Pd_{ar}}{Pd_{ar}},$$

where  $Pd_{ds}$  and  $Pd_{ar}$  are percentages of correctly detected signals based on doubly stochastic and autoregressive models, respectively.

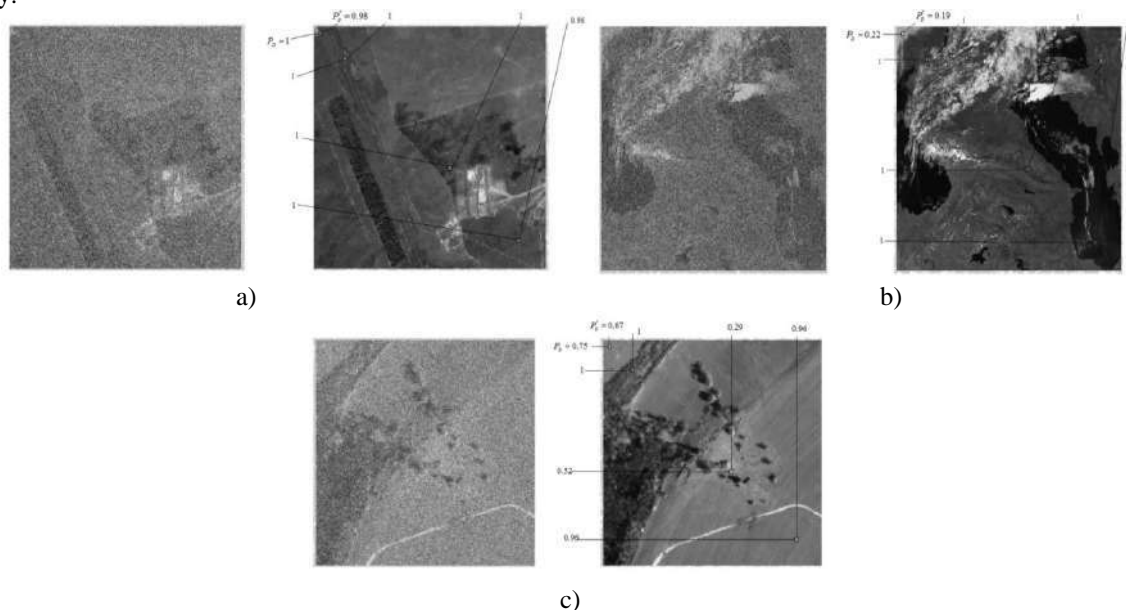


Fig. 1. The noisy image (left) and the source images (right) with the probabilities of correct detection of a square signal: on the left the probabilities for **Algorithm 1** are presented, on the top the probabilities for **Algorithm 2** are presented.

Analyzing the Fig. 1, we can conclude that we also have gains in correct detection probability terms on equal signal-to-noise ratios. Furthermore the probability of correct detection depends not only on the shape and sizes of the signal itself, but also on the brightness values in its immediate neighborhood. In this sense, a more universal algorithm is an algorithm based on doubly stochastic RF models.

## 5. Increase of accuracy of object recognition due to its preliminary detection

Consider the task of recognizing objects in images. Usually, to solve this problem, binarization of the processed image is used. However, the preliminary detection of the anomaly allows us to abandon the complicated segmentation and binarization procedures.

As an example, consider a discrete doubly stochastic model:

$$x_{ij} = \rho_{xij} x_{i-1j} + \rho_{yij} x_{ij-1} - \rho_{xij} \rho_{yij} x_{i-1j-1} + \xi_{ij}, i = 1, \dots, M; j = 1, \dots, N, \quad (3)$$

where  $\{\xi_{ij}\}$  is the field of Gaussian random variables with constant mathematical expectation  $M\{\xi_{ij}\} = 0$  and variation  $M\{\xi_{ij}^2\} = \sigma_\xi^2 = \sigma_x^2(1 - \rho_{xij}^2)(1 - \rho_{yij}^2)$ , changing at every point of image,  $M \times N$  are the image sizes.

It should be noted that the correlation coefficients in the row and column in the model (3) represent the realization of a discrete RF of the following form:

$$\rho_{xij} = \begin{cases} \rho_{x1}, (i, j) \in I_1 \\ \rho_{x2}, (i, j) \in I_2 \end{cases}, \rho_{yij} = \begin{cases} \rho_{y1}, (i, j) \in J_1 \\ \rho_{y2}, (i, j) \in J_2 \end{cases}. \quad (4)$$

Thus, the correlation parameters in expression (4) are a binary RF. Indeed, the elements of each of the fields in (4) can take only two values, so their binarization by converting some values to a minimum value of brightness ( $Y = 0$ ) and others to a maximum value ( $Y = 255$ ) does not cause any special difficulties.

If the anomaly is characterized by a sufficiently high level of brightness, then for its detection and subsequent identification, known methods using brightness characteristics of the image can be used. An example of such methods can be statistical analysis of image histograms. However, the processing results will be unsatisfactory at signal levels comparable to the background level and less than it. To improve the efficiency of processing, it is proposed to perform preliminary detection of objects of interest. Tables 2 and 3 show the results of binarization of the image by brightness and by the selection of the signal area. All the results are obtained against a background of doubly stochastic images. There were cases when one signal was present on the image: square (Table 2) or circle (Table 3). The ratio of the side of the square signal and the diameter of the circular signal to the image length is 10%. For detection, the probability of false alarm was set as  $P_f = 0.01$ .

Table 2. Binarization of an image containing a square signal

Signal-to-noise ratio	0.1	1	3	5	6
Binarization based on detection, %	34	58	94	100	100
Binarization based on brightness, %	0	12	33	78	100

Table 3. Binarization of an image containing a circular signal

Signal-to-noise ratio	0.1	1	3	5	6
Binarization based on detection, %	29	52	92	100	100
Binarization based on brightness, %	0	8	26	39	80

According to Tables 2 and 3, we can conclude that the binarization algorithm, using the results of detection, significantly exceeds the brightness binarization. So at small signal-to-noise ratios, the first algorithm achieves a gain of 60-70%. This gain is observed for both signal forms (square and circle). Note that the effectiveness of algorithms falls with the use of a circle-shaped signal. This is explained by the smaller area of this signal compared to the square one.

Let the anomaly in the image be either circular or square. Then you need to find the area of the object and its center, and then by comparing the fill factor (it is  $d = 1$  for square signal, it is  $d = \pi/4$  for circular signal) with the threshold to assign it to a particular class.

Figure 2 shows the result of the operation of algorithms for images with a high level of brightness of the anomaly. In both cases, the binarization was correct.

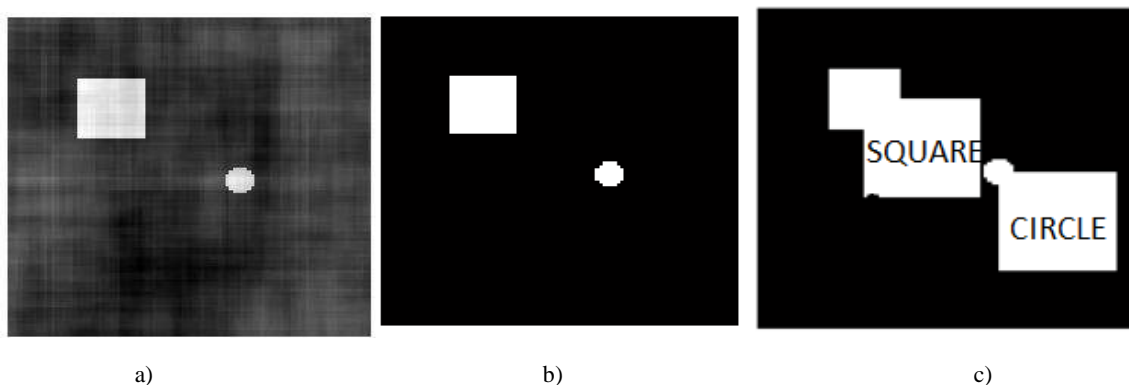


Fig. 2. Recognition of objects on the image: a - the original image, b - binarization, c - the recognition result.

Thus, the proposed algorithm for detecting signals can improve the quality of binarization of images and recognition of anomalies of the simplest geometric shape on them.

## 6. Conclusion

Synthesis was carried out and the efficiency of correct detection based on algorithms using doubly stochastic RF models was studied in the text. Statistical modeling showed that the algorithm using vector Kalman filtering for models with variable

parameters allows to achieve significant gains in comparison with the algorithm based on filtering for models with constant parameters in conditions of imitation of images based on a doubly stochastic model of RF. The main advantage of vector filtering for doubly stochastic images lies in the possibility of estimating the change in image parameters. The developed algorithm is also applicable to the detection of extended signals in images. In this case, the use of detection results allows to significantly improve the detection quality of detectable low-contrast objects.

## Acknowledgements

This work was supported by RFBR grant 16-41-732-027 "Construction of stochastic models and algorithms for processing sequences of inhomogeneous polyzonal images for regional environmental monitoring systems".

## References

- [1] Kazarinov YuM. Radio engineering systems: a textbook for students of universities. Moscow: Publishing Center "Academy", 2008; 592 p.
- [2] Perov AI. Statistical theory of radio engineering systems: a textbook for high schools. Moscow: Radio Engineering, 2003; 400 p.
- [3] Vasiliev KK, Krashennnikov VR. Statistical analysis of images. Ulyanovsk: UISTU, 2014; 214 p.
- [4] Borghys D, Achard V, Rotman SR, Gorelik N, Perneel C, Schweicher E. Hyperspectral anomaly detection: A comparative evaluation of methods. General Assembly and Scientific Symposium, XXXth URSI 2011: 1–4.
- [5] Baghbidi MZ, Jamshidi K, Nilchi AR, Homayouni S. Improvement of Anomaly Detection Algorithms in Hyperspectral Images Using Discrete Wavelet Transform. *Signal & Image Processing: An International Journal (SIPIJ)* 2011; 2(4): 13–25.
- [6] Soofbaf SR, Valadan Zoej MJ, Fahimnejad H, Ashoori H. Efficient detection of anomalies in hyperspectral images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2008; XXXVII(B7): 303–308.
- [7] Denisova AY, Myasnikov VV. Detection of anomalies on hyperspectral images. *Computer Optics* 2014; 38(2): 287–296.
- [8] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Doubly stochastic models of images. *Pattern Recognition and Image Analysis* 2015; 25(1): 105–110. DOI: 10.1134/S1054661815010204.
- [9] Vasiliev KK, Tashlinsky AG, Krashennnikov VR. Statistical Analysis of Multidimensional Image Sequences. *High technology* 2013; 5: 5–11.
- [10] Vasiliev KK, Dementiev VE, Andriyanov NA. Estimation of the parameters of doubly stochastic random fields. *Radiotekhnika* 2014; 7: 103–106.
- [11] Vasiliev KK, Dementiev VE, Andriyanov NA. Analysis of the effectiveness of estimating the changing parameters of a doubly stochastic model. *Radiotekhnika* 2015; 6: 12–15.
- [12] Vasiliev KK, Dementiev VE, Andriyanov NA. Detection of extended signals against a background of doubly stochastic images. *Radiotekhnika* 2016; 9: 23–27.
- [13] Vasiliev KK, Dementiev VE, Andriyanov NA. Application of mixed models for solving the problem on restoring and estimating image parameters. *Pattern Recognition and Image Analysis* 2016; 26(1): 240–247. DOI: 10.1134/S1054661816010284.
- [14] Andriyanov NA. A method for fitting images based on models of random fields with varying parameters. *Uspekhi sovremennoi nauki* 2016; 5(9): 98–100.
- [15] Andriyanov NA. Pseudo-gradient procedures in estimation problems of image model parameters. 26th International Crimean Conference "Microwave Engineering and Telecommunication Technologies" (Crimea, Russia). Sevastopol, September 4-10, 2016; 1: 2705–2710.

# Voice command recognition for noisy environments by means of cross-correlation portraits

A.I. Armer<sup>1</sup>, E.Yu. Galitskaya<sup>2</sup>, N.A. Krashennnikova<sup>3</sup>

<sup>1</sup> Ulyanovsk State Technical University, Severny Venets St., 32, Ulyanovsk, 432027, Russia  
<sup>2</sup> Ulyanovsk Instrument Manufacturing Design Bureau, Krymov St., 10a, Ulyanovsk, 432071, Russia  
<sup>3</sup> Ulyanovsk State University, Lev Tolstoy St., 42, Ulyanovsk, 432017, Russia

---

## Abstract

Methods of voice command (VC) recognition in heavy noise environments are required for precise work of speech information systems on the factory floor and in transport. The paper considers a speaker-dependent way of VC recognition for VCs belonging to a limited vocabulary and being recognized in heavy noise environments. For this purpose, VCs are transformed into cross-correlation portraits (CCPs), i.e. special images. The VC under recognition is referred to a class with a minimal distance (metric) between CCP of this command and model CCPs of the class. The authors elaborated algorithms for VC transformation into CCPs, a method for defining VC boundaries, ways of model command optimization and metric choice. As a result, a rather precise VC recognition in heavy noise environment was obtained.

*Keywords:* voice command; intensive noise; recognition; cross-correlation portrait; metric; precise definition of boundaries; model command; optimization of VC library

---

## 1. Introduction

The growth of production and transport intensity leads to increase in operator burden. To reduce such workload, speech information systems are used. However, these systems often have to recognize VC precisely, especially for noisy environments. At present, a large number of speech recognition systems functioning in nearly noiseless environment have been developed. They include, for example, IBM Via Voice, its recognition accuracy is reported to be 97% and its recognition vocabulary includes up to 2,000 VCs; Dragon NaturallySpeaking or Dragon for PC, this software package accurately recognizes 70% of the vocabulary, which includes nearly 60,000 words; L&H Voice XPress, its accuracy is in the range of 90%-98% and its vocabulary size is nearly 1,000 words, etc. There are also user-friendly systems of continuous speech understanding and processing, such as VocalIQ, Siri, Google Now and Cortana. To compare VocalIQ with Siri, Google Now and Cortana the systems were given multiaspect requests in a natural language [1]. The correct recognition rate was more than 90% for VocalIQ, while Google Now, Siri and Cortana showed only 20% accuracy. Among home-grown technologies it is necessary to mention VoiceCom STC. It is reported to recognize 100-200 VCs in a speaker-dependent version and 30-50 VCs in a speaker independent one with accuracy 98%. However, these systems do not accurately work even in low noise environment. Recognition systems for VCs from a limited vocabulary under acoustic noise are currently being developed mainly for aviation and are used in voice control and flight control devices. Performance quality of such systems today is from 90 up to 98% of accurate VC recognition, depending on the test conditions and vocabulary size. Almost all tested systems are speaker-dependent. According to the Air Force Research Laboratory - Wright-Patterson Air Force Base, flight tests of an ITT VRS-1290 speaker dependent, continuous speech recognition system and a Verbex VAT31 showed the following results: average word accuracy for VRS-1290 was 92-98%, if the vocabulary consisted of 50 commands; average word accuracy for VAT31 was up to 97% (no information on vocabulary size is available). In 1997, flight test results of the VC recognition system produced by National Research Council (Canada) were obtained. The system was integrated into Bell 412HP Avionics Management System and showed an average 95% accuracy for vocabulary consisting of 80 words, which were divided into 24 groups. According to the Smiths Industries Speech Recognition Module system built into the CAMU of the Eurofighter, the accuracy of VC recognition in a standard aircraft flight is at least 95% for a vocabulary consisting of 250 words, 25 of which can be simultaneously active. Currently, Thales Avionics develops a VC recognition system for Rafale fighters. The VC recognition accuracy is required to be above 95% for a vocabulary of 50-300 words. A 5-th generation jet fighter F-35 was equipped with DynaSpeak

VC recognition system developed by SRT International. The developers report the recognition accuracy to be 98%. A multipurpose 4-th generation Eurofighter is equipped with a voice control system developed by Logica. The vocabulary consists of 250 words, and the average VC accuracy is not less than 95%. The developers declare, that for the export version of the Rafale Block 05t, Thales Avionics has developed a speech control system with recognition accuracy not less than 95% for a 300 VC vocabulary, but no information on its implementation is available. Patent US 6529866 B1, 4 March 2003, The United States of America as Represented by the Secretary of the Navy, describes a method and system for transformation of an audio signal into speech. Audio signals are said to contain both VC units and noise, but test and implementation information is not available. Patent WO 1999040571 A1, 3 February 1999, Qualcomm Incorporated, describing a system and method for improving speech recognition accuracy in noisy environment also provides no test or implementation data. Among home-grown technologies the following ones should be noted. First of all, it is a VC recognition system tested on the Mikoyan MiG-29 (Fulcrum). Recognition accuracy is reported to be 56-81%, no information on the vocabulary is available. Patent RF 2267820 1, 25 April 2006, Ulyanovsk State Technical University. Recognition accuracy is reported to be 92%, vocabulary size is 23 VCs, and noise level is 3dB. No information on implementation is available. Patent RF 2271578 2, 10 March 2006, Speech Technology Center. The invention relates to speech analysis under adverse environmental conditions, e.g. in moving transport or high level noisy workplaces. No test information is available. Despite the available developments, there is no information on the actual application of VC recognition systems in avionics, since in real flights the systems developed showed substantially less efficiency than anticipated. Thus, developing VC recognition systems for noisy environments remains a challenging task. This paper examines a speaker dependent technique of VC recognition for a limited vocabulary. A method of VC transformation into portraits, i.e. images, is used.

## 2. Methods of VC recognition

The problems of speech recognition, in particular VC recognition, are widely discussed in modern literature. The first methods of automatic sound recognition were obtained in the first half of the 20-th century [2]. Among speech recognition techniques one can distinguish the following approaches: spectral methods [3, 4, 5, 6, 7], wavelet transform [3], statistical methods [5, 8, 9, 10, 11], and neural networks [12, 13].

This paper deals with VC recognition based on their transformation into portraits, i.e. flat images, and further implementation of image processing techniques [14, 15, 16, 17, 18].

## 3. Autocorrelation portraits

Let  $S = s_0, s_1, s_2, s_3, \dots, s_{N-1}$  be digital VC readouts. Then, a two-dimensional image  $X(i, k) = \{x_{ik} : i = 1, 2, 3, \dots; k = 1..K\}$  will be its autocorrelation portrait (ACP). This image is obtained in the following way. Let us divide VC  $S$  into  $M$  segments and perform the following transformations

$$X(i, k) = \frac{Cov(S_n, S_{n+k})}{\sigma_n \sigma_{n+k}}, \quad (1)$$

where  $Cov(S_n, S_{n+k})$  is a sample covariation of signal  $S$  intervals  $S_n, S_{n+k}$ , which are spaced  $k\Delta t$  apart,  $\sigma_n^2, \sigma_{n+k}^2$  are sample dispersions of segments  $S_n, S_{n+k}$  respectively. Thus, the  $k$ -th element of the  $i$ -th ACP line is equal to the correlation coefficient between the  $i$ -th segment  $S_i$  and the segment shifted left with respect to  $S_i$  on  $k$  readouts. Fig. 1 shows ACP examples.

Let us note some ACP characteristics, which make them favorable for VC recognition. VC portraits are unique, i.e. ACPs of different VCs are unlike, whereas ACPs of the same VCs pronounced at different time intervals are the same. Autocorrelation transformation normalizes a signal, as a result ACPs are nearly insensitive to noisiness and slowly varying additives. If we consider additive white noise with dispersion  $\sigma_\theta^2$ , then its ACPs and VC ACP readouts distorted by noise will differ by a constant factor. However, ACPs also have some negative characteristics, e.g. the dependence of element brightness on the differences in the tone of VC pronunciation, as well as geometric ACP distortions due to variations in speech rate. These distortions can be steadied by modifying ACP development, e.g. taking into account loudness extremum. VC recognition by their ACPs is conducted in the following way. ACPs of model VCs are stored in the memory. VC under recognition is transformed into ACP and it is referred to the class



with a minimal distance between its model portrait and ACP of a recognized VC. This distance (metric) between two ACPs (i.e. images) is calculated as follows. At first, two images are aligned, i.e. for each line of one image a corresponding line of another image is found. The average distance (e.g. Euclidean) between the corresponding lines is considered to be the distance between the portraits. Such a correspondence for ACP of one and the same command means the proximity of VC fragments, so the distance is relatively small, since it only occurs from the difference in pronunciation and surrounding noise. If ACPs of different VCs are compared, then this distance is usually much more visible due to the larger difference in sounds. In the process of command alignment dynamic programming based on minimum distance criterion was applied. While testing the accuracy of VC recognition, commands were pronounced by the speaker in real time. The vocabulary used consisted of ten VC groups, and there were 4-23 aviation commands in each group. In total, the vocabulary included more than 100 VCs. Aircraft engine noise recorded in a flight mode was used as a background and reference noise, the signal-to-noise ratio was 5-0 dB. Four male speakers took part in the tests. Before the experiment each speaker recorded model VCs, each VC belonging to the given vocabulary was pronounced twice. During the experiment on VC recognition each speaker pronounced all the commands from the given vocabulary three times, all in all, more than 1,200 VCs were recorded during the experiment. Average command accuracy was more than 95%. However, further processing has shown that the probability of accurate VC recognition can be significantly reduced in the course of time. This problem is connected with model aging, i.e. speaker's voice pattern can change with time, and previously pronounced command models will not reflect the peculiarities of the speaker's voice at the very time of VC recognition. Therefore, it is required to update the commands from time to time (e.g. before the flight), which, of course, has certain inconveniences. One VC model does not reflect all the possible variants of its pronunciation, so the number of VCs was increased, i.e. the speaker pronounced each VC more than once at different periods of time. The totality of all these patterns somehow reflected pronunciation diversity. However, the increase in model number complicates and slows down the recognition algorithm, but it is permissible only to a certain extent. Therefore, the model number should be limited. Besides, these models should reflect the pronunciation diversity as much as possible. It turns out, that recognition accuracy depends greatly on the correctness of model choice, and recognition deviations can be more than 10%. Thus, among several pronunciations it is necessary to choose a certain number of VCs as model ones, so that the obtained model library contributed to the best VC recognition accuracy. This problem of model library optimization was examined in [19, 20, 21]. Technically it is impossible to conduct complete enumeration of all library patterns. That is why, a method of direct enumeration giving an almost optimal result has been developed. Sometimes it is possible to change the VCs themselves, using their synonyms. This problem was also considered and its solution was found while analyzing the synonym rings.

#### 4. Cross-correlation portraits

Another way to decrease the impact of VC pronunciation variability is to use a different kind of portraits instead of ACPs. In the process of ACP development correlation coefficients between the segments of the same VC (autocorrelation) are found. When ACPs are used for recognition, the distances between the ACP of a recognized command and the ACP of a model command are found. If the distance between the ACP of the command under recognition and the ACP of its model is found, the ACPs of two different pronunciations of this command will be compared. These ACPs can significantly differ from each other (the distance will be large). Therefore, when comparing portraits it is desirable to minimize the difference in pronunciation. For this purpose, it is necessary for pronunciation variability to be somehow reflected in portraits. Let us consider a cross-correlation portrait (CCP), which consists of correlation coefficients between segments of two VCs (cross-correlation) [15, 16, 17, 22]. Let there be two VCs  $S_1$  and  $S_2$ . Let us segment each command into  $M$  segments of the same length and determine the sample correlation coefficients  $x_{ik}$  between the  $i$ -th segment of VC  $S_1$  and a VC segment  $S_2$ , beginning with the  $k$ -th readout of the VC  $S_2$   $i$ -th segment. As a result, we get a two-dimensional array (image)  $X = \{x_{ik}\}$ , called a CCP of VCs  $S_1$  and  $S_2$ . Let us consider CCP development in detail. As an example, let us consider the CCP development of two pronunciations of one avionics VC, the first pronunciation is  $S_1$  and the second pronunciation is  $S_2$ . Let us divide each VC into equal segments, whereas  $N_1$  is the length of each interval for signal  $S_1$ , and  $N_2$  is the length of each interval for signal  $S_2$ . Let  $N = \min\{N_1, N_2\}$  be the minimal of these lengths. While specifying the number of intervals for each command  $M$  it should be taken into account that if the segment length is too small it will not include the whole phoneme; otherwise, if the segment length is big enough it will include several phonemes. Such segmentation will negatively

affect the correlation coefficient between separate phonemes in different VCs while developing CCPs. Let's determine correlation coefficients of signal  $S_1$   $i$ -th segment and signal  $S_2$   $i$ -th segment, shifted  $k = 0..K$  readouts right.

$$x_{ik} = \frac{\frac{1}{N} \sum_{j=0}^{N-1} S_{1iN1+j} S_{2iN2+j+k} - \mu_{1i} \mu_{2i,k}}{\sigma_{1i} \sigma_{2i,k}}, \quad (2)$$

$$\mu_{1i} = \frac{1}{N} \sum_{j=0}^{N-1} S_{1iN1+j}, \quad (3)$$

$$\mu_{2i,k} = \frac{1}{N} \sum_{j=0}^{N-1} S_{2iN2+j+k}, \quad (4)$$

$$\sigma_{1i}^2 = \frac{1}{N} \sum_{j=0}^{N-1} S_{1iN1+j}^2 - \mu_{1i}^2, \quad (5)$$

$$\sigma_{2i,k}^2 = \frac{1}{N} \sum_{j=0}^{N-1} S_{2iN2+j+k}^2 - \mu_{2i,k}^2. \quad (6)$$

While choosing parameter  $K$ , it is necessary to take into account the following fact: if its value increases, than value  $x_{ik}$  decreases. It is connected with correlation reduction of VC readouts along the line. This property proves the inadvisability of using large values  $K$  while developing CCPs (large  $K$  means that  $K > N$ ).

Obviously, if CCPs of the same pronunciation ( $S_1 = S_2 = S$ ) are developed, we get the ACP of a VC  $S$ . It is desirable to examine the CCP of two pronunciations of one and the same command. It depends on two pronunciations, so the pronunciation variability affects the portrait form. Fig. 2 shows CCPs of several VCs. For example, in the picture Manevr3 + Manevr4 'plus' means that this very CCP was obtained from the third and fourth pronunciations of the VC "Manevr".

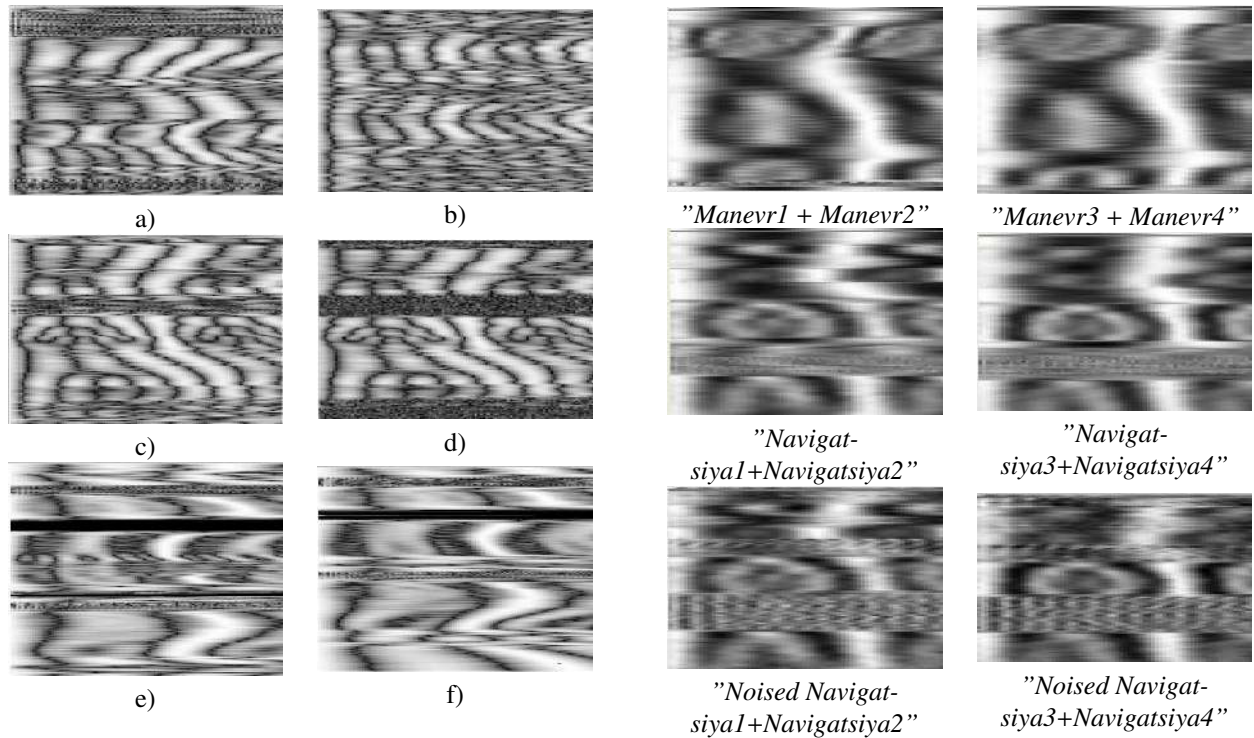


Figure 1: Autocorrelation portraits: a) VC "Svet bol'she", b) VC "Svet bol'she" on the background of aircraft engine noise, c) VC "Konditsioner", d) VC "Konditsioner" on the background of white noise with a mean zero and dispersion equal to five, e), f) VC "Vysota absolutnaya" pronounced at different times.

Figure 2: Cross-correlation portraits of VCs.

Note, that CCP characteristics are similar to those of ACP. But CCPs are less pronunciation dependent, as they combine two different pronunciations. VC recognition by means of CCPs is carried out in the same way as recognition by means of ACPs. For each VC, a model CCP made of two pronunciations of this VC is developed. These model CCPs are stored in the memory. For the VC under recognition CCPs are developed with one of pronunciations of each

command group, then the distance between this CCP and the model CCP is found. The recognized VC is related to the group with the smallest distance.

## 5. Optimization of voice command recognition by means of their CCPs

The CCPs used have a number of characteristics, which come from both the properties of the speech signals themselves and the structure of their CCP development. Let us consider some techniques increasing the recognition accuracy by means of CCPs.

### 5.1. Noisy models

If a VC under recognition is too noisy, it increases the distance from its CCP to its model portrait formed by means of noiseless pronunciations. Therefore, 'noisy models' were used in the experiment, i.e. artificial noise was added to the model commands. It came from the microphone placed far from the operator. As a result, the distorted models and the command under recognition contained approximately the same noise, which significantly increased the recognition accuracy.

### 5.2. Precise definition of boundaries

While developing portraits, it is desirable for the VC time boundaries to be defined as precisely as possible. Then a more accurate portrait alignment can be attained. Among several known techniques of useful signal detection, the one, which shows the most accurate recognition results on the background of noise, was chosen. Besides, after definition of VC boundaries by means of this technique some boundary adjustments were made, which resulted in recognition accuracy.

### 5.3. Pause removal

In some VCs, e.g. those consisting of two words, there are micro-pauses between speech units. These pauses can differ in duration, but they do not contain any information. So, a special method for their removal was developed.

### 5.4. Optimization of portrait width

Portrait width, i.e. the line length, can be chosen arbitrary. So, it is desirable to choose the optimal length, which contributes to the best recognition accuracy. It turned out, the line length in the portrait of a VC under consideration should be equal to  $K = D/(5M)$ , where  $D$  is the length of the recognized VC,  $M$  is the number of lines in a portrait. The line lengths of model CCPs are a bit longer, but they are no less than  $K$ .

### 5.5. Choice of metric

VC recognition by means of their CCP is based on detection of the portraits, which are as similar as possible. Hence, there appears a problem to define the distance between two CCPs, i.e. metric defined on CCP. This distance is considered to be equal to the average distance between the corresponding CCP lines. Moreover, any metric defined on the lines, i.e. on finite sequences or vectors, can be used. Twelve known metrics (namely, Euclidean, Hilbert, Zhuravlev, etc.) and their variants were tested. For the purpose of the problem under consideration, five metrics showed the best results: Zhuravlev method (for  $\varepsilon = 10$ ,  $\varepsilon = 20$  and  $\varepsilon = 30$ ), the Ruzicka distance and the Bray-Curtis distance. Besides, analyzing the recognition results obtained while using these metrics it was found out that certain recognition errors corresponded to certain metrics. Therefore, it is possible to improve recognition accuracy by using, for example, two metrics. If the recognition results coincide, then the command is considered to be recognized; if the recognition results differ, the command should be considered unrecognized. In such a case, the speaker should pronounce the command once again.

### 5.6. Optimization of a model library

As in the case of VC recognition by means of CCPs, the words included in the model library significantly affect the recognition accuracy. Therefore, it is required to optimize the model portrait library while recognizing VCs by means of CCPs.

### 5.7. Fourier analysis

Each CCP line is a sequence of correlation function sample values. Because of speech signal quasi-periodicity, the correlation function turns out to be similar to periodic. This quality was used to improve the portrait quality by removing insignificant harmonics from each CCP line spectrum. This operation was performed by means of FFT. The isolation of the most fundamental harmonics for each CCP line reduced the influence of speech signal pronunciation variability.

## 6. Results

The experiments showed that using CCPs with the described above modifications significantly reduced the effect of VC pronunciation variability and model aging. The recognition accuracy was nearly the same as in the ACP recognition on newly-pronounced models. Thus, to evaluate the efficiency of the suggested method, an experiment was conducted. The recognition was tested on two groups of VCs consisting of 10 commands each. The first group included single-word commands, the second group of VCs included both single-word and two-word commands. Each VC was pronounced 100 times by a woman-speaker. The experimental results are represented in Table 1. The maximum VC recognition accuracy was 95.6%.

Table 1: VC recognition accuracy by means of CCPs.

Commands	Signal/noise ratio (dB)				
	5	4	3	2	1
Group 1	95.6	90.1	87.4	82.9	61.2
Group 2	94.6	92.1	87.4	83.5	67.2

## 7. Conclusion

The present work suggests and examines a speaker-dependent method for recognizing voice commands from a limited vocabulary in conditions of intense acoustic noise, e.g. on the background of an aircraft engine. This method implies transformation of digitized commands into certain images and further application of image processing methods. The method underwent various modifications in order to increase the recognition accuracy. Tests on a large number of voice commands showed rather high efficiency of the suggested method.

## References

- [1] Businessinsider. How apples vocaliq ai works [Electronic resource]. — 2017. — URL: <http://uk.businessinsider.com/how-apples-vocaliq-ai-works-2016-5>.
- [2] Rabiner, L. Tsifrovaya obrabotka rechevykh signalov [Digital processing of speech signals]; translated from English. Edited by M.V. Nazarov and Yu.N. Prokhorov / L.R. Rabiner, R.V. Shafer. — Moscow, Russia. : Nauka, 1981. —P. 495. (in Russian)
- [3] Boykov, F. Primenenie veyvlet-analiza signala v sisteme raspoznavaniya rechi [Wavelet analysis in speech recognition] / F.G. Boykov, Starozhilova T.K. // Trudy mezhdunarodnoy konferentsii Dialog 2003 [Proceedings of the international conference Dialogue 2003]. — Zvenigorod, Russia. —2003. —Pp. 12–19. (in Russian)
- [4] Gudonavichyus, R. Raspoznavanie rechevykh signalov po ikh strukturnym svoystvam [Speech signal recognition by means of their structural characteristics] / R.V. Gudonavichyus, P.P. Kemeshis, A.B. Chitavichyus. —Leningrad, USSR. : Energiya, 1977. —P. 64. (in Russian)
- [5] Myasnikova, E. Ob"ektivnoe raspoznavanie zvukov rechi [Objective recognition of speech sounds] / E.N. Myasnikova. — Leningrad, USSR. : Energiya, 1967. —P. 148. (in Russian)
- [6] Pikone, D. Metody modelirovaniya signala v raspoznavanii rechi [Signal modeling methods in speech recognition] / D. Pikone. —Kemerovo, Russia, 2000. —P. 79. (in Russian)
- [7] Potapova, R. Rech': kommunikatsiya, informatsiya, kibernetika [Speech: communication, information, cybernetics] / R.K. Potapova. — Moscow, Russia.: Radio i svyaz', 1997. —P. 568. (in Russian)
- [8] Sorokin, V. Skrytye markovskie modeli v raspoznavanii rechi [Hidden markov models in speech recognition] / V.N. Sorokin, V.A. Sukhanov // Rechevaya informatika [Speech informatics]. Collected papers edited by V.V. Zyablov. — Moscow, Russia. —1989. —Pp. 104–118. (in Russian)
- [9] Peinado, A. Discriminative codebook design using multiple vector quantization in hmm-based speech recognizers / A. Peinado, J. Segura, A. Rubio [et al.] // IEEE Trans. Speech and Audio Processing. —1996. —Vol. IV, No. 2. —Pp. 89–94.
- [10] Jelinek, F. Statistical Methods for Speech Recognition / F Jelinek. —Cambridge. : MIT Press, 1998.

- [11] Shahshahani, B. A markov random field approach to bayesian speaker adaptation / B. Shahshahani // IEEE Trans. Speech and Audio Processing. "—1997. "—Vol. V, No. 2. "—Pp. 183–191.
- [12] Fedyaev, O. Neyroseteovoy interpretator rechevykh komand dlya upravleniya programmnyimi sistemami [Neural network interpreter of voice commands for program system processing] / O.I. Fedyaev, S.A. Gladunov // Proceedings of the 7th All-Russian conference "Neural computers and their usage", eduted by A.I. Galushkin. "—Moscow, Russia. "—2001. "—Pp. 298–301. (in Russian)
- [13] Lippmann, R. Neural classifiers useful for speech recognition / R. Lippmann, B. Gold // in. Proc. IEEE First Int. Conf. Neural Net. "—Vol. IV. "—1987. "—Pp. 417–422.
- [14] Krasheninnikov, V. Raspoznavanie rechevykh komand na fone intensivnykh шумов s pomoshch'yu avtokorrelyatsionnykh portretov [Speech command recognition on the background of noise using autocorrelation portraits] / V.R. Krasheninnikov, A.I. Armer, N.A. Krasheninnikova, A.V. Khvostov // Naukoemkie tekhnologii. "—2007. "— 9. "—Pp. 65–76. (in Russian)
- [15] Krasheninnikov, V. Cross-correlation portraits of voice signals in the problem of recognizing voice commands according to patterns / V.R. Krasheninnikov, A.I. Armer, V.V. Kuznetsov, E.Yu Lebedeva // Pattern Recognition and Image Analysis. "— 2011. "— Vol. 21, No. 2. "—Pp. 192–194. (in Russian)
- [16] Krasheninnikov, V. Variatsiya granits rechevykh komand dlya uluchsheniya raspoznavaniya rechevykh komand po ikh krosskorrelyatsionnym portretam [Voice command variability for voice command recognition accuracy by means of their cross-correlation portraits] / V.R. Krasheninnikov, E.Yu. Lebedeva, V.K. Kapyrin // Izvestiya Samarskogo nauchnogo tsentra RAN. "—2013. "— Vol. 4(4). "—Pp. 928–930. (in Russian)
- [17] Krasheninnikov, V. Povyshenie veroyatnosti pravil'nogo raspoznavaniya signalov po ikh krosskorrelyatsionnym portretam [Improvement of signal recognition accuracy by means of their cross-correlation portraits] / V.R. Krasheninnikov, N.A. Krasheninnikova, E.Yu. Galitskaya // Radiotekhnika. "—2014. "—Vol. 7. "—Pp. 107–110. (in Russian)
- [18] Vasil'ev, K. Statisticheskii analiz izobrazheniy [Statistical image analysis] / K.K. Vasil'ev, V.R. Krasheninnikov. "— Ulyanovsk, Russia. : UISTU, 2014. "—P. 216. (in Russian)
- [19] Armer, A. Ispol'zovanie ontologii dlya formirovaniya naborov etalonov rechevykh komand v zadache raspoznavaniya rechevykh komand na fone шумов [Using ontologies to generate a set of voice commands in the problem of speech recognition of voice commands in background noise] / A.I. Armer, V.S. Moshkin // Radiotekhnika. "—2016. "—Vol. 9. "—Pp. 72–77. (in Russian)
- [20] Armer, A. Podkhod k formirovaniyu naborov etalonov rechevykh komand s ispol'zovaniem ontologii [Formation of voice command model groups with ontology] / A.I. Armer, V.S. Moshkin // Ontologiya proektirovaniya. "—2016. "— Vol. 6. "—Pp. 270–277. (in Russian)
- [21] Krasheninnikov, V. Optimization of dictionary and model library for recognition of speech commands / V.R. Krasheninnikov, N.A. Krasheninnikova, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. "—2011. "— Vol. 21, No. 3. "—Pp. 505–507.
- [22] Krasheninnikov, V. Optimization of dictionary and model library for recognition of speech commands based on cross-correlation portraits / V.R. Krasheninnikov, N.A. Krasheninnikova, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. "—2013. "— Vol. 23, No. 1. "—Pp. 80–86.

# Development the algorithm of positioning industrial wares in-plant based on radio frequency identification for the products tracking systems

A.V. Astafiev<sup>1</sup>, A.A. Orlov<sup>1</sup>, D.P. Popov<sup>1</sup>

<sup>1</sup>Murom Institute of Vladimir State University, Orlovskaya, 23, 602264, Murom, Russia

---

## Abstract

As the title implies the article describes actuality of algorithm development of positioning industrial wares in-plant based on radio frequency grid for the construction of the products tracking systems. Requirements of international standards regulating the processes of traceability and identification are analysed. The article offers a system hardware solution for positioning of industrial wares in-plant based on radio frequency grid as well as an algorithm for determining the current storage area. Experimental studies of the developed algorithm were conducted.

*Keywords:* positioning; traceability; radio frequency identification; RFID

---

## 1. Introduction

Is the identification mechanism which provides the traceability of products during whole technological cycle of production. According to MN ISO 9001-87 requirements a supplier, if necessary, must set and support a method of product identification on all stages of production [1]. Traceability in production helps to provide compliance of requirements of government and international standards of quality, execute a rapid and targeted track of products during the technological cycle that, in turn, allows minimizing financial consequences. Especially important the question of tracking of products becomes if the technological cycle consists of large number of stages located on large territorial areas.

In the last few years instead of graphic marking and systems of technical vision systems companies prefer to use the radio frequency identification method. Currently radio frequency identification is one of the best information technologies used for constructing inventory control systems. Radio frequency identification is used for accounting tasks in different areas of activity, for example in logistics, libraries, shops, etc. However, the task of development and deployment of a complete system for product tracking in production still remains unsolved. Based on this one can conclude that development of new algorithms for identification and positioning of industrial wares in-plant based on radio frequency grid for the construction of the systems of product tracking is an actual scientific and technical task.

## 2. Setting the production requirements for the process of radio frequency identification in the products tracking systems

Let us consider base concepts of this area. Radio frequency identification [radio frequency identification; RFID] is technology of automatic identification and capture of data that uses electromagnetic or inductive connection carried out by means of radio waves for interaction with a radio frequency mark and an unambiguous read-out of its identification data by applying different types of signal modulation and data encoding. Interrogation – interaction of reader / survey device with an RFID tag to read data from it. Backscatter – the process that an RF tag uses to respond to the signal and to react to the electromagnetic field of a reading / interrogation device by modulating and re-radiating it, without changing the carrier frequency. [2]

As the operating frequency of the RFID-tag and the system there are the following ranges: low frequency (LF) - 125-134 kHz, high frequency (HF) – 13.56 MHz, ultra high frequency (UHF) - 860-960 MHz, microwave (SHF) – 2.4 GHz. Accordingly, for each range there is a corresponding standard, which specifies requirements for it. For example, the general requirements for the air interface for on 860-960 MHz frequency band can be found in the standard [3].

There are a number of standards which establish the structure of RFID tags. Standard [4] reviews unique radio frequency tags that are used for the purposes of: quality control of integrated circuits, which are used in RF tag manufacturing process; RFID traceability during their production and during their term of service; completing the process for reading information of RFID system configuration, including multiple antennas; implementation of anti-collision algorithm for inventory plurality of RFID tags, while in the zone of a reading / interrogation device; traceability of the object with the RF tag on. The standard [5] lists the requirements for the selection of RFID tags, as well as other data carriers, adhesive, face of the label material and ink. This standard specifies methods for reducing the influence of electrostatic discharge and damage to the RFID tag, as well as methods of data verification of the RFID tag. The Standard lists RFID placement and attachment rules.

It is important to point out that RFID-technologies can be subject to an attack. The most common attacks are: RFID-Zapper, cloning, Dos-attacks, attack via other RFID-tags, substitution of RFID-tags memory contents. For the protection of RFID systems experts give the following recommendations: while creating new software publish the code to third-party developers who for a fee can help find bugs, admitted in the development, and remove unnecessary functions [5].

During the system analysis of interstate and international standards requirements to the process of RF identification of products were established. Thus, to develop a radio frequency identification algorithm we must:

1. Define the task of identifying and selecting the appropriate method.
2. Develop a model and to consider all the requirements for it.
3. Determine which keywords should be used, their parameters and the range of operating frequencies.

4. To protect the system from the attacks of various kinds.

### 3. Subject overview

A great contribution to the development of radio frequency identification technology and product movement control systems in various spheres of life was made by Bondarevsky A.S., Zolotov R.V., Do Zuy Nyat, Kamozi D.U., Manish B., Shahram M., Ke-Sheng Wang, Worapot Jakkhupan, Somjit Arch-int, Yuefeng Li, Mahir Oner, Alp Ustundag, Aysenur Budak and many others.

Application of these knowledge-intensive technologies makes it possible to automate the processes of controlling the product movement at industrial plants, ultimately, to increase the efficiency and reliability of transportation control and warehouse inventory control of manufactured products.

However, they are not without flaws. The use of existing software and hardware solutions is more aimed at organizing automated warehouse inventory control and less suitable for automating product movement control, in the absence of universal methods and algorithms. In support of this, at a number of industrial enterprises, developers of RFID systems attempted to organize traceability of products by automatic movement control based on radio frequency identification. As a result, it became clear that automatic control of the product movement is possible only in certain areas of the production process. Such areas are conveyor lines and transport tunnels, where the transportation of products is carried out along the permanently installed radio frequency identification equipment (RFID tunnels). In other production and warehouse areas, automatic control over the movement of products is impossible. This is due to the lack of universal methods and algorithms for product identification in the process of its transportation along unmapped routes.

Positioning of objects and people using information technology is quite a substantial task. These technologies can be used to solve social, industrial and other types of tasks. Currently, there are a large number of approaches to positioning using a large number of technologies, among them:

- Satellite navigation technologies (GPS, GLONASS);
- Local positioning technology (infrared and ultrasonic);
- Technology of technical vision;
- Radio-frequency technologies.

The use of satellite navigation technology and positioning are tightly integrated into our daily lives. They are used for navigation and transport tracking, monitoring and coordination of various kinds of events. The accuracy of positioning is 10-15 meters outdoors. Unfortunately, the application of this technology inside production facilities is almost impossible. An exception is the installation of expensive equipment for organizing GPS-positioning indoors, the unit of which can cover no more than 10 square meters, which is unacceptable for most industrial plants, whose sizes can be tens of kilometers.

Local positioning technologies are highly accurate - about 2 centimeters, but with a short range of 5-10 meters. With these attributes, they are used to achieve local accurate results and, in general, are used for flaw detection (analysis of welds, detection of chips, dents, etc.). The use of local positioning technology for small-scale mechanization is not economically effective and will lead to huge financial costs.

The use of vision technology for solving positioning problems is a relatively young concept. Currently, there are a huge number of methods and algorithms for solving localization and positioning problems, but their effectiveness depends a lot on meeting a large number of requirements, which include the quality of materials used for production of visual labels, cleanliness and lighting of premises, staff attentiveness, etc. Failure to comply with even one of the requirements can lead to a significant reduction in positioning accuracy or make it completely inoperative.

Radio-frequency technologies have found wide application in sales (organization of security in stores). Positioning based on radio frequency technologies can be divided into two categories: positioning on passive RFID tags (distance up to 5 meters) and active RFID beacons (distance up to 80 meters), but all of them are based on the principle that the moved object is marked with an RFID tag and the reading equipment is stationary. This approach allows to effectively automate production processes, where the product movement routes are strictly limited, for example, conveyor lines. However, for the positioning of chaotically moving small mechanization means, this approach will lead to a significant increase in the cost of the positioning system. Instead of a few readers they would need ten times as many.

Considering all the information stated above it is possible to draw a conclusion, that development of technology and software for the construction of positioning and control systems for small mechanization in industrial plants based on radio frequency identification methods is a substantial scientific and technical task.

The development of software and hardware for movement control systems is carried out by: PCT-Invent (Russia, Saint-Petersburg), AiTiProject (Russia, Moscow), Impinj (USA, Seattle), Motorola (USA, Morrisville), Nordic ID (Finland, Salo), FEIG (Germany, Weilburg).

Development of positioning systems based on radio frequency identification is carried out by the following scientific organizations:

- Human positioning systems, in particular patients in medical institutions: Shonan Institute of Technology (Japan, Fujisawa), Institute of Medicine (Kathmandu, Nepal), National Patient Safety Foundation (USA, Boston) and others.
- Systems for positioning moving non-metallic objects: East China Jiaotong University (China, Nanchang), Universiti Sains Malaysia (Malaysia, Nibong Tebal), University of Adelaide (Australia, Adelaide), Wellness Convergence Research Center (Korea, Daegu) and many others.

However, the tasks of developing and implementing automatic systems for tracking products in production are still unresolved. Currently, industrial enterprises still have a number of problems, the solution of which is not realized with the help of modern product movement control systems.

#### 4. Development of the project hardware positioning system for wares in-plant based on the basis of radio-frequency grid

As the basic data for the implementation of functional tracking of industrial products arises in the course of its movement through the territory of the plant, it is advisable to develop a hardware solution for receiving and processing of the data. The main types of traffic information are the information about who moves the products, how, departure point and point of arrival. Receiving and processing of this information will allow to organize a permanent automatic traceability of industrial products in the plant.

The paper proposes the development of a stand-alone device, consisting of reading equipment, processing and transmission of information. The developed device is mounted on the transport device, and to ensure that the product gets delivered, plant territory is marked with RFID tags creating a radio frequency grid.

Thus, the hardware part of the system can be divided into 5 levels:

1. The RF tag for labeling storage areas.
2. Equipment to read RFID tags.
3. Equipment for collecting and processing statistical data.
4. The equipment for the transmission of data to the enterprise server.
5. Company's software and hardware.

Laboratory prototype was developed for testing the project hardware industrial products positioning system on the territory of the enterprise on the basis of radio-frequency grid for experimental studies. Laboratory prototype consists of a microcontroller, a manual RF reader, power supply unit and a laptop (Figure 1).



Fig. 1. Type of laboratory prototype.

#### 5. The algorithm for determining the current storage area

In order to determine the current position of the transport device for continuous automatic monitoring of transported goods has been developed an algorithm to determine the current storage area.

The algorithm is based on statistical analysis of the number of recognitions of radio frequency tags for certain time periods  $t$ . The period of time  $t$  is the average period of time during which the transport device is moved from the beginning of the current scan area to the end (Figure 2).

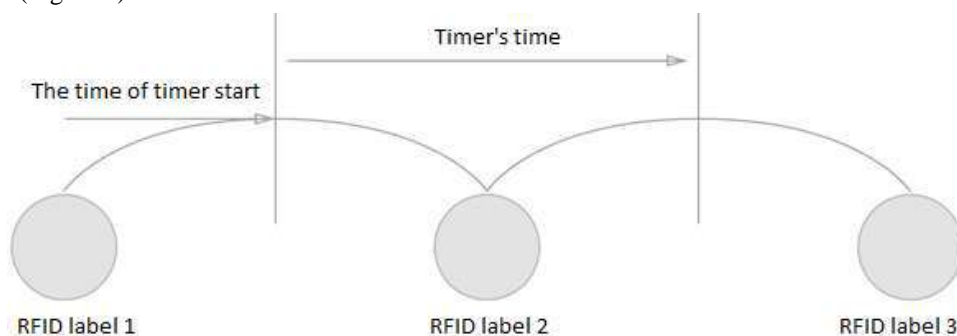


Fig. 2. Is a chart of determination of temporal interval of  $t$ .

Let the identifiers of warehousing zones be presented as a vector of  $I$  :



$$I = (I_1, I_2, I_3, \dots, I_n)$$

The amount of recognitions of radio frequency identifiers for the moment of time of  $t$  is presented in a kind:

$$C = (C_1(t), C_2(t), C_3(t), \dots, C_n(t))$$

Then determination of current position takes place by the calculation of index of  $k$  using a formula:

$$C_k(t) = \begin{cases} \max_i C_i(t), & \text{if } C_i(t) > p \\ \text{undefined,} & \text{otherwise} \end{cases}$$

$C_k(t)$  is the maximal amount of recognitions of radio frequency  $k$ -th identifier. It means that in current moment “ $t$ ” a transporting device is above the zone of warehousing market “ $k$ ”.

## 6. The experimental results of algorithm to determine the current storage area

During the experimental studies many different types of situations that are close to production were modeled (Figure 3). Among them:

- movement between two storage areas;
- move between three or more storage areas;
- movement between storage areas with the presence of "noise" (the other RFID tags, which are not labeled storage areas)
- movement between storage areas with partial overlap with non-metallic and metallic barriers.

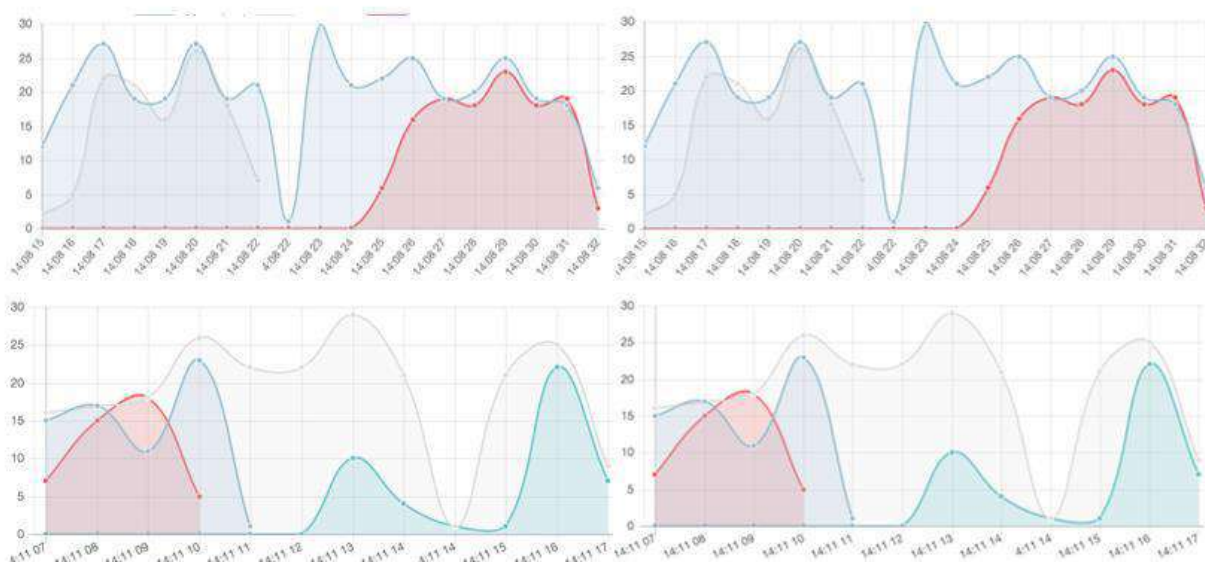


Fig. 3. Results of experimental studies.

Experimental studies have shown the correctness of the algorithm to determine the current storage area in the laboratory.

## 7. Conclusion

The article showed relevance of developing an algorithm of positioning industrial wares in-plant based on radio frequency grid for to create the products tracking systems. We analysed the requirements of international standards regulating the processes of traceability and identification. The article offered a system hardware solution for positioning of industrial wares in-plant based on radio frequency grid as well as an algorithm for determining the current storage area. Experimental studies of the developed algorithm were conducted.

## Acknowledgements

Acknowledgements and Reference heading should be left justified, bold, with the first letter capitalized but have no numbers. Text below continues as normal.

## References

- [1] ISO 8402:1994 Quality management and quality assurance – Vocabulary. URL: <http://www.docload.ru/Basesdoc/5/5812/index.htm> .
- [2] GOST R ISO/IEC 19762-3-2011 Information technology. Technology automatic identification and data collection (AISD). The harmonized dictionary. Part 3. radio frequency identification (RFID), 2012-05-01; 20 p.
- [3] GOST R ISO/IEC 18000-6-2013 Information technology. Radio frequency identification for control items. Part 6. The parameters of the radio interface for the frequency range 860 - 960 MHz. General requirements, 2014-01-01; 20 p.
- [4] GOST R ISO/IEC 15963-2011 Information technology. Radio frequency identification for control items. Unique identification of RFID tags, 2012-01-01; 28 p.

- [5] GOST R 54621-2011 Information technology. Radio frequency identification for control items. Recommendations for use. Part 1. Labels and packaging radio-frequency tag according to ISO/IEC 18000-6 (type C), 2012-06-01; 71p.
- [6] Provotorov A, Privezentsev D, Astafiev A. Development of Methods for Determining the Locations of Large Industrial Goods During Transportation on the Basis of RFID. *Procedia Engineering* 2015; 129: 1005–1009. DOI: 10.1016/j.proeng.2015.12.163.
- [7] Astafiev AV, Orlov AA, Provotorov AV. The method of combining the results of localization algorithms for character and bar code labels. *Stability and Control Processes in Memory of V.I. Zubov (SCP)*. International Conference, St. Petersburg, 5-9 Oct. 2015: 618–619. DOI: 10.1109/SCP.2015.7342240.
- [8] Astafiev AV, Orlov AA, Provotorov AV. The localization algorithm of symbolic and bar-code labels on industrial products for the control of product movements. *Stability and Control Processes in Memory of V.I. Zubov (SCP)*. International Conference, St. Petersburg, 5-9 Oct. 2015: 615–616. DOI: 10.1109/SCP.2015.7342230
- [9] Orlov A, Astafiev A. Development of algorithm for localization of production markings with the use of analysis of the color data on digital images. *Geoconference on informatics, geoinformatics and remote sensing conference proceedings*. Albena, Bulgaria 2014; 1: 113–118.
- [10] Orlov AA, Antonov LV, Astafiev AV. Development and experimental investigation of algorithms for distinguishing fuzzy boundaries of objects in photographs of industrial materials. *Pattern Recognition and Image Analysis* 2015; 25 (3): 509–513.
- [11] Terekhin AV. The Algorithm for Generating Pairs of Projections of Three-Dimensional Objects on Two Images. *J. Applied Mechanics and Materials* 2015; 770: 604–607.
- [12] Antonov LV, Orlov AA. Document Research and development of algorithms for objects detection on images of industrial materials. *Proceedings of 2014 International Conference on Mechanical Engineering, Automation and Control Systems, MEACS 2014* .
- [13] Muhammad Shahzad, Liu AX. Identification of Active RFID Tags with Statistically Guaranteed Fairness. *IEEE 23rd International Conference on Network Protocols (ICNP)* 2015: 279–290. DOI: 10.1109/ICNP.2015.23.
- [14] Muhammad Shahzad, Liu AX. Probabilistic Optimal Tree Hopping for RFID Identification. *IEEE/ACM Transactions on Networking* 2015: 796–809. DOI: 10.1109/TNET.2014.2308873.

# The reliability of pattern-match searching for the fragment on image using set of pseudo-gradient procedures

L.Sh. Biktimirov<sup>1</sup>, A.G. Tashlinskii<sup>1</sup>

<sup>1</sup>Ulyanovsk State Technical University, ul. Severnyi Venets 32, 432027, Ulyanovsk, Russia

---

## Abstract

The effectiveness of pattern-match searching for the fragment on image using set of pseudo-gradient procedures covering all initial image by their workplaces is studied. Procedures control is managed to reduce computational expenses and based on analysis of penalty function and giving priority of making next iteration to procedure that has minimum value of penalty. It is considered that required fragment belongs to the domain that has a procedure which is reached prescribed limit of iterations. If it is prior unknown if there is any required fragments on the initial image, the hypothesis of their absence should be tested. If the hypothesis is not confirmed than the scan for domains with fragments should be performed. Concerning there are fragments on the image the missing probabilities are found using penalty function and limit of iterations. Herein, the proposal is considered for single and multiple required fragments.

*Keywords:* digital image; fragment searching; fragment missing; pseudo-gradient procedure; probability; first and second type errors

---

## 1. Introduction

Pattern search of the single or multiple equal fragments belongs to the field of digital image processing [1-4]. There is type of search algorithms based on pseudo-gradient procedures (PGP) [5,6]. But PGP has relatively small working range [7], so it is necessary to split the high-resolution images to a multiple domains, with its own procedures. At this time appears a task to choose domains containing required fragments.

There is a missing fragment probability assigned to a search process. If all procedures are in equal conditions and make equal number of iterations than missing fragment probability depends on the preselected criteria of choice. It can be the best value of goal function of quality estimating on the last iteration. But this method provides low choice veracity cause goal function estimations calculates based on small-size local sample. Besides, number of search domains can be up to tens of thousands [8,9], so, to make all procedures achieve the prescribed limit of iterations, needs huge computational cost. To increase the probability it should be used more reliable criteria of choice, the same as using estimations of PGP on last iteration, for example, maximum of correlation index between required fragment template and its probable location on the image [10]. But this causes even more computational costs.

To reduce computational costs in [11] the algorithm for manipulating of set of PGP's is proposed where on the each step the priority to make next iteration to procedure that has best value of some penalty function (PF)  $X$  [12]. In this case step of the algorithm means complex of the operations: making ordinary iteration by procedure with best PF value, calculating new PF value and finding procedure with best PF value. The domain with procedure that firstly made prescribed number of iterations  $T$  is being considered as a required one (domain that probably is containing required fragment). To search  $k > 1$  fragments it should be chosen  $k$  domains containing procedures that achieved limit of iterations faster than others. Further, probability of wrong domain choice will be considered as probability of error caused by mentioned method of manipulating ensemble of procedures.

Generally, it is unknown, if there are any required fragments in domains under investigation. Thus, search procedure must contain some kind of testing the hypothesis of absence of fragments. During testing process there can be first  $P^{(1)}$  and second  $P^{(2)}$  type errors. So, if prior probability of location of the fragment among considering domains, is equal to  $P_F$  than decision about presence of the fragment is accepted with probability

$$P_F(1 - P^{(2)}) + (1 - P_F)P^{(1)}$$

and about absence – with probability

$$P_F P^{(2)} + (1 - P_F)(1 - P^{(1)})$$

Work [13] considers issue of testing hypothesis of fragment's absence. If the hypothesis rejected than next step is to determine the position of the fragment. Here, the probability of choosing domain with fragment after declining hypothesis with prescribed second type error probability:

$$P = P_F P_{ER} + (1 - P_F)P^{(2)}, \quad (1)$$

where  $P_{ER}$  is relative probability of wrong fragment choice in case it really is on the image.

Let us consider the probability  $P_{ER}$  of wrong choice of image domain with fragment in case it is on the image. Also the probability of making an error selecting  $k > 1$  domains when there are  $k$  identical fragments on the image, for example, images of equal objects (biological or technical) will be considered. Taking into account differences in search processes for single or multiple fragments these processes would be investigated separately.

## 2. Error probability in case of searching for a single fragment

Let image (or just part of it) where should be found the fragment is divided to  $N$  domains and there is just one of them to contain required fragment. Let's find error probability of that domain identification  $P_{ER}$ . Assume that goal function and PF of PGP are pre-defined. Let us call «  $x^+$  » value of PF  $X$  for the procedure in domain with fragment and «  $x^-$  » in domain without fragment.

If there are only two domains then domain without fragment will be chosen if its procedure will be first to make  $T$  iterations, i.e.  $x_T^- < x^+$ , where  $x_T^-$  is equal to PF value on  $T$ th iteration. Here, second procedure may perform from 1 to  $(T-1)$  iterations. Then, if value of PF on  $T$ th iteration is equal to  $x_0$ , to assume that choice is wrong it should be two conditions simultaneously: PF of the procedure in domain with required fragment exceed  $x_0$  and in domain without fragment PF value should be equal to  $x_0$ . Proposing that these events are independent the probability of wrong choice will equal to:

$$P_{ER} = \int_{x_0}^{\infty} w(x_T^+) dx \int_0^{x_0} w(x_T^-) dx.$$

But value of  $x_0$  is prior unknown and wrong choice probability generally is equal to probability that on  $T$ th iteration  $x_T^+ > x_T^-$

$$P_{ER} = \int_0^{\infty} w(x_T^-) (1 - F(x_T^+)) dx = \int_0^{\infty} w(x_T^+) F(x_T^-) dx, \quad (2)$$

where  $F(x_T) = \int_0^x w(x_T) dx$  is integrated distribution function.

It should be noticed that densities of distribution  $w(x^+)$  and  $w(x^-)$  are depending on initial approximation of search parameters [14] estimated by procedure and in this context are relative. Proposing that initial parameters' approximation for procedure in fragment's domain gets worst convergence in work range of procedure  $P_{ER}$  will be the upper limit of wrong fragment choice's probability.

If number of separated domains is equal to  $N$ , than assuming independence of procedure's PF (taking into account (2)):

$$P_{ER} = \int_0^{\infty} w(x_T^+) \left(1 - (1 - F(x_T^-))^{N-1}\right) dx. \quad (3)$$

The assumed restriction about PF independence is not strict cause samples from domains that don't have a fragment have weak correlation with samples from fragment.

## 3. Error probability in case of searching for multiple fragments

In previous case the required domain was assumed domain with procedure achieved limit of  $T$  iteration first of all others. Here, to provide low error probability of fragment's search with high signal/noise ratio it is necessary to specify large number of iterations. It is possible to decrease error probability with low  $T$  choosing several domains where procedures made limit of the iterations. Then probability of occurrence of domain with fragment among chosen domains increases. But it is not true for probability of right choice of the fragment.

There are criteria allowing to identify the fragment with low error probability, such as above mentioned maximum of correlation index that can be calculated on whole image, or extremes of information-theoretical measures of images similarity [15]. But using such criteria causes large computational expenses. Moreover, if image is divided into a lot of search domains and computing resources are strictly limited using of these criteria is not reasonable. But for small number of domains (for example, two) using these criteria is acceptable. On this basis, the probability of location the fragment among  $n$  domains where corresponding procedures firstly made prescribed limit of iterations.

If local samples to estimate all procedure's goal function value are independent and suppose best value of PF has domain without fragment a random event, than task may be reduced to Bernoulli scheme. So for probability  $P_{ER}^{(n)}$  of missing domain with fragment during choice  $n$  domains with procedures first reached prescribed limit of iterations using binomial law it can be written:

$$P_{ER}^{(n)} = \int_0^{\infty} w(x_T^+) \sum_{i=n}^{N-1} C_i^{N-1} F(x_T^-) (1 - F(x_T^-))^{N-i-1} dx, \quad (4)$$

Where  $C_i^{N-1}$  is a number of combinations from  $(N-1)$  of  $i$  elements. It should be noticed that cause number of investigating domains in general is less than general number of domains of separated image ( $n \ll N$ ) so it is reasonable to use next expression:

$$P_{ER}^{(n)} = \int_0^{\infty} w(x_T^+) \left( 1 - \sum_{i=0}^{n-1} C_i^{N-1} F(x_T^-) (1 - F(x_T^-))^{N-i-1} \right) dx. \quad (5)$$

Fig. 1 shows an example of graphs of dependency for probability of missing required fragment  $P_{ER}^{(n)}$  on the number of iteration when choosing one (solid-line curve) and two (dashed-line curve) domains. Calculation carried out for relay-type PGP with working range requiring to split two different-size images on 36 (curve 1 and 2) and 625 domains (curve 3 and 4). Initial parameters of mismatch were equal to 6 steps of parallel shift and 20 degrees of turn.

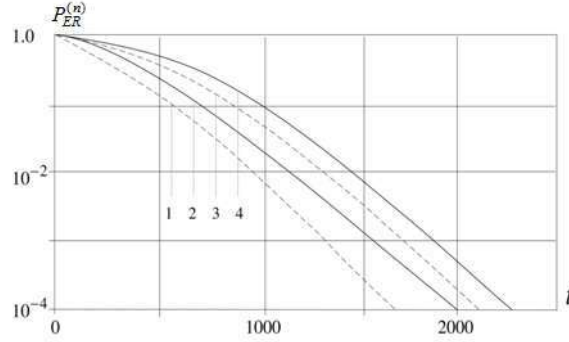


Fig. 1. Probability of missing required fragment against the number of iteration when choosing one and two domains.

During calculation was used expression (5) where meaning of expression:

$$\sum_{i=0}^{n-1} C_i^{N-1} F(x_T^-) (1 - F(x_T^-))^{N-i-1}$$

was computed as a ratio of incomplete  $B_q(N-n, n) = \int_0^q x^{N-n-1} (1-x)^{n-1} dx$  and complete  $B(N-n, n) = \int_0^1 x^{N-n-1} (1-x)^{n-1} dx$  beta-functions [16], where  $q = 1 - F_{PT}(\psi)$ . Hence while representing complete beta-function through gamma-function [17]

$$B(N-n, n) = \frac{\Gamma(N-n)\Gamma(n)}{\Gamma(N)},$$

will get

$$P_{N-1}\{0 \leq i \leq n-1\} = \frac{\Gamma(N) \int_0^q x^{N-n-1} (1-x)^{n-1} dx}{\Gamma(N-n)\Gamma(n)}$$

It is obvious from graphs that if  $n = 2$  probability  $P_{ER}^{(n)}$  is essentially low. For example, if  $N = 36$  and  $T = 1000$  probability of error in choice depending to situation  $n = 1$  is decreasing by 5.5 times, and if  $T = 2000$  – by 9 times. In this case, of course, computational expenses are increasing too.

#### 4. Error probability in case of searching for several equal fragments

Let's consider probability of missing at least one of domains with fragments location during search position of  $k > 1$  similar fragments. In this case it should be at a minimum  $k$  procedures to make limit number of  $T$  iterations. Same as before, it will be considered that presence of the fragments is known in advance.

In particular, in case  $n = k$  similar to (3) probability of all  $k$  procedures first made  $T$  iterations will correspond to domains containing fragments:

$$P^{(k)} = 1 - P_{ER}^{(k)} = 1 - \int_0^{\infty} w_k(x_T^+) \left( 1 - (1 - F(x_T^-))^{N-1} \right) dx, \quad (6)$$

where  $w_k(x_T^-)$  means probability density function for maximum of  $k$  PF's values of procedures from domains with fragments.

To decrease probability of missing fragments the number of domains to choose can be more than the number of required fragments ( $n > k$ ). In this case probability  $P_{ER}^{(j,k)}$  of missing  $j$  domains with fragments from  $k$  considering ratio (4) and uniqueness condition for each domain is equal to:

$$P_{ER}^{(j,k)} = \int_0^{\infty} w_k(x_T^+) \left( 1 - \sum_{i=0}^n C_i^{N-k} F(x_T^-) (1 - F(x_T^-))^{N-k-i} \right) dx, \quad (7)$$

where  $w_k(x_T^\pm) = C_j^K w(x_T^-) \frac{(1 - F(x_T^\pm))^{j-1} F^{k-j-1}(x_T^\pm)}{jk - j^2} (j - kF(x_T^\pm))$  is probability density function of  $j$  th ranged by maximum PF  $k$  procedures in domains with fragments  $i = \overline{1, k}$ .

## 5. Conclusion

One of the class of the algorithms to search fragment on the image by template is based on the PGP. But these procedures have relatively small working range of search that makes it necessary to split the image into array of domains each of them is containing own procedure. Here is a task about finding domains with required fragments. If all search procedures are working in same conditions and will make equal number of iterations than it require huge computational expenses. To reduce these expenses the algorithm of managing ensemble of PGP [12] can be used. In this case on the each step priority of making next iteration is giving to procedure that has best value of some penalty function. Domains with procedures made prescribed limit of iterations first are chosen as a required ones (probably containing fragments).

If it is prior unknown if there are required fragments on the image it is necessary to test the hypothesis about absence of fragments with prescribed error probabilities of first and second types. This issue and statistical criteria of hypothesis validity are considered in papers [13, 18]. If the hypothesis is rejected then choosing domains of fragments' location carried out. In this case error probability of choosing domains of fragment's location is a conditional probability and with prescribed second-type error probability in general determines by expression (1).

Probability of wrong choice assuming there are required fragments on the initial image depends on their count. If there is only one fragment and procedure then probability of wrong choice of domain with fragment determines by expression (3). To reduce error of right domain it can be chosen several domains instead one (Supposing using additional criteria to choose final domain among selected). Here, probability of presence of required domain among selected determines by expression (5).

If required fragments is more than one and each of them has corresponding one search procedure then probability of case when all procedures of domains with objects will make limited number of iterations first determines by expression (6). If number of choosing domains is greater than number of required fragments then probability of missing prescribed count of domains with fragments determines by expression (7).

## Acknowledgements

The study was carried out with financial support of the RFBR grant 16-01-00276.

## References

- [1] Ipatov YuA, Krevetsky AV. Modeling methods of detection and spatial localization of group point objects. Science. Technology. Production 2014; 2: 7–11.
- [2] Gerasimova NI, Verkhoturova AE. Search of the image fragment with application of Kohonen neural network. Information technologies in science, management, social sphere and medicine. Tomsk: TPU 2014; 1: 68–70.
- [3] Nikolenko AA, Babilunga OYu, Zaykovskij VN. Localization of specific image fragments based on two-dimensional wavelet filters. Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. Kharkov: NTU KhPI 2011; 36: 122–127.
- [4] Chambon S, Crouzil A. Dense matching using correlation: new measures that are robust near occlusions. British Machine Vision Conference, Norwich, Great Britain 2003; 1: 143–152. DOI: 10.5244/C.17.15.
- [5] Tsyarkin JZ. Information theory of identification. Moscow: Nauka; Fizmatlit, 1995; 336 p.
- [6] Zitova B, Flusser J. Image registration methods: a survey. Image and vision computing 2003; 21(11): 977–1000. DOI: 10.1016/S0262-8856(03)00137-9.
- [7] Tashlinskii AG. Estimation of parameters of spatial deformations of image sequences. Ulyanovsk, UISTU, 2000; 131 p.
- [8] Szeliski R. Image alignment and stitching: A tutorial. Foundations and Trends in Computer Graphics and Vision 2006; 2(1): 1–104. DOI: 10.1561/0600000009.
- [9] Tashlinskii AG. Pseudo-gradient Estimation of Digital Images Interframe Geometrical De-formations. Vision Systems: Segmentation & Pattern Recognition. Vienna, Austria: I Tech Education and Publishing 2007: 465–494. DOI: 10.5772/4975.
- [10] Pankova TL, Reznik AL. The effectiveness of algorithms for precision alignment of digital images . Optoelectronics, Instrumentation and Data Processing (Avtometriya) 1991; 5: 39–43.
- [11] Tashlinskii AG, Muratkhanov DS. Structural optimization of algorithms of parameter estimation of geometric image deforming. The physics and technical applications of wave processes 2001: 102–110.
- [12] Tashlinskii AG, Muratkhanov DS. Structural Optimization of pseudo-gradient Algorithms for Measuring Interframe Image Deformations. Pattern Recognition and Image Analysis 2003; 13(1): 177–178. DOI: 10.1134/S1054661806020088.
- [13] Biktimirov LSh, Tashlinskii AG. Estimating the probability of absence of target fragment on image for algorithm with control of multiple search procedures. Radioengineering 2016; 9: 6–10.
- [14] Tashlinskii AG, Kaveev IN, Voronov SV. Image registration method in conditions of intensive noise. Radioengineering 2012; 9: 45–49.
- [15] Voronov SV, Tashlinskii AG. Efficiency analysis of information theoretic measures in image registration. Pattern recognition and image analysis 2016; 26(3): 502–505. DOI: 10.1134/S1054661816030226.
- [16] Levin BR. Theoretical bases of statistical radio engineering. M.: Radio and communication, 1989; 656 p.
- [17] Koroljuk VS. A Handbook on Probability Theory and Mathematical Statistics. M.: Science, 1985; 640 p.
- [18] Biktimirov LSh, Tashlinskii AG. Criteria of testing the hypothesis of absence of target fragment on image. Modern problems of design, production and operation of radio engineering systems. Ulyanovsk: UISTU, 2016; 137–140.

# Development of informative neighborhood selection technology for modeling texture images

E. Biryukova<sup>1</sup>, R. Paringer<sup>1,2</sup>, A. Kupriyanov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

The paper proposes a method for constructing an informative neighborhood for modeling texture images. To describe the characteristic features of textures used assumptions underlying model representation texture images described by using a Markov random field. The results of the conducted experimental researches confirm that application of the developed approach allows to reduce the dimensionality of the features space while preserving the reliability of the classification.

*Keywords:* Markov random field; Gaussian Markov field; texture image classification; co-occurrence matrix; causal neighborhood

---

## 1. Introduction

Texture analysis widespread in the processing of various types of images. However, despite the fact that even in 1979 Haralick noted that the methods of distinguishing textures are developed individually for each specific case [1], there is no clear definition of texture or a particular concept in solving problems analysis of texture images.

Haindl wrote that the texture is a surface property, which is the spatial information contained in the object's surface [2].

The literature describes three approaches to texture analysis [1, 3, and 4]:

- A statistical approach, wherein the set of features used to provide texture image characteristics.
- Structural modeling allows us to consider texture as two-dimensional images composed of many primitives or subpatterns that are arranged accordance with a certain rule.
- Stochastic modeling suggests that the texture is the realization of a stochastic process that is characterized by certain parameters. This approach allows you to get good results for the generation of realistic natural texture images using Markov random fields [5].

To classification texture images, we will apply the model image as a realization of a random Markov field, that is, a stochastic approach to texture analysis. Great contribution to the development of this model has made by Haralick, who introduced the statistical and structural approaches to the description of texture [6] and suggested using of features based on the matrix of mutual probability distribution. The proposed is the gray level co-occurrence matrix [1]. It describes the spatial relationships of brightness pairs of texture elements.

## 2. Representation of the image according to the model of the Markov random field

Introduction of stochastic models and random fields models have led to the development of image reconstruction algorithms, segmentation, modeling and texture classification. In particular, Markov random fields is very useful for modeling spatial relationships, as well as for the study of stochastic interaction between the observed values, including the analysis of medical images and interpretation of remote sensing images [7].

The theory of Markov random field (MRF) provides a convenient and consistent method for modeling communication between dependent entities, such as image pixels and correlated features. Convenience is achieved due to the characteristic mutual influence among such objects, when using conditional distribution of MRF. The practical use of the model Markov random field obtained thanks to the theorem of equivalence between MRF and the Gibbs distribution, which was introduced by Hammersley and Clifford in 1971 [8]. This is because the joint distribution required for most applications, but the conclusion of the joint distribution of the conditional is very difficult for MRF. Equivalence theorem of Markov random fields and Gibbs points out that the joint distribution of MRF is the simplest form of the Gibbs distribution.

We will consider the model of a Gaussian Markov random field (GMRF), which is a particular case of MRF, where the value of the pixel in the position  $(i, j)$  is statistically independent of neighboring pixels. This means that the model takes into account the spatial interaction between the various components within each color component, and interaction of [9]. Image is represented on a rectangular lattice  $S = M * N$  with  $p$  number of bands.

Let  $X(i, j) = [x_1(i, j)x_2(i, j) \dots x_p(i, j)]$  is a vector in a texture region  $R$ . It is assumed that the vector at a position  $(i, j)$  represents the linear combination of the color components of neighboring pixels and additive Gaussian noise. Let  $\mu_1, \mu_2 \dots \mu_p$  denote the mean color intensity, and  $e_1, e_2 \dots e_p$  the spatial interaction of pixels and  $v_{xy}$  be the expected value of  $e_x e_y$ .  $x, y$  takes on the values from 1 to  $p$ . Let  $\phi_{xy}$  the associated parameters of the model and  $\sum$  the co-occurrence matrix.

Spatial interaction of color pixels is defined as:

$$\begin{aligned}
e_1(i, j) &= (x_1(i, j) - \mu_1) - \sum_{(m,n) \in N_{11}} \phi_{11}(m, n)(x_1(i + m, j + n) - \mu_1) \\
&- \sum_{(m,n) \in N_{12}} \phi_{12}(m, n)(x_2(i + m, j + n) - \mu_2) - \dots \\
&- \sum_{(m,n) \in N_{1p}} \phi_{1p}(m, n)(x_p(i + m, j + n) - \mu_p).
\end{aligned}$$

Similarly it is defined for  $e_2(i, j), e_3(i, j) \dots e_p(i, j)$ . The generalized form is given by:

$$\begin{aligned}
e_k(i, j) &= (x_k(i, j) - \mu_k) - \sum_{(m,n) \in N_{k1}} \phi_{k1}(m, n)(x_1(i + m, j + n) - \mu_1) \\
&- \sum_{(m,n) \in N_{p2}} \phi_{k2}(m, n)(x_2(i + m, j + n) - \mu_2) - \dots \\
&- \sum_{(m,n) \in N_{kp}} \phi_{kp}(m, n)(x_p(i + m, j + n) - \mu_p), k = \overline{1, p},
\end{aligned}$$

where  $N_{xy}$  denote neighboring pixels. If  $x=y$ , then the neighboring pixels will correspond to the same color component. Otherwise, the neighboring pixels are of the other components.

The co-occurrence matrix is defined as follows:

$$\sum = \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1p} \\ v_{21} & v_{22} & \dots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \dots & v_{pp} \end{pmatrix}.$$

The expected value  $v_{kl}$  is represented as:

$$v_{kl} = E[e_k e_l] = \frac{1}{M_R} \sum_{(i,j) \in R} e_k(i, j) e_l(i, j).$$

Having described all the terms, the probability density function of  $X(i, j)$  is found to be:

$$P(X(i, j)|R) = \frac{1}{((2\pi)^p |\Sigma|)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} (e_1(i, j) e_2(i, j) \dots e_p(i, j)) \sum (e_1(i, j) e_2(i, j) \dots e_p(i, j))^t\right\}.$$

### 3. Choice of informative neighborhood

Winkler in "Image Analysis, Random Fields and Dynamic Monte Carlo Methods" [10] writes the restoration images and modeling textures with random fields, in detail the finite random fields, including MRF applies Monte Carlo methods for Markov chains. Chohen for example textile fabrics control automation task [11] solves the problem of detection and localization of various kinds of defects, which uses Gaussian Markov random field and the non-causal neighborhood. Kovtun in [12] proposes a model image, a feature of which is that the segmentation and each texture are set independent random fields. His work is an attempt to highlight the problem of texture segmentation of the general class of problems of generation and modeling of Markov random fields.

Thus, in [3, 5, 7-12] is said about using the Markov random field model to describe and generate texture images. One of the parameters of the described model is the probability distribution of the brightness of neighboring pixels. In this case, the choice of the neighboring pixels, in works devoted to this subject, using non-causal neighborhood (Fig. 1).

The paper proposes a new method of selecting the informative neighborhood to describe the characteristics of the texture.

The following algorithm can represent description of the main stages of the technology of selecting the informative neighborhood:

1. Choosing raw data: the shape of the neighborhood, a set of features calculated from the surroundings and the separate images on the textural classes.
2. Calculation features of the selected neighborhood for each image. Form the initial sample.
3. Calculate individual separability criteria for each feature [13]. We assess informative features, based on the value criterion [14].
4. Excluded from the original sample with features of lower value separability criterion.
5. We exclude from the neighborhood of the pixels corresponding to the non-informative characters.

Thus, the remaining pixels constitute informative neighborhood.

Experimental technology study carried out based on texture images "Kylberg Texture Dataset v. 1.0"[15]. Consider the application of technology to the two classes of images and rice1 rice2 selected database. Figure 2 shows examples of the images under consideration.



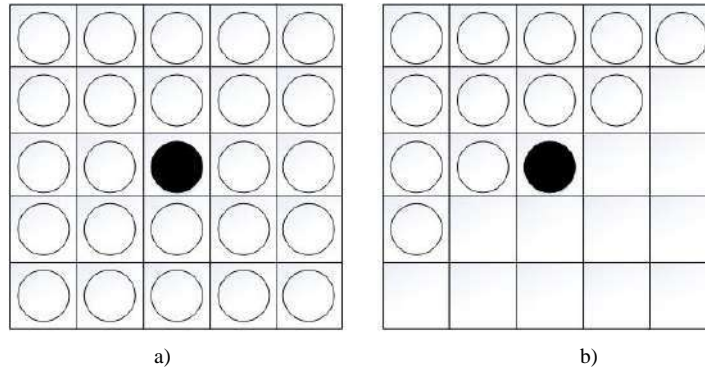


Fig. 1. Example of the surrounding area: (a) the non-causal, (b) causal.

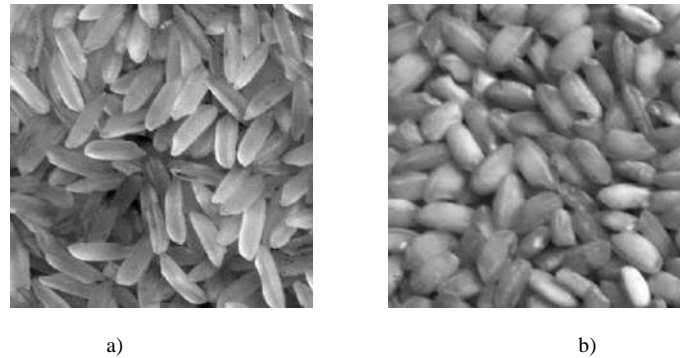


Fig. 2. Example images: (a) - rice1, (b) - rice2.

To distinguish the classes of texture images, we used statistical features calculated by the formula:

$$\lambda(\Delta x, \Delta y, n) = \frac{\sum |f(x, y) - f(x + \Delta x, y + \Delta y)|^n}{N}$$

where  $f$  is the image intensity function,  $N$  is the number of image pixels. In the following research we used the features  $\lambda$ , calculated at  $\Delta x, \Delta y = 0, \pm 1, \pm 2, n = 1, 2, 3$ . Because the features are symmetrical used causal neighborhood.

Individual criteria of separability for each feature were calculated (Figure 3).

For a sample consisting of  $n$  elements, divided into classes  $g$  and comprising a  $p$  features separability criterion is calculated using the following formulas:

$$J = tr((T)^{-1}B),$$

where  $T = B + W$ .

$B$  – is the intergroup dispersion matrix. The elements of this matrix are calculated according to the formula:

$$b_{ij} = \sum_{k=1}^g n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j), i, j = \overline{1, p},$$

$W$  – is the intragroup dispersion matrix. The elements of the matrix are calculated according to the formula:

$$w_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk}), i, j = \overline{1, p},$$

$x_{ikm}$  – is the value of the  $i$ -th feature for the  $m$ -th element of  $k$  class,  $\bar{x}_{ik} = 1/n_k \sum_{m=1}^{n_k} \bar{x}_{ikm}$  – is the mean value of the  $i$ -th feature of  $k$  class,  $\bar{x}_i = 1/n \sum_{k=1}^g n_k \bar{x}_{ik}$  – is the mean value of the  $i$ -th feature in all the classes, and  $n_k$  is the number of elements in  $k$  class.

0.74	0.77	0.82	0.77	0.55
0.58	0.74	0.85	0.61	
0.58	0.61	●		
0.53				

Fig. 3. Mean value of the separability criterion.

The higher the value of the criterion is, the more the separability of the classes grows.

After calculating the individual criteria for separability, features with a low value criterion were excluded. Analysis of the feature space, led to the conclusion that some of the neighboring pixels carry information about the features of the texture (pixel information are highlighted in Figure 3). It was excluded from the neighborhood of the pixels corresponding to non-informative features (calculated at  $(\Delta x = 2, \Delta y = 2)$ ,  $(\Delta x = -2, \Delta y = 1)$ ,  $(\Delta x = 1, \Delta y = 1)$ ,  $(\Delta x = -2, \Delta y = 0)$ ,  $(\Delta x = -1, \Delta y = 0)$ ,  $(\Delta x = -2, \Delta y = -1)$ ). Thus, we resins are informative neighborhood new form. Modified neighborhood for the test classes is shown in Figure 4.

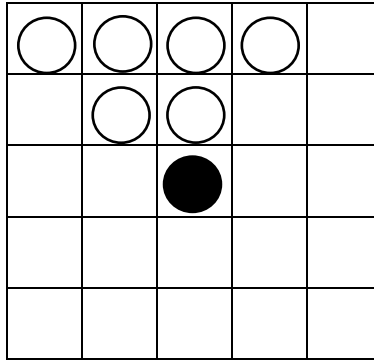


Fig. 4. Modified neighborhood.

To study the effectiveness of the technology was evaluated the quality of the selected neighborhood. The evaluation was conducted by calculating the clustering error on the based of k-means algorithm, where the centers of the starting classes used as initial conditions [16]. Under the error of clustering is understood the proportion of images that were not attributed to their class. Clustering error in the case of using features calculated by causal neighborhood was 0.21, the modified 0.19, which confirms the information content of the modified neighborhood.

Table 1 shows the values of the clustering error in the case of features, calculated using the causal neighborhood and modified to distinguish other classes of images from the selected base textures.

compared classes	causal neighborhood	modified neighborhood
blanket1, and canvas1	0.03	0.03
scarf1, and scarf2	0.18	0.16
Linseeds, and sesameseeds	0.46	0.40

As Table 1 shows that clustering error value using the modified neighborhood does not exceed the error value using a causal neighborhood, which indicates the information content received surroundings, and hence the effectiveness of the proposed technology.

Figure 5 shows the mean values of separability criteria for the cases considered in Table 1, the modified neighborhoods are highlighted in color.

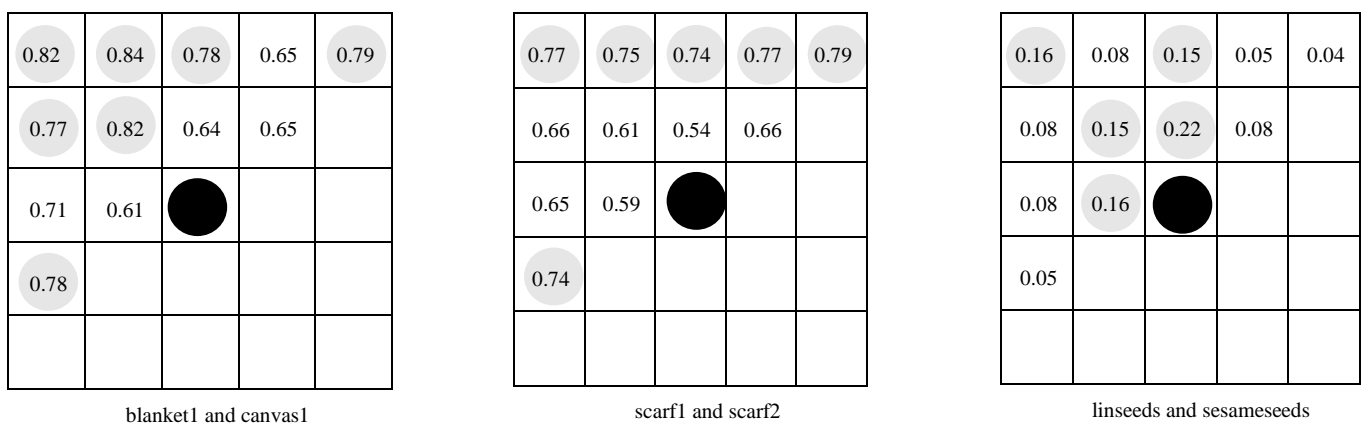


Fig. 5. The mean values of the separability criterion for different classes.

#### 4. Conclusion

The paper presents the technology of choice informative neighborhood, which has shown to be effective for the considered classes of texture images. The features space and the clustering error were reduced by reducing the number of neighboring pixels. The proposed technique can be used to modeling texture images, wherein for the calculation of the model parameters using a Markov random field neighborhood.

## Acknowledgements

This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world's leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29-03823, # 15-29-07077, # 16-41-630761, # 17-01-00972); by the ONIT RAS program # 6 "Bioinformatics, modern information technologies and mathematical methods in medicine" 2017.

## References

- [1] Haralick RM. Statistical and structural approaches to texture. *Proceedings of the IEEE* 1979; 67(5): 786–804.
- [2] Haindl M. Texture synthesis. *CWI Quarterly* 1991; 4: 305–331.
- [3] Dubes RC, Jain AK. Random field models in image analysis. *Journal of Applied Statistics* 1989; 16(2): 131–164.
- [4] Ahuja N, Rosenfeld A. Mosaic models for textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1981; PA MI 3: 1–10.
- [5] Plastinin AI. The method of forming features texture images based on Markov models. *Dis. kan. those. Sciences, Samara*, 2012.
- [6] Haralick RM. Statistical and structural approaches to the description of textures. *TIIRE* 1979; 5: 98–118.
- [7] Li SZ. *Markov Random Field in Image Analysis*. Springer-Verlag, 2009; 362 p.
- [8] Hammersley JM, Clifford P. *Markov field on finite graphs and lattices*, unpublished, 1971
- [9] Mridula J. *Feature Based Segmentation of Colour Textured Images using Markov Random Field Model*. Master of Technology, Odisha, India, 2011.
- [10] Winkler G. *Image Analysis, Random Fields and Dynamic Carlo Methods Monte*. Novosibirsk: Branch "Geo" Publishing RAS, 2002; 343 p.
- [11] Cohen FS, Fan Z, Attali S. Automated Inspection of Textile Fabrics Using Textural Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1991; 8(13): 803–808.
- [12] Kovtun IV. Texture image segmentation based on Markov random fields. *Control systems and machines* 2003; 4: 46–55.
- [13] Fukunaga K. *Introduction to statistical pattern recognition theory*. Moscow: Nauka, 1979; 270 p.
- [14] Biryukova E, Paringer R, Kupriyanov A. Development of the effective set of features construction technology for texture image classes discrimination. *CEUR Workshop Proceedings* 2016; 1638: 263–269. DOI: 10.18287/1613-0073-2016-1638-357-363.
- [15] Kylberg G. *Kylberg Texture Dataset v. 1.0*, 2014. URL: <http://www.cb.uu.se/~gustaf/texture>.
- [16] Mandel ID. *Cluster analysis*. Moscow: Finance and Statistics, 1988.

# Comparison of classification algorithms in the task of object recognition on radar images of the MSTAR base

A.A. Borodinov<sup>1</sup>, V.V. Myasnikov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

The present work is devoted to the analysis of local objects on radar images. In comparison, the following algorithms are used: decision tree; Bayesian classifier for normal distribution; Nearest neighbor method; Support Vector Method (SVM). As preliminary processing of images provided by a synthetic aperture radar. The research is carried out on the objects from the base of radar images MSTAR. The paper presents the results of the conducted studies.

*Keywords:* Classification of images; Synthetic aperture radar; Classification; Decision tree; C4.5; CART; SVM; MSTAR

---

## 1. Introduction

Radar satellite imagery obtained with synthetic aperture radars allows obtaining images of good quality in difficult weather conditions, as well as in cases of low illumination. A certain complexity in the processing of the obtained images is the speckle noise that is present on the radar images. The recognition of images on radar images is used in various fields, such as agriculture, forestry, relief analysis, oil spill monitoring and equipment recognition. Studies of various algorithms for the classification of radar images has been conducted previously, but often they had been compared the obtained data with the data from other articles. Such an approach may lead to inaccurate results of the analysis of the results obtained. Also, most articles use for comparison only the most popular classification algorithms, such as SVM, AdaBoost and neural networks. In this paper, the study adopted classifiers, which are used in works on this topic less often: decision trees, k nearest neighbors method, naive Bayesian classifier. The purpose of this paper is to fill this gap. All tests were conducted using the public database of radar images of MSTAR military equipment.

## 2. Statement of the classification problem

The task of object recognition on an image can be divided into two main subtasks:

- search for an object in the image and selection of areas of interest;
- recognition and classification of the found object or area of interest. [1]

The first subtask is aimed at finding objects for classification. Often information about the location, size, orientation, availability and number of goals is initially missing. In this case, it is necessary to determine the unknown parameters required for further selection of the object or local area of interest.

The second subtask is applied to the entire image and allows you to decide which of the several classes the image being processed belongs to. The goal of the classification is the construction of a decision function. The decision function for each feature vector relates the corresponding class. In this article, we consider only the classification problem.

In connection with the need to process a large number of images for training and testing, as well as low performance of some algorithms, there is a need to reduce the dimensionality of the feature space. There are various methods used to solve this problem. Such methods include the most popular ones: the method of principal components, factor analysis, the method of independent components, self-organizing maps of Kohonen and others. In this paper, the principal component method is applied.

## 3. Principal component analysis

The Principal Components Analysis (PCA) method is one of the most widely used methods for reducing the dimensionality of a feature space with the loss of the least amount of information. This method reduces to calculating the eigenvalues of the covariance matrix of the analyzed image. [2] Algorithms for calculating the covariance matrix operate in the line-by-line mode of reading the image, which allows achieving high performance and low requirements for RAM. [3]

## 4. Evaluation of classification results

To assess the results of the classification, the sliding control method is used. Sliding control (cross-validation, CV) is a statistical method for assessing the generalization of the quality of classification. It is a more reliable and thorough assessment method, compared to the usual sequential division of a data set into a training and test sample. With the sliding control, the data is repeatedly divided into training and test sets and fed to the classifier's input.

The paper uses a modified method of sliding control with multiple partitioning [4] in which the entire volume of data is divided into a specified number of parts of N (equal to 1). The number of iterations of learning in this algorithm corresponds to

the number of blocks  $N$ . There is also a stratification of classes and samples, allowing reducing the dispersion of estimates of sliding control. This leads to a decrease in the confidence interval and a more accurate classification quality. Applying class stratification makes it possible to break each class in a given ratio. At each iteration of the algorithm,  $K$  parts are randomly selected as the training sample and  $L$  parts as the test sample. This partition can be described as follows:

$$N \geq K + L, \Omega_O \cup \Omega_T = \Omega', \Omega' \subseteq \Omega, \Omega_O \cap \Omega_T = \emptyset,$$

$$\Omega_O = \bigcup_{i=0}^{K-1} \Omega_{O_i}, \Omega_T = \bigcup_{j=0}^{L-1} \Omega_{T_j}.$$

Where  $\Omega_O$  is a training sample,  $\Omega_T$  is a test sample,  $\Omega$  is the original sample.

For each partition obtained, the classifier is set up on the training sample and the quality value of the classifier is calculated on the test sample.

The functional quality of the algorithm on the sample has the following form:

$$CV(\mu, \Omega') = \frac{1}{N} \sum_{p=1}^N \frac{1}{N} \sum_{q=1}^N Q(\mu(\Omega' \setminus \Omega_{Opq}), \Omega_{Opq}).$$

$\mu$  is the learning method.

## 5. Algorithms of classification

### 5.1. Bayesian Classifier Gaussian Case

The classification method based on the naive Bayesian classifier is a learning algorithm with the teacher, in which the Bayes theorem is applied with a strict (naive) assumption of independence between each pair of characteristics [5]. The assumption of independence makes it possible to get rid of a complex scheme for evaluating the parameters of the classifier. This allows us to apply the algorithm to large samples. Also, the classification is quite accurate: insufficient for high-precision classification systems, but satisfactory for rough estimation and comparison with other algorithms. Proceeding from the Bayes theorem:

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(y|x_1, \dots, x_n)}{P(x_1, \dots, x_n)}$$

and the independence assumption, we obtain:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y),$$

that can be rewritten:

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)},$$

and get the resulting classifier-function  $\hat{y}$ :

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y).$$

and use the a posteriori maximum estimate to estimate  $P(y)$  and  $P(x_i|y)$ .

Naive Bayesian classifiers differ, mainly, by the assumptions they make about  $P(x_i|y)$ .

In this paper, the densities used are the Gaussian case, which is based on the use of the probability density of the form:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}},$$

where  $\mu_y$  and  $\sigma_y^2$  are the mathematical expectation and the correlation matrix.

### 5.2. KNeighbors

The nearest neighbor's algorithm  $k$  refers to metric classification algorithms with training sample  $\Omega_O$ . Such algorithms refer object  $u$  to that class  $y \in Y$ , for which the total weight of the nearest objects from the training sample is maximal:

$$a(u, \Omega_O) = \arg \max_{y \in Y} \Gamma_y(u, \Omega_O), \text{ и } \Gamma_y(u, \Omega_O) = \sum_{i=1}^K [y_u^{(i)} = y] \omega(i, u).$$

Where the weight function  $\omega(i, u)$  estimates the degree of importance of the  $i$ -th neighbor for the classification of the object  $u$ . The function  $\Gamma_y(u, \Omega_0)$  is an estimate of the closeness of the object  $u$  to the class  $y$ . The importance function is chosen to be non-negative and not increasing in  $i$ . The selection criteria are due to the fact that the smaller the distance between the sampled objects  $u$  and  $x_u^{(i)}$ , the greater the probability of a correct classification. In the algorithm  $k$  of the nearest neighbors, the object  $u$  is referred to a class with more elements among the  $k$  nearest neighbors  $x_u^{(i)}$ ,  $i = \overline{1, k}$ :

$$\omega(i, u) = [i \leq k] \omega_i, a(u, \Omega_0, k) = \arg \max_{y \in Y} \sum_{i=1}^k [y_u^{(i)} = y] \omega_i.$$

As a metric, the Euclidean metric is most often chosen because of its simplicity and comprehensibility. Three metrics are studied: Euclidean, Minkowski and Manhattan distance.

The Euclidean distance between two points  $x, y$  is defined in Euclidean  $n$ -dimensional space as:

$$r(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

The Manhattan distance is defined as the sum of the moduli of the coordinate differences:

$$r(x, y) = \|\vec{x} - \vec{y}\| = \sum_{i=1}^n |x_i - y_i|.$$

Another metric on the Euclidean space, which is investigated in the paper is the Minkowski metric. It can be regarded as a generalization of the Euclidean and Manhattan distances. For the parameter  $p = 2$ , the Minkowski distance is generalized to the Euclidean distance, and for  $p = \infty$  to the Chebyshev distance. This metric is defined by the following formula:

$$r(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}.$$

The drawbacks of metric algorithms include storage of the entire training sample.

### 5.3. Decision Tree (C4.5, CART)

A decision tree is a structure of a hierarchical type, in which branches a partition of the feature space is defined, and the sheets are elementary classification functions. There are various methods for constructing trees. In this paper, the algorithms C4.5 [6] and CART [7] will be considered.

C4.5, receiving the input sample  $\Omega_0$ , builds the source tree, based on the following rules. If all objects in the sample belong to the same class or the sample is small, then the tree is a sheet marked with the most common class in the sample. Otherwise, a split criterion is selected that divides the sample into two or more samples. Then the criterion is chosen for the obtained partitions. This procedure is recursively applied for each sample received. One of the criteria is used to minimize the entropy value of the obtained sample partitions. The resulting source tree is then trimmed to avoid retraining. Based on the received tree, a decision function is constructed for classifying objects.

In the CART algorithm, a binary decision tree is recursively constructed. The tree is created to the maximum size without using the stopping rule, and then it is clipped. The algorithm builds not one but a sequence of nested truncated trees. The best division is selected based on the sliding control. The partition criterion is based on the Gini index.

### 5.4. SVM

The support vector machine is one of the most reliable methods among all known algorithms and is most often used for comparison with new algorithms. The function separating the classes is a separating hyperplane. The algorithm maximizes the shortest distance between the points closest to the points on the hyperplane [8]. In this paper, the linear separating function and the radial basis function are used as the separating function.

## 6. Experimental research

All the experimental studies were conducted on a PC Intel Core i5-4460, 16 GB RAM. All classification algorithms were written in the programming language Python 3.6. Also used were frameworks and libraries scikit-learn, openCV, numpy. As objects of classification, samples of military equipment from the public database of radar images MSTAR, presented in Figure 1, served.

For recognition, the magnitudes of the images of BMP-2, BTR-60, BTR-70 and T-72 were used. As preprocessing of images, the orientation of objects on centered images was normalized and cropped from  $128 \times 128$  to  $60 \times 60$  pixels. The resulting images are shown in Figure 2.

The target shooting angle is 15 and 17 degrees. The initial sample consists of 3438 images of different classes of objects. The number of images of each class is shown in Table 1.

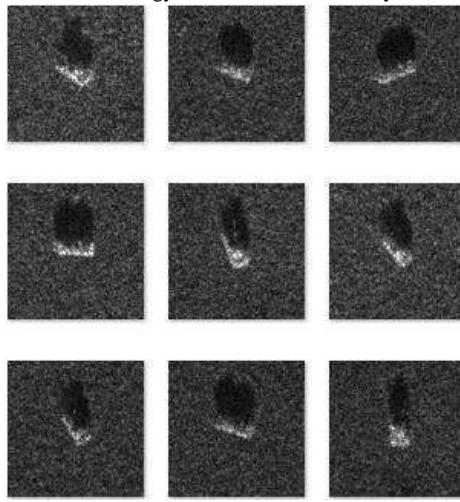


Fig. 1. Images of MSTAR objects.

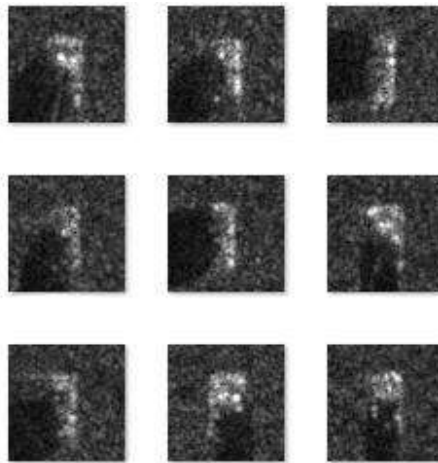


Fig. 2. Processed MSTAR images.

Table 1. Number of classification objects.

Object / Angle	15°	16°
BMP-2	587	698
BTR-60	195	256
BTR-70	196	233
T-72	582	691

For all images from the general sample, the dimension was reduced. The list of investigated classifiers is given in Table 2.

Table 2. Investigated classifiers.

GaussianNB	Naive Gaussian Bayesian classifier
KNeighbor_1	Nearest neighbor method, Euclidean metric
KNeighbor_2	The nearest-neighbor method, the Manhattan distance
KNeighbor_3	The nearest-neighbor method, the Minkowski distance
CART	The decision tree based on the CART algorithm
C4.5	The decision tree based on the C4.5 algorithm
SVM_1	The method of support vectors for a linear separating function
SVM_2	The method of support vectors for a radial basis function

The value of the classification quality will be calculated as the average relative number of correctly classified objects from the test sample  $\Omega_T$ . For the sliding control method, we specify the number of partitions and the number of iterations  $N = 10$ ,  $K = 6$  and  $L = 4$ , dividing the total sample in the ratio 6: 4. The method of the main components will reduce the dimension to 20 eigenvectors, retaining a significant part of the radar image information necessary for the classification of objects. Detailed classification results are presented in Table 3.

Table 3. Results of classification.

	BMP-2	BTR-60	BTR-70	T72	Avg.
GaussianNB	0.10895	0.62011	0.78899	0.74476	0.56570
KNeighbor_1	0.98156	0.95438	0.98680	0.97407	0.97420
KNeighbor_2	0.98586	0.95476	0.98721	0.97448	0.97558
KNeighbor_3	0.98152	0.96915	0.98481	0.97208	0.97689
CART	0.85964	0.78115	0.77543	0.86927	0.82137
C4.5	0.87424	0.75564	0.85257	0.87975	0.84055
SVM_1	0.92018	0.66914	0.89281	0.87513	0.83932
SVM_2	0.96351	0.96602	0.95939	0.99392	0.97071

The support vector machine with a radial basis function and the k nearest-neighbor method (with Minkowski distance) showed the best result of the classification of radar images.

## 6. Conclusion

It can be seen from the results of the conducted research that the best indicators of the classification of radar images of the MSTAR base are given by the method of the nearest neighbors, and by the support vector machine. In subsequent studies, it is planned to apply boosting algorithms, such as AdaBoost [9], and neural networks. Over the past few years, there have been many publications using neural networks [10] for the classification of radar images, so their study and comparison of classification results with the results obtained in this paper is of great interest.

## References

- [1] Gashnikov M, Glumov NI, Ilyasova NYu, Myasnikov VV, Popov SB, Sergeev VV, Soifer VA, Khramov AG, Chernov AV, Chernov VM, Chicheva MA, Fursov VA. Methods of computer image processing. Ed. Soifer VA. 2 nd ed. rev. Moscow: Fizmatlit, 2003; 784 p.
- [2] Method of the main components URL: [http://www.machinelearning.ru/wiki/index.php?title=Method\\_head\\_component](http://www.machinelearning.ru/wiki/index.php?title=Method_head_component) (01/26/2017).
- [3] Kuznetsov AV, Myasnikov VV. Comparison of algorithms for controlled element-by-element classification of hyperspectral images. *Computer Optics* 2014; 38(3): 495–502.
- [4] Vorontsov KV. Combinatorial approach to the evaluation of the quality of learning algorithms. *Mathematical problems in cybernetics*. Ed. Lupanov OB. Moscow: Fizmatlit, 2004; 13: 5–36.
- [5] Naive Bayes. URL: [http://scikit-learn.org/stable/modules/naive\\_bayes.html](http://scikit-learn.org/stable/modules/naive_bayes.html) (23.01.2017).
- [6] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers Inc., 1993; 302 p.
- [7] Wu X, Kumar V, Ross Quinlan J. Top 10 algorithms in data mining. *Knowledge and Information Systems* 2008; 14(1). DOI:10.1007/s10115-007-0114-2.
- [8] Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning* 1995; 20(3): 273–297.
- [9] Sun Y, Liu ZP, Todorovic S, Li J. Adaptive Boosting for SAR Automatic Target Recognition. *IEEE Trans. Aerosp. Electron. Syst.* 2007; 43(1): 112–125.
- [10] Profeta A, Rodriguez A, Clouse HS. Convolutional neural networks for synthetic aperture radar classification. *Algorithms for Synthetic Aperture Radar Imagery XXIII* 2016; 9843. DOI:10.1117/12.2225934.



# Spatio-temporal analysis through remote sensing and GIS in Moscow region, Russia

Komal Choudhary<sup>1</sup>, M.S. Boori<sup>1,2,4</sup>, A. Kupriyanov<sup>1,3</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>American Sentinel University, 2260 South Xanadu Way, Suite 310, Aurora, Colorado 80014, USA

<sup>3</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

<sup>4</sup>Bonn University, Meckenheimer Allee 166, D-53115 Bonn, Germany

---

## Abstract

Spatio-temporal analysis is a process for city development with growing population and economy for better implementation of planning policies with advance technology. In this research work, three dates (1995, 2005 & 2016) satellite images were used to mapping and monitoring of Moscow region, Russia. This study focuses on the further classification of the study area into different categories on the basis of use and association by implementing a rule-based classification system on remotely sensed data. This research provides useful and up-to-date information to local land use planners, managers and policy-makers to step up towards sustainable development in Moscow region, Russia.

*Keywords:* Spatio-temporal; land use/cover; remote sensing; GIS

---

## 1. Introduction

Planning is a widely established approach for managing resource and decision making. It includes the use of collective intelligence and knowledge of future requirements and need to improve environment in which people work and spend their leave time [1]. Hence studying the spatial and temporal LULC changes might provide a prominent basis for more effective land use planning that would keep the ecosystem in balance. At this time research on urban growth has become a very important factor for the interpretation of global environmental changes it has effects on the local environment and the economy growth can be defined as the spread of new developments in urban areas to the surrounding land [2]. Urban growth is responsive for the disorganized use of land resources and energy intrusion into agricultural land. Unplanned urban growth has been responsible for many problems such as poor quality of life, polluted drinking water noise pollution, air pollution etc. The combination of spatial data and analytical methods will provide support to city planners, ecologists and resource managers in their planning and decision making [3-4]. Dynamic spatial urban models provide an enhanced capability for evaluating future development and generating planning.

The technology of remote sensing and GIS includes both aerial and satellite based examination with high resolution and high temporal frequency [5-6]. In this research an attempt has been made to diction the spatio-temporal urban growth dynamics of the Moscow region. To achieve this Landsat satellite data from 1995, 2005 and 2016 for the month of February were analyzed for land use mapping. The urban expansion of Moscow over all 15 years period 1995-2016 was mapped using remote sensing and GIS images.

## 2. Study area

Moscow region is the one of the most densely populated regions in the country and is the second most populated federal region. The Oblast has no official administrative center, it is public authorities are located in Moscow and across other locations in the oblast. As of the 2010 Census, its population was 7,095,120 and 7,231,068 recorded in the 2015 Census. The latitude of the city is 55° 45' 7" N and longitude is 37° 36' 56" E. The region is highly industrialized, such as metallurgy, oil refining, mechanical engineering, food, energy and chemical industries [7].

The climate of Moscow region is humid continental, short but warm summers and long cold winters. The average temperature is 3.5 °C (38.3 °F) to 5.5 °C (41.9 °F). The coldest months are January and February average temperature of -9 °C (16 °F) in the west and -12 °C (10 °F) in the east. The minimum temperature is -54 °C (-65 °F). Here are more than three hundred rivers in Moscow region. The first largest river is Volga, most river belong to the basin of the Volga. Which itself only crosses a small part in the north of Moscow Oblast. They are mostly fed by melting snow and the flood falls on April-May. The water level is low in summer and increases only with heavy rain. The river freezes over from late November until April.

## 3. Material and methods

The Landsat program is a series of Earth-observing satellite mission jointly managed by NASA and the U.S. geological survey [8]. The first Landsat satellite was launched in 1972 and the most recent one Landsat 8 was launched on February 11, 2013. Data from Landsat 8 has eight spectral bands with spatial resolutions ranging from 15 to 60 m. The Landsat satellite data of 1995, 2005 and 2016 have been used in this study with a spatial resolution of 30 m. The satellite data were checked completely before classification into land use groups [9-10]. There are many techniques available for detecting and recording differences,

ratios and correlation. The data used in this paper were divided into two categories first satellite data and second ancillary data. Satellite data for the other hand consisted of multi- spectral data acquired by Landsat satellite provided by USGS gloves [11]. Ancillary data include ground truth data for the land use/cover classes and topographic maps. Spectral charts were prepared to distinguish and find out the difference in pixel values of different land use/cover classes in different bands. Primary land use classes were defined, such as agriculture, barren land, forest, settlements, scrubland, water body and wetland. The land use classes are defined in Table 1.

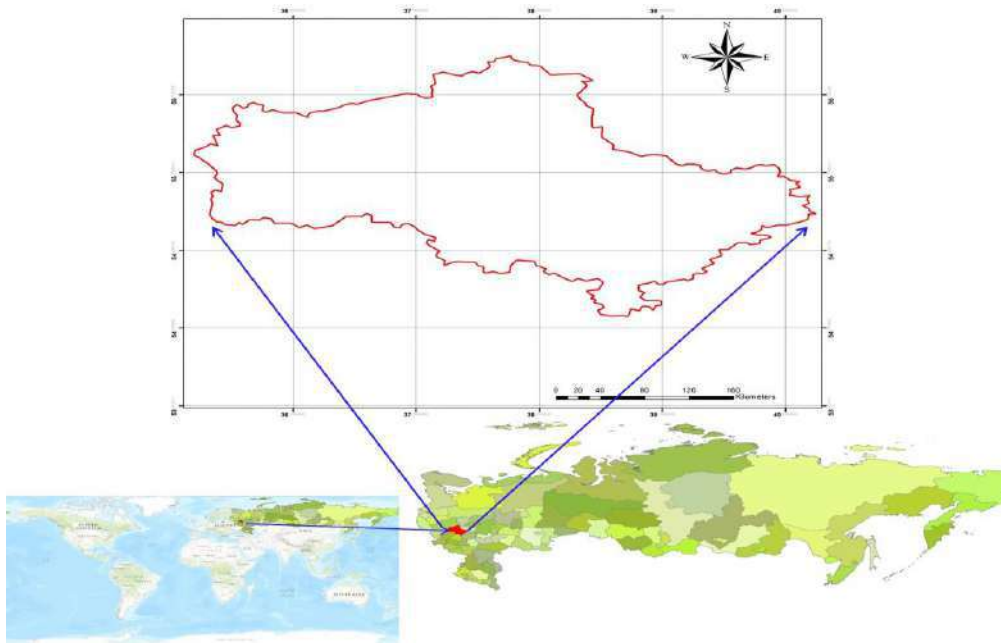


Fig. 1. Location map of the study area in Moscow Region, Russia.

Table 1. Land use classes definitions.

LULC Classes	Definition of Land Use Classes
Agricultural	Cultivated areas, crop lands, grass lands, vegetables, fruits etc.
Barren land	This contains open lands mostly barren but also small vegetation.
Forest	Small trees and shrub vegetation area except for vegetation.
Scrubland	Scrub is a plant community describe by vegetation shrubs, often also including grasses and herbs.
Settlements	Includes construction activities along the coastal dunes as well as sporadic houses within the local village and some governmental buildings.
Water body	All the water within land mainly river, ponds, lakes etc.
Wetland	A wetland is a land area with standing water and low soil fertility.

### 3.1 Database preparation

Any study of land use changes will involve the analysis of both conventional and remotely sensed data. Conventional data is more accurate and site specific, but its collection is time consuming, manpower hungry and difficult to extrapolate over a larger area. Remotely sensed data, on the other hand, has several advantages due to its repetitive and synoptic coverage of large and inaccessible areas in a quick and economical fashion. In the present study both conventional and remotely sensed data were used. The specific satellite images used were Landsat ETM+ (Enhanced Thematic Mapper plus) for 1995 and 2005, Landsat OLI (Operational Land Imager) for 2016, an image captured by a different type of sensors at a resolution of 30m were used for land use/cover classification. These data sets were imported in ArcGIS 10.2 software. Satellite images were making by processing software to create composites. A Trimble hand-held GPS with an accuracy of 10 meters was used to map and collect the coordinates of important land use features during pre- and post-classification field visits to the study area in order to prepare land-use and land-cover maps.

### 3.2 Image classification

Land cover classes are typically mapped from digital remotely sensed data using some sort of supervised, digital image classification. The overall objective of the image classification procedure is to automatically categorize all pixels in an image into land-cover classes or themes and the maximum likelihood classifier quantitatively evaluates both the variance and covariance of the category's spectral response patterns whenever it classifies an unknown pixel. This is why it is considered to be one of the most accurate classifiers - it is based on statistical parameters. Supervised classification was performed here using ground checkpoints and digital topographic maps

### 3.3 Land use/cover change detection and analysis

Land use maps shows in figures 2, were prepared using Landsat data. The accuracy of these classified maps was checked using the GIS tools. The accuracy for these periods is 90% respectively. There is a big change in land use during this time period. To order increase the accuracy of the land use mapping of the two images, ancillary data, and the result of visual

interpretation was integrated with the classification results using Arc GIS [12, 13]. The classification of imagery from each individual year, a multi-date, post-classification comparison, change-detection algorithm was used to determine changes during two intervals from 1995-2005 and 2005-2016. This is perhaps the most common approach to change detection. The post-classification approach provides 'from-to' change information which facilitates easy calculation and mapping of the kinds of landscape transformations that have occurred [14]. Accuracy assessment was then carried out at 85 points, 65 from the field data and 20 from existing topographic maps and the land cover map. Specification of these 85 points used a stratified, random method so that all of the different land-cover classes would be represented. In order to increase the accuracy of the land-cover mapping of the two images, ancillary data as well as the result of visual interpretation was integrated with the classification results using GIS [14]. The aim of this was to improve the classification accuracy of the classified image.

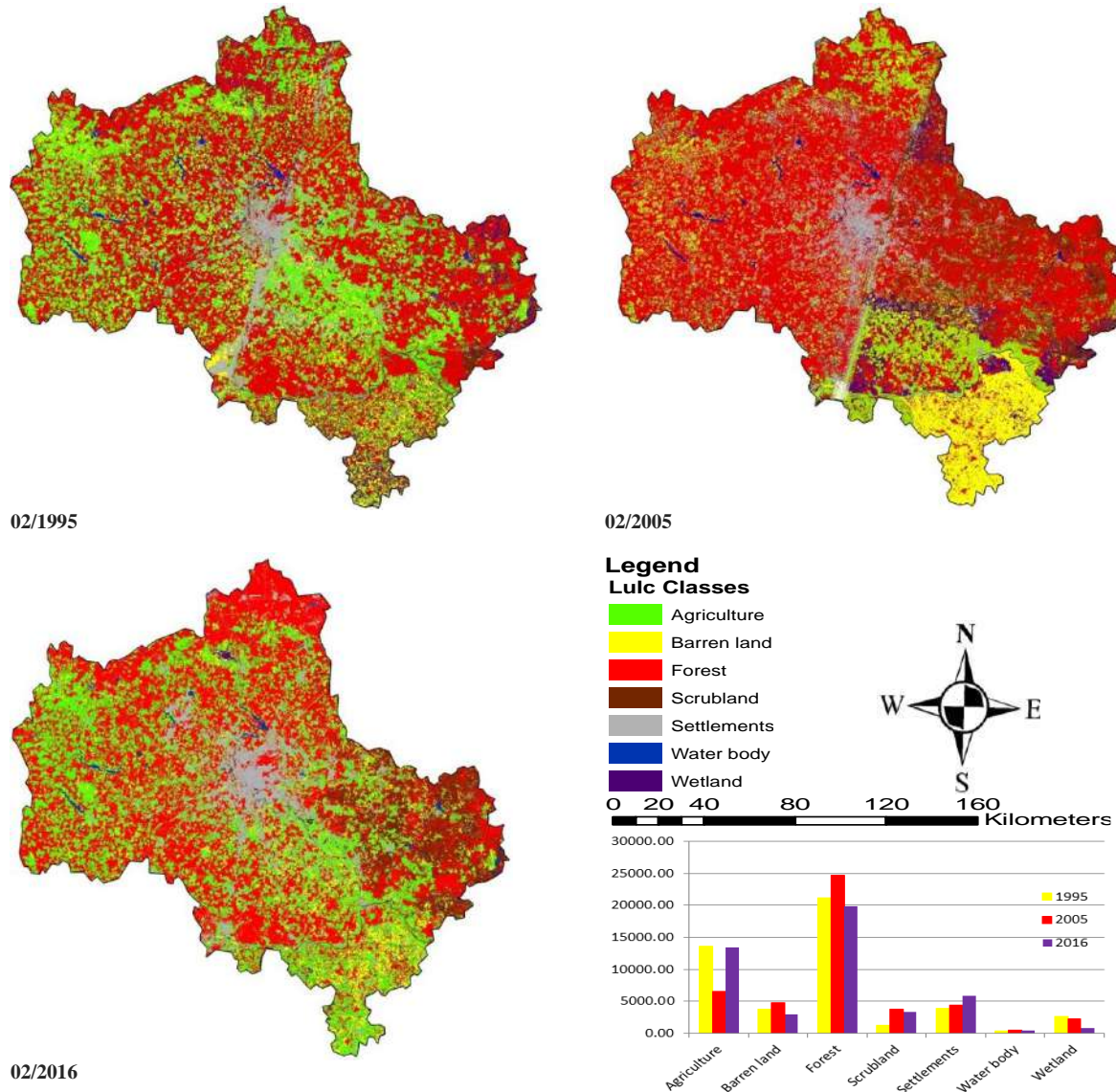


Fig. 2. Land use of Moscow Region, Russia; (a) in 1995, (b) in 2005 and (c) in 2016.

#### 4. Results and Discussion

Figure 2 shows land use image after supervised classification. These images provide pattern of land use/cover of the study area. The green color represent agricultural, yellow color barren land, red color forest, gray color settlements, brown color shows the scrubland, blue color shows water body and purple color shows wetlands. All land cover class maps were compared with reference data. Over all classification accuracy of the study area was more than 90% all three dates.

There is a big change in land use during this time period, as show in the graphical representation of the data in figure 2. Classification maps were generated for all of the sixteen years shown in figure and the individual class area and change statistics are summarizes in table 1. In 1995 the urban area covered 3898.31 km<sup>2</sup> (8.34 %), but by 2005 it had increased to approximately 4361.75 km<sup>2</sup> (9.33 %) and in 2016 had increased to 5852.00 km<sup>2</sup> (12.51). The agricultural area first half decreased from 13673.51 km<sup>2</sup> (29.24 %) in 1995 to 6504.00 km<sup>2</sup> (13.91 %) by 2005 and then increased to 13403.62 km<sup>2</sup> (28.66 %) by 2016. The forest area increased from 1995 21135.18 km<sup>2</sup> (45.19 %) to 24671.31 km<sup>2</sup> (52.75 %) by 2005 and then it was decreased

Image Processing, Geoinformation Technology and Information Security / Komal Choudhary, M.S. Boori, A. Kupriyanov from 2016 to 19896.64 km<sup>2</sup> (42.54 %). The barren land area was 3802.63 km<sup>2</sup> (8.13 %) in 1995, in 2005 had increased 4717.74 km<sup>2</sup> (10.09 %) and then it had decreased 2993.18 km<sup>2</sup> (6.40 %) by 2016.

All the urban categories increased continuously, with the urban area increasing by 1953.69 km<sup>2</sup> (4.17%) since 1995. Results show that forest area has been most dominant class in the study area for all three dates. The land use transition during the 1995-2016 periods is shown in table 2.

Table 2 shows both positive and negative land use/cover changes in the study area from 1995 to 2005, the major change was in agriculture and forest area. Forest was increased 3,536.13 km<sup>2</sup> (7.56) and agriculture was decreased 7169.51 km<sup>2</sup> (15.33%) of the total study area due to harsh climatic conditions. From 2005 to 2016 total agriculture area was increased from 6,899.62 km<sup>2</sup>. In the same time period other classes such as barren land, scrubland, settlements, water body and wetland increased respectively. From 2005 to 2016 total agricultural area was increased from 6,899.62 km<sup>2</sup> and other classes settlements and waterbody were increased.

Table 2. Area and amount of change in different land use categories in the study area during 1995 to 2016.

Class	1995		2005		2016	
	Area KmSq	%	Area KmSq	%	Area KmSq	%
Agriculture	13673.51	29.24	6504.00	13.91	13403.62	28.66
Barren land	3802.63	8.13	4717.74	10.09	2993.18	6.40
Forest	21135.18	45.19	24671.31	52.75	19896.64	42.54
Scrubland	1268.97	2.71	3791.82	8.11	3377.49	7.22
Settlements	3898.31	8.34	4361.75	9.33	5852.00	12.51
Water body	408.96	0.87	430.13	0.92	449.45	0.96
Wetland	2580.57	5.52	2291.37	4.90	795.75	1.70
Total	46768.12	100.00	46768.12	100.00	46768.12	100.00

Table 3. Land use change showing land encroachment of the study area.

1995-2005	CLASS	AGRICULT	BARREN_L	FOREST	SCRUBLAN	SETTLEME	WATER_BC	WETLAND	Total
	Agriculture	3820.91	2119.56	4633.85	1396.83	1318.99	15.29	503.14	13808.57
	Barren land	867.28	1091.05	910.37	215.43	522.59	1.39	127.87	3735.99
	Forest	990.98	583.75	15982.20	1517.75	605.99	44.48	1384.32	21109.46
	Scrubland	198.75	244.62	365.54	205.70	82.00	20.85	125.09	1242.55
	Settlements	458.66	276.59	1178.62	198.75	1599.75	13.90	151.50	3877.76
	Water body	8.34	4.17	41.70	22.24	40.31	293.26	2.78	412.79
	Wetland	150.11	418.35	1662.29	137.60	122.31	2.78	87.56	2581.00
	Total	6495.04	4738.09	24774.56	3694.29	4291.94	391.95	2382.25	46768.12
2005-2016	CLASS	AGRICULT	BARREN_L	FOREST	SCRUBLAN	SETTLEME	WATER_BC	WETLAND	Total
	Agriculture	3926.40	622.67	1067.43	137.60	717.18	1.39	40.31	6512.97
	Barren land	2711.65	964.57	414.18	27.80	528.15	2.78	88.95	4738.09
	Forest	4401.14	480.16	15697.28	2155.70	1798.50	41.70	69.49	24643.96
	Scrubland	1129.97	500.88	906.20	915.93	300.21	38.92	29.19	3821.29
	Settlements	1099.39	291.30	451.71	43.09	2354.45	27.53	45.87	4313.33
	Water body	59.75	0.00	16.68	11.12	36.14	326.62	0.00	450.30
	Wetland	220.88	116.75	1238.38	193.19	230.72	12.13	276.12	2288.17
	Total	13549.18	2976.32	19791.85	3484.42	5965.35	451.06	549.92	46768.12

The results show that from 1995 to 2005, 3820.91 km<sup>2</sup> agriculture areas were stable but 990.98 km<sup>2</sup> areas converted from forest to agriculture (table 3). In the same time period 15982.20 km<sup>2</sup> forest areas were stable but 1662.39 km<sup>2</sup> wetland area was encroached by forest. Maximum stable class was water body, where 293.26 km<sup>2</sup> areas were stable from 1995 to 2005. In the second half from 2005 to 2016 3926.40 km<sup>2</sup> agriculture area was stable and 2711.65 km<sup>2</sup> barren land, 4401.14 km<sup>2</sup> forest and 1129.97 km<sup>2</sup> scrubland area converted into agriculture land due to increase of market demand. In this time period there is not a big change in wetland and maximum bare land area 276.12 km<sup>2</sup> was stable. Scrubland 906.20 km<sup>2</sup> and wetland 1238.38 km<sup>2</sup> area was converted into forest area which shows governmental protection from 2005 to 2016. Since 2005 to 2016, 2354.45 km<sup>2</sup> settlements area was stable but 1798.50 km<sup>2</sup> forest area was converted into settlements. In the second half again water body area was highly stable area around 326.62 km<sup>2</sup>.

As shown by our study, land cover change is mainly driven by the expansion of socio-economic activities. The increase of agricultural areas, if poorly managed has impacts above those previously mentioned changes in the soil water cycle, nutrient



depletion and an increased risk of soil erosion and land degradation even though the expansion of croplands leads to a growth in agricultural outputs like food and fibers to positively impact on the country's economy and human well-being.

As well as the huge increase in agricultural area there has also been a considerable increase in urban settlements. Such changes require rapid adjustments to land management in order to avoid crises in food. From a socio-economic point of view this means not only a loss of ecosystem services, but also a decline of earned money and cultural values, not to mention a subsequent reduction of income from tourism. A consequence of this is to make protected areas some of the few remaining zones where fuel wood, rich pastures and game resources are left and so they attract more and more legal activities.

## 5. Conclusion

In this new period of globalization cities should have quality infrastructure, energy and environment condition to sustain growth and attract foreign investment. The planning authorities should adopt new technologies such as remote sensing and GIS to address these issues. Remote Sensing and GIS are adequate of providing the necessary information and intelligence for planning proposal. In this research remote sensing and GIS have been unified to exhibit the changes in urban development and its future growth trends. This study focuses on discovering the expansion of the urban area of Moscow region. A large percentage of barren land was transformed into urban area during the study period. The urban growth shows maximum detail on the outskirts of the region. This expansion also indicates of industrial growths. Only remote sensing data can provide complete spatial information for the efficient assignment of urban growth in developing countries over the time period.

## Acknowledgements

This data work is financially supported by the Russian Scientific Foundation (RSF), grant no. 14-31-00014 "Establishment of a Laboratory of Advanced Technology for Earth Remote Sensing".

## References

- [1] Baker WL. A Review of Models of Landscape Change. *Landscape Ecology* 1989; 2: 111–133.
- [2] Bauer T, Steinnocher K. Per Parcel Land Use Classification in Urban Areas Applying a Rule-Based Technique. *GeoBIT/GIS* 2001; 6: 24–27.
- [3] Boori MS, Choudhary K, Kupriyanov A. Vulnerability analysis on Hyderabad city, India. *Computer Optics* 2016; 40(5): 752–758. DOI: 10.18287/2412-6179-2016-40-5-752-758.
- [4] Alemayehu M. Forage Production in Ethiopia: A case study with implications for livestock production. Ethiopian Society of Animal Production (ESAP), Addis Ababa, Ethiopia, 2002.
- [5] Allan J. Sensors, Platforms and Applications; Acquiring and Managing Remotely Sensed Data. Application of Remote Sensing in Agriculture. Butterworths, London, 1990.
- [6] Fang S, Gertner GZ, Sun Z, Anderson AA. The impact of interactions in spatial simulation of the dynamics of urban sprawl. *Landscape and Urban Planning* 2005;73: 294–306.
- [7] Henriquez C, Azocar G, Romero H. Monitoring and modeling the urban growth of two medium-sized Chilean cities. *Habitat International* 2006; 30: 945–964.
- [8] Asner G. Contributions of multi-view angle remote sensing to land-surface and biogeochemical research, *Remote Sensing of Environment* 2000; 18: 137–162.
- [9] Boori MS, Choudhary K, Soifer VA, Sugimoto A. Computer simulation of satellite data for urban expansion analysis. *International Journal of Mathematics and Computers in Simulation* 2016; 10: 142151.
- [10] Boori MS, Kuznetsov AV, Choudhary K, Kupriyanov A. Satellite image analysis to evaluate the urban growth and land use changes in the city of Samara from 1975 to 2015. *Computer Optics* 2015; 39(5): 818–822. DOI: 10.18287/0134-2452-2015-39-5-818-822.
- [11] Dewan AM, Yamaguchi Y. Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Applied Geography* 2009; 29: 390–401.
- [12] Boori MS, Choudhary K, Kupriyanov A, Sugimoto A, Kovelskiy V. Monitoring land use/cover change detection through remote sensing and GIS techniques in Eastern Siberia, Russia. *SGEM 2016 Conference Proceedings* 2016; 2: 971–978. DOI:10.5593/SGEM2016/B22/S10.124.
- [13] Boori MS, Choudhary K, Kupriyanov A, Sugimoto A, Evers M. Natural and environmental vulnerability analysis through remote sensing and GIS techniques: A case study of Indigirka River basin, Eastern Siberia, Russia. *Proc. Of SPIE* 2016; 10005(100050U): 1–10. DOI:10.1117/12.2240917.
- [14] Choudhary K, Boori MS, Kupriyanov A. Landscape Analysis through Remote Sensing and GIS Techniques: A Case Study of Astrakhan, Russia. *SPIE 2017 Conference Proceedings*, SPIE 2017; 10225: 102251U-1. DOI: 10.1117/12.2266245.

# Program-algorithm complex for image imposition in aircraft vision systems

A.I. Efimov<sup>1</sup>, A.I. Novikov<sup>1</sup>

<sup>1</sup>Ryazan State Radio Engineering University, Ryazan, 390005, Russia

---

## Abstract

One of the most important tasks being solvable on the aircraft board is a task of imposition of real images and images synthesized according to the digital terrain map. Complex of auxiliary tasks and actual imposition task should be solved on a real time basis (with frequency 25 frames per second) and with strict requirements to accuracy of the heterogeneous image imposition. Traditional correlation-extremal methods of imposition ensure a necessary accuracy but require unacceptably high expenditures of computer time. The paper describes an algorithm of imposition based on affine transformations of the synthesized image to the plane of a real video image and also algorithms for solution of auxiliary tasks.

*Keywords:* preprocessing; skeleton; image enhancement; affine transformations; projective transformation; image imposition

---

## 1. Introduction

Necessity to improve aircraft flight security, to ensure safety of landing requires a development of new methods for integration and interpretation of information obtained from onboard technical vision systems (TVS) of various spectral ranges and also from navigation devices and digital terrain map (DTM) [1,2]. Video information obtained from TVS together with the synthesized image of terrain relief, cartographic and navigation information obtaining in real time help the crew to pilot and land under conditions of low visibility. Onboard TVS can contain a television (TV) camera, infrared imager, lidar and radar which form TV, thermal imaging (TI) and location images of the underlying surface respectively.

## 2. Object of research

Object of the research is a process of imposition of a real television (TV) image obtained from the television camera installed on an aircraft board and a synthesized image in onboard TVS. Synthesized image is formed into an onboard calculator according to the digital terrain map. Imposition of real and synthesized images in onboard TVS is one of the most important and complicated tasks being solvable into the onboard computer complex. Issues occurring under its solution are caused by several reasons. One of the main reasons is errors in detection of current coordinates of an aircraft as a material point in the air space (latitude  $\lambda$ , longitude  $\varphi$  and height  $h$ ) and also errors in determination of aircraft orientation as an extended object in the space. Errors in measurement of parameters of the yaw  $\psi$ , pitch  $\theta$  and roll  $\gamma$  also belong to them. Errors can be in the digital terrain map (DTM). Also another source of errors can be sensors forming images. Different nature of real and synthesized images is one more reason complicating solution of the image imposition tasks. Positioning errors can be added by geometrical distortions appearing on processed images at stages of the boundary detection of brightness jump and formation of closed circuits.

## 3. Methods of research

Widely known correlation-extremal methods of image imposition, as practice of their application shows, ensure enough good quality of imposition [3]. However, in the present case search of a global extremum of the objective function is joined with formation of  $10^6$  angles (sets of 6 numbers - vector coordinates  $\mathbf{v} = (x, y, h, \psi, \theta, \gamma)$  of navigation parameters) and, as a consequence, with unacceptably high expenditures of computer time. There are known approaches to reduce a spatial dimension due to usage of the extended angle under formation of the synthesized image and application of a pyramid of different-scaled images for consecutive refinement of a point of the objective function global optimum under the correlation-extremal approach to image imposition [4]. Such modernization of correlation-extremal methods of imposition leads to reduction of computation efforts but does not solve the issue up the end both regarding time and accuracy of the task solution.

Limiting values of errors in determination of navigation parameters set a corresponding parallelepiped in the 6-dimensional space of parameters. Within this parallelepiped, a grid with nodes is formed. A synthesized image is constructed by values of the vector of parameters at each node. This image is overlapped onto a real image; a value of the objective function being a measure of the imposition quality is calculated and stored. After enumeration of all nodes the objective function global extremum and vector of navigation parameters  $\mathbf{v}_{opt}$  where this optimum is achieved are found.

Alternative method to impose images is reduced to a search of the same objects on a pair of heterogeneous images, their comparison, calculation of the geometrical transformation of one image to the plane of other one and imposition of images for representation to a pilot. Algorithms based on the geometrical transformation of the synthesized image to the plane of a real one require less computation efforts for their implementation than correlation-extremal methods and provide acceptable quality of image imposition. However, these methods are applied only in cases if contours of continuous presence objects (water objects, roads, large buildings etc.) are stably distinguished on the underlying surface. Besides, bottleneck of these methods is a search of

key (corresponding) points onto a pair of imposed images by means of which transformation of the synthesized image is constructed to the plane of a real one. Correctness of selection of key points determines a degree of precision of “imposed” images, i.e. quality of image imposition [5]. Methods for enhancement of image imposition by means of projective transformations under presence of less informative areas on the image and, as a consequence, presence of incorrect pairs of key points are suggested in papers [6,7]. Alternative and widely spread method for construction of a qualitative projective transformation under presence of some subset of incorrect pairs in a set of pairs of key points is a usage of the algorithm RANSAC [8].

Imposition algorithm considered below is based on analysis and comparison of contours on a pair of images. Its basis is affine transformations of the synthesized image to the plane of a real video image. Affine transformations do not take into consideration projective distortions which inevitably appear under aerial photography [9]. Their advantage is simplicity of implementation, low computational efforts providing functioning in real time and satisfactory quality of image imposition. Suggested algorithm holds an intermediate position between correlation-extremal methods of imposition and methods of projective geometry. This algorithm is a certain compromise under conditions when mentioned methods cannot be implemented in automatic mode with acceptable accuracy and time characteristics.

Although imposition of images is a final and very important procedure but possibility and quality of image imposition significantly depend on how successfully auxiliary tasks are solved. These tasks include tasks of detection of contours on images, enhancement of images, identification of contours and setting of one-to-one correspondence between contours on a pair of heterogeneous images, formation of a set of pairs of key points.

### 3.1. Imposition algorithm on the basis of transformation in the complex plane

For realization of the algorithm it is necessary to have contours of continuous presence objects both on a real image and on a responding synthesized image. Specific requirements are imposed on quality of contours used in the algorithm. They are considered below.

Algorithm is based on an affine transformation of points on the complex plane according to formula  $z_k^{(r)} = z_{np} \cdot z_k^{(s)}$ ,  $z = x + iy$ , where  $z_k^{(s)}$  – a point on the synthesized image contour,  $z_k^{(r)}$  – a result of transformation to the real image plane,  $z_{np} = x_{np} + iy_{np}$  – a complex number providing transformation of points of one image to the plane of other one. It is required to find a pair of corresponding (key) points on the first and second contours in order to determine a complex number  $z_{np}$ .

Let's  $D$  be some area on the image with boundary  $\partial D$ . Suggested variant of the algorithm takes points belonging to ends of the area diameters as corresponding points. i.e.  $\{M_1, M_2\} = \arg \max_{M_i, M_j \in \partial D} \rho(M_i, M_j)$ . Such points are found for corresponding

objects both on the real and responding synthesized images. Vectors  $\mathbf{a}_1 = (x_2 - x_1; y_2 - y_1)$  and  $\mathbf{a}'_1 = (x'_2 - x'_1; y'_2 - y'_1)$  are placed in correspondence with found points  $M_1(x_1, y_1)$  and  $M_2(x_2, y_2)$  on the real image and responding points  $M'_1(x'_1, y'_1)$ , on the synthesized one. Complex numbers  $z_1 = x_2 - x_1 + i(y_2 - y_1)$  and  $z'_1 = x'_2 - x'_1 + i(y'_2 - y'_1)$  respond these vectors on the complex plane. Complex number  $z_{np}$  executing transformation of all points of the virtual image to the plane of the real one is determined according to formula  $z_{np} = \frac{z_r}{z_s} = \frac{z_r \cdot \bar{z}_s}{z_s \cdot \bar{z}_s}$ .

Algorithm for detection of points  $M_1, M_2$  belonging to ends of the diameter is the following. Let's take a random point  $M$  on the contour, choose a direction of the contour tracing and a near point  $\tilde{M}$  is found in this direction where local maximum is realized

$$\rho(M, \tilde{M}) = \max_{M_j \in \partial D} \rho(M, M_j) \quad (1)$$

Then point  $\tilde{M}$  is taken as initial one and here the following point is found for it where the local view maximum is achieved (1). After full contour tracing in the chosen direction a global maximum is separated from a set of local maximums and, as a consequence, unknown points  $M_1, M_2$  are found. Large volume of experimental researches of the algorithm for determination of points  $M_1, M_2$  belonging to ends of some contour diameter confirm its correctness. For search of points  $M_1, M_2$  belonging to ends of the diameter of a certain contour it is required that the contour should be closed. Contours of the real image obtained as a result of the algorithm for separation of boundaries of brightness jump can contain breaks. On the basis of information from the digital terrain map we know a type of the object and in particular it is clear that contours of its boundaries should be closed. Algorithm for formation of contours and additional processing is described below. The algorithm is aimed at enhancement of the contour image and, in particular, removal of breaks of small length which contours should be closed.

Quality of image imposition by means of the described algorithm depends on quality of separation of brightness jump boundaries. As a rule, skeletons obtained as a result of separation of brightness jump boundaries contain a great number of short lines. They significantly complicate a search of interested objects and determination of one-to-one correspondence between such objects (between object contours) on real and responding virtual images. For elimination of these disadvantages the algorithm of additional processing of contour images has been developed. It allows eliminating both closed and unclosed lines of short length. The algorithm is described in [10]. Fig.1 shows the original TV image, boundaries separated using Canny detector [11] and enhanced contour image, and also synthesized image responding to the original TV image correspondingly.

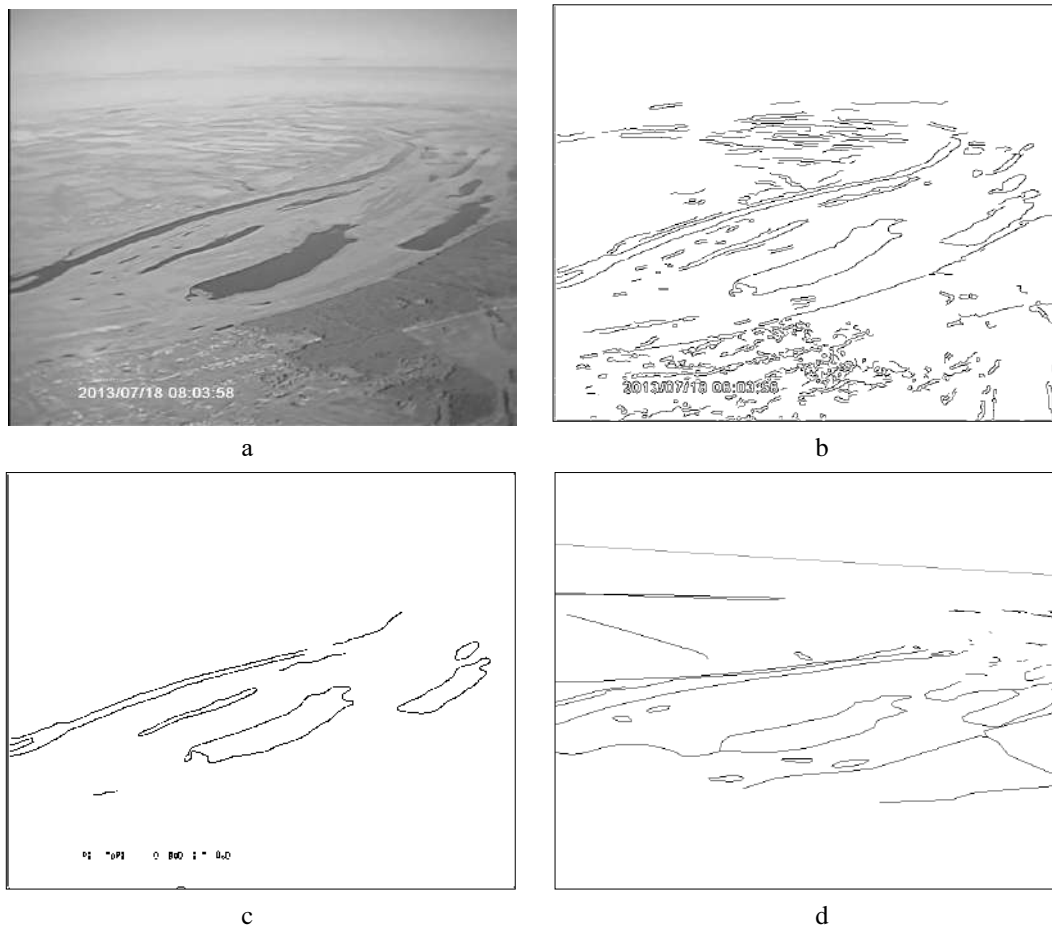


Fig. 1. Images at various stages of the technological chain execution: a – original TV image; b – boundaries separated using Canny detector; c – improved contour image; d – synthesized image constructed by the digital terrain map.

After achievement of the enhanced image of boundaries (Fig.1c) it is necessary to obtain a description of contours as a connected set of points. This procedure is realized as following:

- 1) pixels of the bitmap black-and-white image obtained as a result of the previous processing is looked through;
- 2) if black pixel is found, it is taken as a start of the contour and marked by the current pixel for analysis; iterative execution of steps 3-4 starts. Black pixels previously included into content of some contour is removed from consideration.
- 3) pixels adjoining the current one are looked through according to the order shown in Figure 2. In the case of detection of a neighbor in positions 1-8 it is added to the contour and marked as current one; operation is repeated;
- 4) if there are no black pixels in positions 1-8, contour tracing is stopped and return to step 1 occurs;
- 5) if all pixels of the image are looked through, algorithm operation completes.

5	1	6
3	X	4
8	2	7

Fig. 2. Order of point review under contouring.

In practice, there are often cases when contour of the extended object contains insignificant breaks. For this purpose, operation of additional combination is applied to contour descriptions obtained by the above-mentioned approach if distance between end points (the beginning and the end) of a certain pair of contours does not exceed a threshold value (threshold value is accepted as 7 pixels). Operation of additional combination is addition of pixels of one contour to list of pixels of another contour if above-mentioned condition is met. It allows removing small breaks and increasing quality of following procedures of imposition.



As a result, after processing of the whole image we have a set of connected contours. It is natural and logical to remove contours of short lengths, i.e. contours where a number of pixels is less than the threshold value (experimental researches have determined that for images having resolution  $704 \times 576$  pixels, contour length should be more than 120 pixels). Finally, we obtain connected contours of long length corresponding to extended objects on the original image. In addition to checking for exceeding of the minimum threshold length, we examine satisfaction to a range of additional conditions (these values will be described below in details):

- coordinates  $(x, y)$  of the object «center of mass» should be located on some surrounding of even one of objects from the virtual map, otherwise correspondences are not found for the mentioned object that negatively influences on the final result of image imposition;

- length  $d$  of the contour diameter should be not more than  $2/3$  of its length  $L$ ;

- width  $w$  of the contour is not less than 12 pixels.

Execution of mentioned conditions guarantee that only contours of extended closed objects which can be used for following imposition will be on images.

### 3.2. Automatic identification of contours and match making between them on real and virtual images

Digital terrain map contains information on object types which contours are reflected on the synthesized image. This information allows identifying corresponding objects on the synthesized image. Contour analog of the real image may not contain contours of some objects which, nevertheless, are present on the synthesized image. And otherwise contours can be present on the real image which contain specific distinctions from the corresponding contours on the synthesized image. Distinctions can be explained both by outdated state of the digital terrain map and disadvantages of algorithms of contour separation at all stages of the real image processing.

At the visual level, one-to-one correspondence between object contours is determined enough easily. Task to determine such correspondence by a computer in automatic mode is enough complicated. For solution of the present task we suggest the algorithm based on usage of some numerical characteristics. They are calculated for each contour on the real and synthesized images:

- coordinates  $(x, y)$  of the object «center of mass»;

- length  $L$  of the contour;

- length  $d$  of the contour diameter;

- width  $w$  of the contour.

Let's designate  $M_i^r(x_i^r, y_i^r)$ ,  $M_j^s(x_j^s, y_j^s)$ ,  $i = \overline{1, I}$ ,  $j = \overline{1, J}$  object centers of mass which contours are selected on real and synthesized images correspondingly. Here  $i$  - a number of the object on the real image and  $j$  - a number on the synthesized one. Coordinates of "centers of mass" are found as mean values according to the corresponding coordinate by all pixels of the contour. Matrix of distances between centers of mass of the size  $I \times J$  is constructed for a set of objects on real and virtual images, i.e.

$$\begin{pmatrix} \rho(M_1^r, M_1^s) & \rho(M_1^r, M_2^s) & \dots & \rho(M_1^r, M_J^s) \\ \rho(M_2^r, M_1^s) & \rho(M_2^r, M_1^s) & \dots & \rho(M_2^r, M_J^s) \\ \dots & \dots & \dots & \dots \\ \rho(M_I^r, M_1^s) & \rho(M_I^r, M_1^s) & \dots & \rho(M_I^r, M_J^s) \end{pmatrix}.$$

Distances are located in the Euclidean metric.

Analysis of correspondences between object centers of mass is based on enough really supposition that shift of the object on the synthesized image with regard to the object responding to it on the real image does not exceed a certain limiting value  $T$ . So, if in some line  $i_0$  of the distance table all distances are longer than this value then no object on the synthesized image corresponds with the object with number  $i_0$  on the real image. Correspondingly, if such situation has a place in  $j_0$ -column then no object on the real image corresponds with the object with number  $j_0$  on the synthesized image. Such objects will not further participate in the procedure for determination of correspondences between objects.

After removal of objects not having a corresponding object on other image from the matrix, procedure for determination of correspondences between objects begins. Let's  $I_1, J_1$  be numbers of residual objects on real and synthesized objects correspondingly. For all residual objects, both on real and virtual images the following is calculated:

- lengths  $L_i^r, L_j^s$ ,  $i = \overline{1, I_1}$ ,  $j = \overline{1, J_1}$  of contours;

- lengths  $d_i^r, d_j^s$ ,  $i = \overline{1, I_1}$ ,  $j = \overline{1, J_1}$  of diameters;

- values of object width  $w_i^r, w_j^s, i = \overline{1, I_1}, j = \overline{1, J_1}$ .

In this algorithm, a contour length is considered as a number of pixels in the contour. Area width is considered as a length  $|\mathbf{b}|$  of the vector  $\mathbf{b} = \overrightarrow{N_1 N_2}$  with ends on the contour in its average part and orthogonal to the vector  $\mathbf{a}$  – an area diameter.

After all numerical characteristics of all objects are found, a chain of computational procedures and comparisons is performed. Sequentially, in the cycle by  $i$  from 1 to the end of the real object list, objects are chosen and following actions are performed for each of them:

1) in the cycle by  $i$  execution of inequality  $\rho(M_i^r, M_j^s) < T$  is checked. Objects  $j_1, j_2, \dots, j_k$  which this inequality is fulfilled for, participate in the following comparison, the rest ones – not;

2) by each of three parameters  $L, d, w$  for  $i$ - contour on the real image, the nearest “neighbor” is searched on the synthesized image

$$\begin{aligned} j_1^* &= \arg \min_j \left\{ |L_i - L_{j_1}|, |L_i - L_{j_2}|, \dots, |L_i - L_{j_k}| \right\} \\ j_2^* &= \arg \min_j \left\{ |d_i - d_{j_1}|, |d_i - d_{j_2}|, \dots, |d_i - d_{j_k}| \right\} \\ j_3^* &= \arg \min_j \left\{ |w_i - w_{j_1}|, |w_i - w_{j_2}|, \dots, |w_i - w_{j_k}| \right\} \end{aligned} \quad (2)$$

Nearest “neighbor” in each of three conditions in (2) should comply with inequality  $|L_i - L_{j_1^*}| < \delta \cdot L_i, |d_i - d_{j_2^*}| < \delta \cdot d_i, |w_i - w_{j_3^*}| < \delta \cdot w_i$ . Here  $\delta = 1,3$ . These inequalities are based on suppositions that values of each of three parameters for corresponding objects on real and virtual images cannot differ more than  $\delta - 1$  per unit.

If  $j_1^* = j_2^* = j_3^* = j^*$ , then the decision is made that the object with number  $j^*$  on the virtual image corresponds to the object with number  $i$  on the real image. Even if all values  $j_1^*, j_2^*, j_3^*$  are different then we make a decision that the responding object on the virtual image is not found for the object with number  $j_1^*, j_2^*, j_3^*$  on the real image.

#### 4. Results

Fig. 3 shows contours of objects selected on real and synthesized images after the stage of image enhancement.

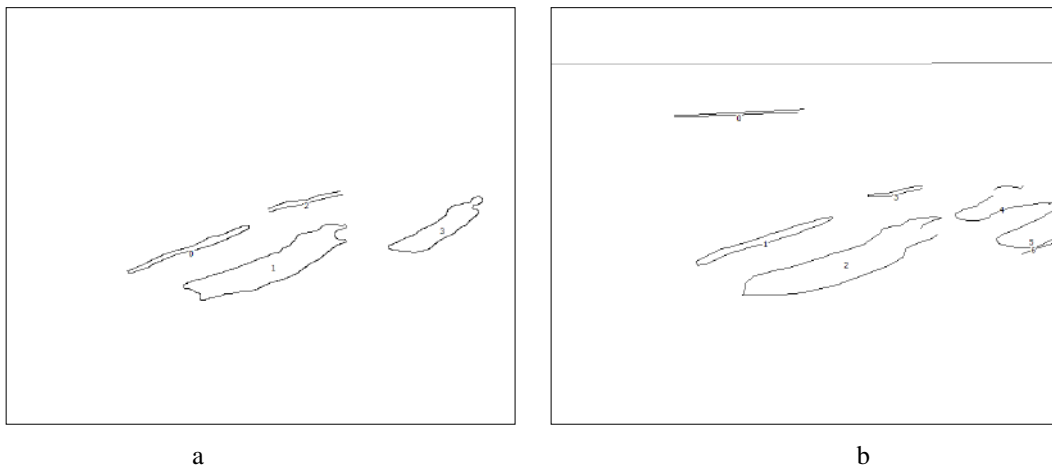


Fig. 3. Contours of objects selected on real (a) and synthesized (b) images.

Distances between each object on the real image and all objects on the virtual images are calculated according to the algorithm for determination of correspondence on a pair of images. These distances are shown in Table 1.

Object SI-0 (object with number 0 on the synthesized image) should be removed from the procedure for determination of correspondence between objects because distance from it to each of four objects on the real image is longer than the threshold value  $T$  ( $T=100$  was set as a threshold value for this experiment). Analysis of data from Table 1 allows supposing that by

minimum criterion of distances between objects on real and synthesized images (Fig.3) there are the following correspondences:  $RI-0 \Leftrightarrow SI-1$ ;  $RI-1 \Leftrightarrow SI-2$ ;  $RI-2 \Leftrightarrow SI-3$ ;  $RI-3 \Leftrightarrow SI-4$ .

Table 1. Distances between objects on the real and synthesized images.

Object	RI-0	RI-1	RI-2	RI-3
SI-0	187	233	196	376
SI-1	42	80	130	307
SI-2	152	39	82	200
SI-3	234	145	63	134
SI-4	371	266	208	35
SI-5	408	297	253	62
SI-6	411	299	259	69

However, control of determined correspondences by other parameters of objects according to algorithm (2) leaves only three correspondences for the final imposition that are:  $RI-0 \Leftrightarrow SI-1$ ;  $RI-1 \Leftrightarrow SI-2$ ;  $RI-2 \Leftrightarrow SI-3$ .

Now we can perform the final stage of processing – imposition of contours of the synthesized image on the real TV-image. Fig.4a shows a result of simple overlapping of the synthesized image on the real one. We can see significant discrepancies between contours which are expressed by the shift of the synthesized image in relation to the real one and by contour dimensions. Fig. 4b shows a final result of imposition performed according to the algorithm. At the visual level, we can estimate quality of imposition as satisfactory.

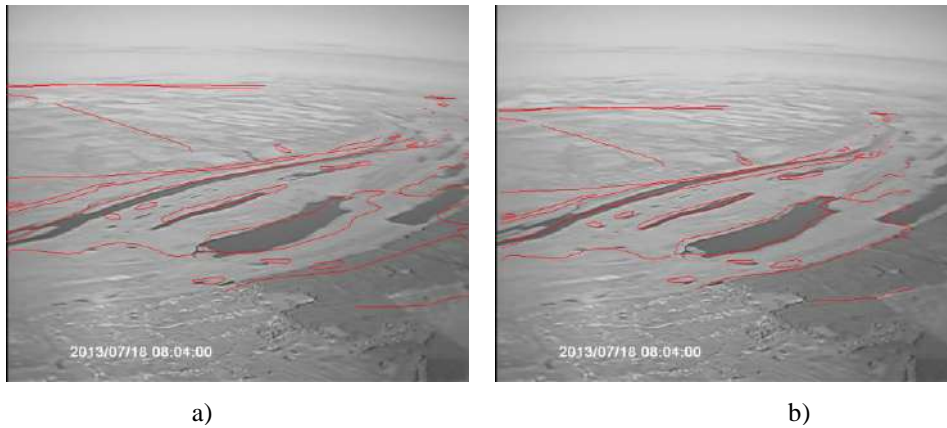


Fig. 4. a – result of overlapping of real and synthesized images, b – result of imposition.

Imposition of each eighth frame of the video sequence responding to imagery within 4 sec of flight has been performed to estimate algorithm efficiency. Results of imposition quality estimation by 13 pairs of real and synthesized by DTM frames from this video sequence are shown in Table 2. Imposition quality estimation has been performed using index  $\alpha$  introduced in paper [7].

Main idea of the suggested method is the following. Image is divided into square blocks (cells) of the specified size, e.g. 100x100 pixels. It provides a possibility to obtain not only integral estimation of imposition quality but also local estimations in each of separated square blocks. In each cell for all informative (other than background color) points of one of images, informative points of other image locating in certain square surrounding of size  $(2k+1) \times (2k+1)$  having its center in the processed informative point is searched. As a rule,  $k=1$  or  $k=2$ . Value  $k=1$  is equal to thickening of the thin line in one pixel of the first contour up to two pixels, and  $k=2$  – up to three pixels.

Sliding window of the chosen size ( $5 \times 5$  in the considered experiment) is moved along lines of the image. As soon as an informative point of the first image gets into the center of this surrounding then informative pixels of the second image getting into this surrounding and not marked at the previous stages are searched and marked. After scanning of the whole image is completed, in each square blocks a number of  $m_i$  marked points of the imposed (second) contour is calculated and we found a

ratio of this number to the total number of informative points of the first contour  $M_i$ , i.e.  $\alpha_i = \frac{m_i}{M_i}$ . Let's call coefficient  $\alpha_i$  as

an index of the imposition quality in  $i$ -block of the image and coefficient  $\alpha = \frac{\sum_i m_i}{\sum_i M_i}$  – an integral index of the whole contour

imposition quality.

Results of imposition quality estimation by 13 pairs of real and synthesized by DTM frames from this video sequence are shown in Table 2. Within processing of the first frame there was an imposition disruption because correspondence between contours on real and responding synthesized images could not be determined. Enhancement of the imposition quality was not achieved in two frames. Imposition quality increased within the range from 24% to 108% in 10 frames.

In paper [12] imposition of heterogeneous images on these 13 frames was performed using broken-linear transformations which allow taking projective distortions into consideration. This paper achieved higher indices of the image imposition quality. Technology of the image imposition described in [12] requires greater computational efforts and it has not yet been automatized up to the end.

Table 2. Results of the image imposition estimation.

Number of a pair of frames	Index $\alpha$ before imposition	Index after imposition	$\alpha$ Change of Absolut. /percent.	Expert estimation of image imposition
1	0.281	---		Imposition disruption
9	0.282	0.402	0.12/ 42.6%	enhanced
17	0.247	0.421	0.174/ 70.4%	enhanced
25	0.231	0.481	0.25/ 108.2%	enhanced
33	0.232	0.418	0.186/ 80.2%	enhanced
41	0.334	0.236	-0.098/ -29.3%	worsened
49	0.300	0.389	0.089/ 29.7%	enhanced
57	0.313	0.241	-0.072/ -23.0%	worsened
65	0.229	0.364	0.135/ 59.0%	enhanced
73	0.324	0.426	0.102/ 31.5%	enhanced
81	0.279	0.381	0.102/ 36.6%	enhanced
89	0.261	0.362	0.101/ 38.7%	enhanced
97	0.194	0.240	0.046/ 23.7%	enhanced
Mean values	0.268	0.363	0.095/ 35.4%	

Algorithm was also examined on real video sequences obtained from a TV camera consisting of OVS within long time interval. Fig.5 shows a diagram of quality estimation for 900 frames (36 sec flight). Estimation of mathematical expectation of the imposition quality in the mentioned video sequence was 0.27, estimated variance – 0.004. Research of influence of the scene nature on results of the imposition quality estimation has been performed. It determined that quality estimation of the imposition algorithm results based on transformation in the complex plane exceeds the index before imposition on the same fragments where we can separate and determine correspondence as minimum between two objects. For the frames where correspondence was not determined, quality estimation remains invariable.

## 5. Description of the program-algorithm complex

Program-algorithm complex of the image imposition in aviation vision systems which constituent is an algorithm for the image imposition based on transformation in the complex plane contains the following main blocks:

- 1) block for registration of images obtained from a vision sensor (in the case of program realization on a bench – frame capture from the video sequence);
- 2) block for construction of a virtual image by the virtual terrain model;
- 3) block of pre-processing of real and virtual images;
- 4) block for imposition and removal of geometrical mismatch of real and virtual images between each other;
- 5) block for visualization of the imposition result.

Blocks 2, 3 and 4 are most important for the whole complex and complicated from the point of view of their construction.

Construction of the virtual terrain model is realized by the separate software module in the developed realization of the image imposition complex. Its main functions are positioning of the virtual camera according to specified position coordinates and construction of the image by the available digital map in format sxf. It is possible to receive a frame both in thin lines and with overlapped textures.

Block of preprocessing realizes auxiliary operations required for execution of the geometrical imposition: separation of boundaries, removal of low informative lines, obtaining of connected contours – for real images; removal of lines of short extension and obtaining of connected contours – for virtual images. Quality of preprocessing greatly determines effectiveness of the following imposition.

Key element of the program-algorithm complex for image imposition is a block of geometrical imposition. The algorithm of imposition based on transformation in the complex plane as enough fast and reliable approach is suggested to be applied as one of algorithms for removal of geometrical mismatch. High-speed performance of the algorithm of image imposition based on transformation in the complex plane is the following: total costs for preprocessing do not exceed 0.3 sec per frame, for procedures of imposition – 0.05 sec. Calculations are performed using a computer equipped with processor Intel i7-3630QM,

2.40 GHz, random-access memory 8 Gb, realization of algorithms for preprocessing and algorithm for imposition is performed in the C++ programming language. Circuit of the program-algorithm complex organization is shown in Fig.6.

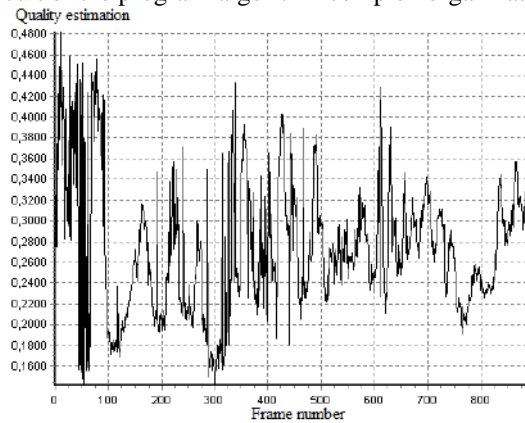


Fig. 5. Diagram of quality estimation for the video sequence.

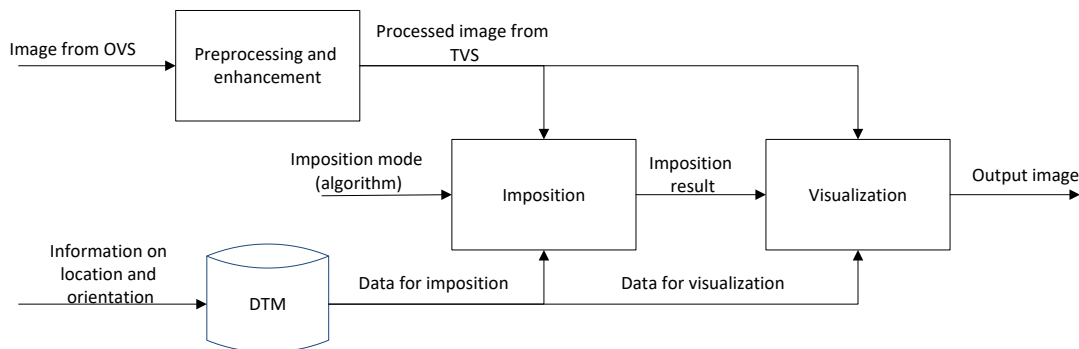


Fig. 6. Circuit of the program-algorithm complex organization for image imposition.

## 6. Conclusion

As it was mentioned before, correlation-extremal methods of imposition provide the necessary accuracy but require unacceptably high costs of computer time. Broken-linear transformations of the synthesized image to the plane of real video image are vulnerable because of issues with search of a set of key point pairs in the automatic mode [9]. Suggested algorithm is a certain compromise between requirements to accuracy of algorithms for imposition of heterogeneous images and requirements for their implementation in the automatic mode and in real time.

Considered algorithm can operate under large errors of navigation parameters only if similarity of contours of corresponding objects on real and synthesized images is saved.

Developed algorithm can be applied independently for imposition of images and it can be used in combined schemes with algorithms of a higher level for pre-imposition.

## References

- [1] Elesina SI. Imposition of images in correlation-extreme navigation systems. Edited by Kostyashkin LN, Nikiforov MB. Moscow: Radio Engineering, 2015; 208 p.
- [2] Wisilter YuV. Aviation systems of enhanced and synthesized vision: analytical review of materials of foreign information sources. State Scientific Center of the Russian Federation, State Scientific and Research Institute of Aviation Systems (Federal State Unitary Enterprise "GosNIIAS"), Scientific Information Centre; Edited by Fedosov EA. M., 2011; 77 p.
- [3] Baklitsky VK. Correlation-extreme methods of navigation and guidance. Tver: Book Club, 2009; 360 p.
- [4] Babayan PV, Ershov MD. Algorithms for removal of mismatches in the onboard vision system. Vestnik of RSREU 2015; 4(2): 32–38.
- [5] Crum WR, Hartkens T, Hill DLG. Non-rigid image registration: theory and practice. The British Journal of Radiology 2014; 77: 140–153.
- [6] Goshin EV, Kotov AP, Fursov VA. Two-stage formation of spatial transformation for image imposition. Computer Optics 2014; 38(4): 886–891.
- [7] Efimov AI, Novikov AI. An algorithm for multistage projective transformation adjustment for image superimposition. Computer Optics 2016; 40(1): 258–266. DOI: 18287 / 2412-6179-2016-40-2-258-266
- [8] Hast A, Nysjö J, Marchetti A.. Optimal RANSAC – Towards a Repeatable Algorithm for Finding the Optimal Set. Journal of WSCG 2013; 21(1): 21–30.
- [9] Gruzman IS, Kirichuk VS, Kosykh VP, Peretryagin GI, Spector AA. Digital image processing in information systems. Novosibirsk: Publishing house of NSTU, 2002; 351 p.
- [10] Novikov AI, Sablina VA, AI Efimov. Image Superimposition Technique in Computer Vision Systems Using Contour Analysis Methods. 5th Conference on Embedded Computing (MECO) Proceedings 2016: 132–137.
- [11] John Canny. A Computational Approach to Edge Detection. IEEE Transactions on Pattern and Machine Intelligence 1986; PAMI-8(6): 679–698.
- [12] Novikov AI, Sablina VA, Nikiforov MB. Algorithms for automatic identification of objects on heterogeneous images and image imposition. Collection of papers of the III International Conference and Youth School "Information Technologies and Nanotechnologies" 2017: 599–607.

# The algorithm of the high-capacity information embedding into the digital images DCT domain using differential evolution

O.O. Evsutin<sup>1</sup>, A.O. Osipov<sup>1</sup>

<sup>1</sup>Tomsk State University of Control Systems and Radioelectronics, 40 Lenina Prospect, 634050, Tomsk, Russia

---

## Abstract

Methods of the steganography are characterized by such efficiency rates as invisibility, robustness and capacity. There is considered the maximum capacity support of the information embedding into the DCT-domain. It is investigated the known algorithm that realizes the adaptive information embedding into the digital images frequency domain. The adaptivity is reached due to the image partition into the unequal blocks using a quad-tree. There is received the improved modification of the algorithm based on the reference point variation in case of the image partition into the blocks. The received modification allows to provide the better invisibility at the same capacity.

*Keywords:* digital steganography; data hiding; digital images; DCT; optimization; differential evolution

---

## 1. Introduction

Digital steganography is one of modern directions of the informational security. Steganographic methods of protection of information allow one to solve such problems as the organisation of safe transmission of classified information and protection of copyrights of digital objects [1]. There is a common feature of all steganographic methods: hidden embedding of additional information in digital objects through embedding some modifications in the data elements composing a digital object. The properties that are required for the given process depend on a specific task.

The methods and algorithms of digital steganography are characterized by the following indexes of embedding efficiency: invisibility, robustness and capacity. A separate steganographic algorithm cannot ensure the maximum values of all specified parameters. The ratio between them can be described by the scheme presented on Fig. 1.

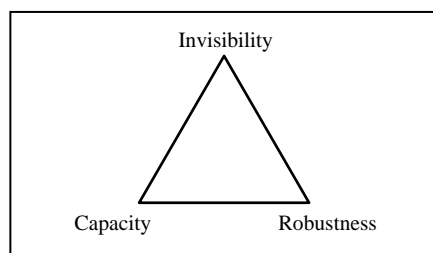


Fig. 1. Ratio between parameters of steganographic embedding efficiency.

Providing of an acceptable level of invisibility of embedding is the mandatory requirement to all steganographic methods and algorithms. Therefore it is possible to consider the contrast of two indexes of embedding effectiveness instead of three: capacity and robustness. The given contrast corresponds to the separation of methods of information embedding in digital objects into methods of arbitrary message embedding and methods of digital watermark embedding.

The methods which refer to the first class correspond to the classical concept of steganography and are designed for providing of confidentiality of the embedded information. In the second case, it is a question of copyright protection of digital objects. Digital watermarks represent special marks which contain information on the owners of the given objects. Digital watermarks are used for authentication of owners of digital objects, as well as for authentication of digital objects themselves including detection of falsifications.

It is necessary to note, that often embedding of digital watermarks does not refer to digital steganography and is not considered as a separate direction in the field of data hiding. However, such division is not always appropriate. Many methods of embedding of digital watermarks can be also used for embedding of limited capacity arbitrary messages into digital objects. Besides, there are methods enabling the control of the ratio between capacity and robustness of embedding. Therefore, further, we will equate methods of data hiding and methods of digital steganography.

Apart from tasks being solved, methods of digital steganography are classified according to the types of digital objects they process. Mainly, they are audio and video data and digital images. In the given paper we consider digital images as digital objects.

The methods of digital steganography operating with digital images divide on two big groups on domain of data embedding: embedding in the spatial domain and embedding in the frequency domain. The pixel matrix of a digital image is named as the spatial domain, and the frequency domain is the matrix of values received from a digital image by application of any frequency transform. The given data are also named as coefficients of frequency transform. In digital image processing including the embedding information into images the following transforms are used: discrete Fourier transform (DFT), discrete cosine transform (DCT), Walsh-Hadamard transform (WHT), various versions of discrete wavelet transform (DWT).

In the present paper, providing the maximum capacity of embedding in the frequency domain of discrete cosine conversion at maintenance of comprehensible quality of the cover image is considered. The known algorithm of high-density embedding is investigated and a new improved algorithm is offered.

## 2. 2. Methods of embedding information in the frequency domain of digital images

There exist many algorithms where information embedding is carried out in the frequency domain of digital images. Frequency transforms associate the matrix of pixels of a digital image with the matrix of frequency coefficients. Frequency coefficients can be divided on significant (carrying the basic information of a source image), and insignificant (that can be discarded or modified without any noticeable distortions in the initial image) [2]. Therefore frequency embedding allows to better choose data elements which can be used for not noticeable recording of additional information.

Let us note some research papers of last years.

Algorithms based on DFT are mainly used for embedding of digital watermarks. It is connected to properties of the given transform which do not permit to ensure the high capacity of embedding; however, they ensure resistance against some types of attacks on cover images. The majority of such algorithms operate with elements of the amplitude Fourier spectrum.

In the algorithm presented in [3], space of hiding is formed of the middle frequency elements with values in the set range. For embedding of one bit of a digital watermark a pair of symmetrically allocated elements varies so that the difference between them accepts certain value depending on the embedded bit.

In paper [4] a digital watermark is formed as the amplitude Fourier spectrum with elements accepting values from set  $\{0, 1\}$ . Significant elements form a circumference in the area of middle frequencies. It ensures stability in case of geometry attack like “turn of image”.

In [5] for formation of the binary digital watermark with circle symmetry, log-polar mapping is used. When being embedded those elements of the peak Fourier spectrum of the digital image that correspond to elements of the digital watermark with values 1 are converted by averaging over neighborhood  $3 \times 3$  with multiplication by the coefficient of amplification.

Another large class of steganographic algorithms works with DCT frequency domain. This class contains both algorithms of digital watermark embedding and algorithms of arbitrary message embedding. Besides, all algorithms of embedding of information into compressed JPEG images also operate with DCT coefficients because the given frequency transform is the basis of an appropriate compression method.

The papers [6, 7] can be considered as instances of classical papers in the given field. Paper [6] presents a resistant method of digital watermark embedding. The resistance is attained due to small capacity of embedding. One block of DCT coefficients of the size of  $8 \times 8$  contains one bit of a digital watermark. Embedding consists in determination of certain ratio between the pair of DCT coefficients depending on the value of the built-in bit.

The QIM method presented in [7] has capacity. It operates with low-frequency DCT coefficients. The given method uses two different quantizers for embedding zero and on-bits of the secret message into DCT coefficients of the cover image.

Paper [8] presents an instance where the increase of the effectiveness of digital watermark embedding is considered as an optimisation problem. The genetic algorithm is applied to its solution. It is used for sampling of an optimal order of embedding of parts of a digital watermark into DCT coefficients of the cover image.

Papers [9–11] present algorithms based on discrete wavelet transform.

Paper [9] considers embedding of biometric data of the owner into a digital image. The digital watermark represents a picture of a retina of an eye. Details wavelet coefficients are used for embedding. The digital watermark is converted to a character sequence existing in alphabet  $\{-1, 1\}$ , and embedding is carried out in an additive way. The choice of certain coefficients for embedding is carried out by means of a key.

Paper [10] presents the algorithm of embedding of semi fragile digital watermarks. Such digital watermarks are used for protection of images against falsification. They are resistant to usual processing of images (compression, resizing, filtering), but are destroyed in the case of modification of the image content, for example, when adding or removing of objects. The algorithm presented in [10] divides the image into blocks of an equal size, transfers them in the frequency domain by means of DWT, then low-frequency components of certain blocks are embedded in the high-frequency components of other blocks. The recombination of blocks is carried out by means of the generalized cat map. In [10] an elementary representative of DWT set — Haar transform — is used as frequency transform.

The algorithm of embedding presented in [11] is based on block quantization of DWT coefficients in the quadrant middle-frequency sub-band LH2. One bit of the message is built in into the block of  $k$  DWT coefficients. Embedding consists in the modification of summarised energy of coefficients of the block so that depending on the value of the built-in bit, it meets certain condition. The modification of value  $k$  changes the ratio between capacity and robustness. If  $k$  is increased, the built-in message obtains properties of a digital watermark.

In paper [12] embedding is carried out in the WHT frequency domain. For this purpose the image is divided into blocks by  $4 \times 4$  pixels; and WHT is applied to each block. The algorithm of embedding is built using a linear predictor function. Values of AC-coefficients of WHT of each block are predicted on the basis of DC-coefficient values of 8 adjacent blocks. The message bits are built in prediction errors according to the LSB method. To determine the weighting coefficients of the linear predictor function the neural network is used.

A series of publications [13–15] represents results of research directed on reaching the maximum capacity of embedding in the frequency domain of discrete cosine transform.

In paper [13], the cover image is divided into non-overlapping blocks by the size of  $m \times m$  pixels; DCT is applied to each block. For embedding, a part of the DCT-coefficient block is used, that forms a square in the right lower angle. This square corresponds to the least significant high-frequency coefficients and has a different size for different blocks. The size of embedding area is defined by the quantization matrix. Embedding consists in replacement of DCT-coefficients in the area of embedding by elements of the secret message. The secret message is also a digital image; and pixels of this image are exposed to additional quantization before embedding.

The given approach is developed in papers [14, 15]. The algorithm presented in [14] represents a different method of the secret image processing before embedding it. In [15], the cover image is divided into homogeneous blocks of pixels having unequal size by using quad-tree, that allows to raise the efficiency of embedding.

The present paper develops the offered in [13–15] approaches to high-capacity embedding of information into the frequency domain of discrete cosine transform. In the following section of the paper a more detailed description of algorithm [15] is given; probable ways of its improvement are defined and a new more effective algorithm is offered.

### 3. New algorithm on the basis of the approach to high-capacity embedding of information into the frequency domain of discrete cosine transform

#### 3.1. Adaptive algorithm of embedding using a quad-tree

Let's consider the QTAR embedding algorithm presented in article [12] in more details.

##### Input:

Square cover image  $I$ ; secret image  $S$ ; homogeneity threshold of block  $Th$ ; minimum block size  $m$ ; square matrix of quantization of a size  $8 \times 8$   $Q$ ; scale factor  $k$ .

##### Output:

Cover image containing a secret image  $I'$ .

**Step 1.** To execute recursive partition of each inhomogeneous square pixel block of the cover image into four equal sub-blocks. The cover image is taken as an initial block. The block partition stops if its size (the square side) is less or equal to  $m$  or if it is homogeneous. The block is recognized as inhomogeneous if the difference between the maximum and minimum values of pixels is higher than  $255Th$  value.

**Step 2.** To execute the scaling of the secret image pixels by formula  $\tilde{s}_i = k/255 s_i$ .

**Step 3.** For  $j = \overline{1, N}$ , where  $N$  — is amount of blocks in the quad-tree to execute as follows:

**Step 3.1.** To execute two-dimensional DCT of the  $j$ -th block of pixels of the size  $m_j \times m_j$ .

**Step 3.2.** To expand matrix  $Q$  to the extent of  $m_j \times m_j$  using interpolation and to divide the DCT-coefficients of the block into elements of the given matrix with the subsequent round-off.

**Step 3.3.** To select a square area of the greatest possible size  $n_j \times n_j$ , consisting only of nulls in the right lower angle of each block of the quantized DCT-coefficients.

**Step 3.4.** In the initial block of DCT-coefficients (before quantization) to substitute area of embedding  $n_j^2$  with pixels of modified secret image  $\tilde{S}$ .

**Step 3.5.** To execute inverse two-dimensional DCT.

**Step 4.** To return stego image  $I'$  and key sequence  $(n_1, n_2, \dots, n_N)$  and complete the algorithm.

The algorithm of extraction of the secret message is as follows.

##### Input:

stego image  $I'$ ; key sequence  $(n_1, n_2, \dots, n_N)$ ; threshold of block homogeneity  $Th$ ; minimum block size  $m$ ; scale factor  $k$ .

##### Output:

extracted secret image  $S'$ .

**Step 1.** To represent the stego image in the form of a quad-tree out of  $N$  blocks with the size not less than  $m \times m$  pixels with the threshold value  $Th$ .

**Step 2.** For  $j = \overline{1, N}$  to execute as follows:

**Step 2.1.** To execute two-dimensional DCT of the  $j$ -th block of pixels with the size  $m_j \times m_j$ .

**Step 2.2.** To select in the right lower angle of the received block of DCT coefficients a square block of embedded data elements with the side  $n_j$ .

**Step 2.3.** To execute an inverse scaling of the selected block elements using the formula  $s'_p = 255/k \tilde{s}_p$ ,  $p = \overline{1, n_j^2}$ , to derive the block of pixels of the secret image.

**Step 3.** To restore secret image  $S'$  from separate blocks of pixels.

**Step 4.** To return the extracted secret image  $S'$  and to complete algorithm.

Generally, extracted image  $S'$  does not coincide with initial secret image  $S$ . The pixels of the secret image are restored inaccurately because of the round-offs originating at the scaling, but these distortions do not lead to considerable losses of



quality.

Here it is necessary to note that the use of not compressed digital image as an secret message is an atypical solution because digital images differ with high redundancy. However, in the case of QTAR algorithm, the given solution is reasonable as the mentioned redundancy of images allows one to avoid considerable distortions during the scaling. Besides, restoring of the cover image from the modified DCT spectrum requires a round-off at the transition from real values to integer pixels, which leads to additional distortions of the embedded data.

Fig. 2 shows the partition of image “Lenna” onto homogeneous blocks with a quad-tree and the selection of embedding areas for various values of the threshold of homogeneity  $Th$ . It is accepted that the minimum block size in each case equals to 8.

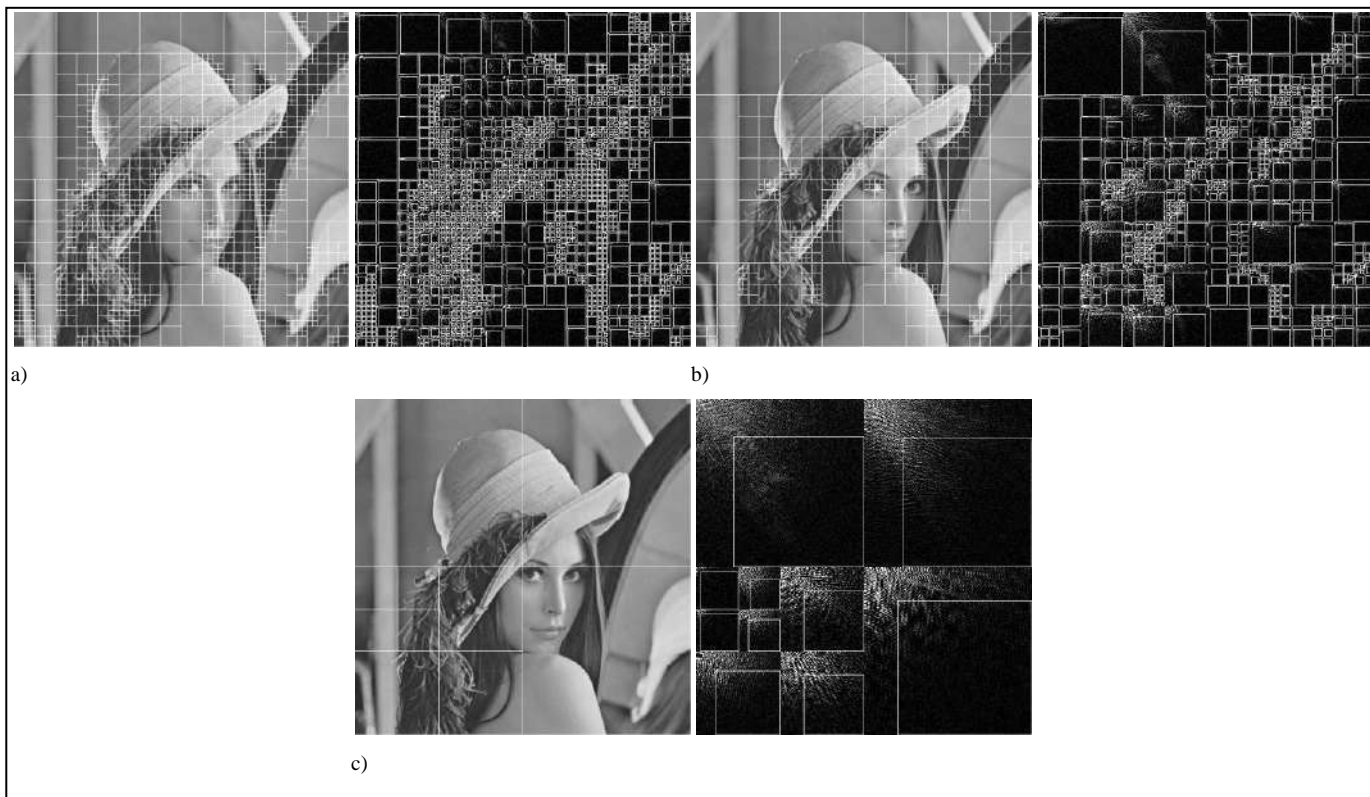


Fig. 2. Partition of the image “Lenna” on blocks: a)  $Th = 0,4$ ; a)  $Th = 0,6$ ; a)  $Th = 0,8$ .

The increase of the homogeneity threshold of the block leads to quad-tree simplification. Thus, the capacity decreases. In the instances presented in Fig. 2, the image capacity “Lenna” equals to 5,77, 5,76 and 4,86 bits per pixel accordingly.

### 3.2. Possible ways of QTAR algorithm improvement

The QTAR algorithm considers the cover image as the initial square block of pixels. Coordinates of the top left corner of the given initial block are named as an index point and designated as  $(x, y)$ . In the initial algorithm the given point has coordinates  $(0, 0)$  and cannot be changed. However, if the digital image is presented in the form of torus, for example as in cellular automata models [16], it is possible to choose any point of the cover image as an index point. The index point modification will change the form of the quad-tree and will affect the distribution of parts of the secret image on the cover image blocks.

The example is shown in Fig. 3. The index point is marked white. Other parameters of the algorithm of quad-tree construction in the above-mentioned example coincide with the analogous parameters of the example shown in Fig. 2a.

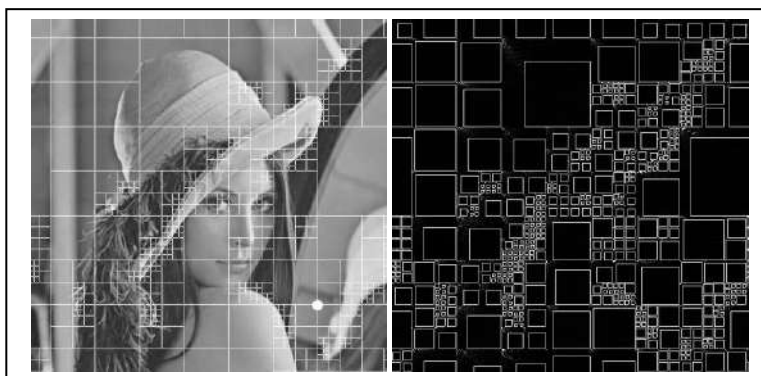


Fig. 3. Partition of the image “Lenna” on blocks with the modified index point ( $x = 412$ ;  $y = 412$ ).

It is possible to see that the index point modification changes the quad-tree form, and it leads to the modification of indexes of capacity and quality of embedding.

Fig. 4 shows the modification of the given indexes at the index point modification by the example of four various images: “F15”, “Clouds”, “Jellyfish”, and “House”. The image “Baboon” was a secret image in each case. The parameters of algorithm of quad-tree construction were set as follows:  $Th = 0,4$ ,  $m = 8$ . The embedding quality index is the peak signal-to-noise ratio (PSNR). The index of embedding capacity is the bits per pixel amount (BPP).

It is possible to see that the maximum capacity for the image “F15” is by default given by the index point, but the PSNR value can be increased approximately by 1,0 dB when maintaining capacity, which is essential. The capacity of embedding distinct from the maximum is given by the index point for the images “Clouds” and “Jellyfish”. Besides, in the case of the image “Clouds” the appropriate PSNR value is close to greatest possible one for the given capacity, and in case of the image “Jellyfish” the PSNR value can be essentially increased. The image “House” represents a cover image instance where the index point by default gives the greatest possible capacity and the PSNR value which is the greatest possible for the given capacity. However, in this case, it is possible to obtain quality improvement of embedding by 0,5–1,0 dB at the expense of insignificant decrease of capacity.

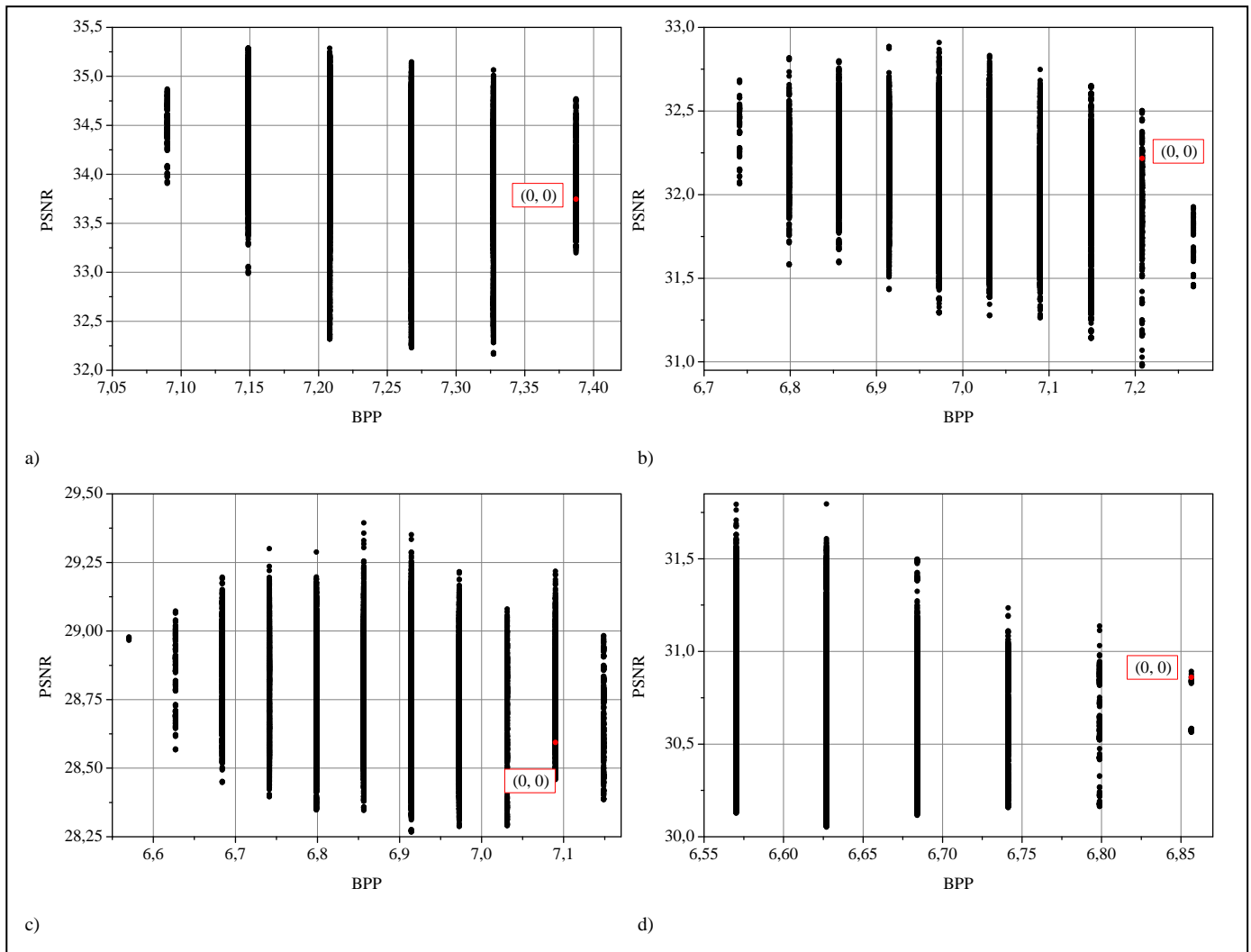


Fig. 4. Modification of PSNR and BPP at modification of the index point: a) for the image “F15”; b) for the image “Clouds”; c) for the image “Jellyfish”; d) for the image “House”.

The presented instances show that the index point modification allows us to increase value PSNR at equal or comparable BPP value, and on occasion to raise both given indexes.

The second possible approach to improvement of QTAR algorithm is connected to selecting of the threshold value. Since the brightness of an image makes essential impact on perception of the given image by human sight, in the present paper it is offered to introduce different threshold values for blocks of the image with different brightness. For this purpose, let us divide all brightness range of pixels on three equal sub-bands  $[0, 255] = [0, 85] \quad [85, 170] \quad [170, 255]$  define the threshold value of block homogeneity for each part and designate them as  $Th_1$ ,  $Th_2$ ,  $Th_3$  accordingly.

The instance is shown in Fig. 5. For the cover image “House” and the secret image “Baboon” the introduced threshold values were set as follows:  $Th_1 = 0,9$ ,  $Th_2 = 0,1$ ,  $Th_3 = 0,4$ . The graph shows the modification of PSNR and BPP indexes at the index point modification by analogy with the previous instances. It is possible to see that the given graph differs from the graph shown

in Fig. 4d and is obtained for the same pair of images. Pareto frontier was displaced: the range of capacity values slightly displaced towards decrease, and maximum PSNR value increased.

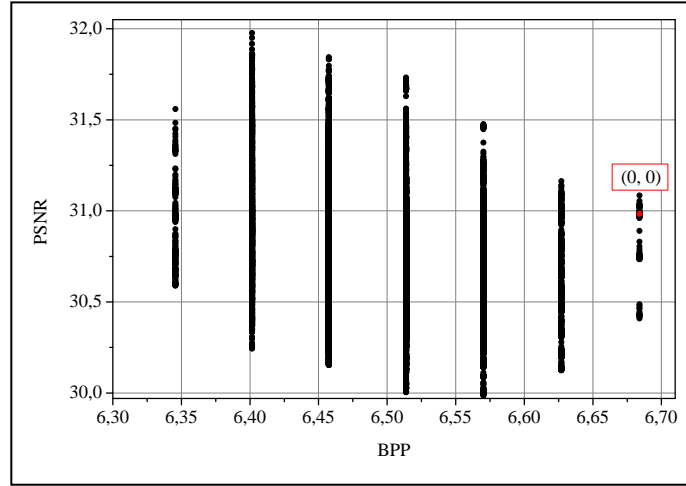


Fig. 5. Modification PSNR and BPP at an index point modification ( $Th_1 = 0,9$ ;  $Th_2 = 0,1$ ;  $Th_3 = 0,4$ ;  $m = 8$ ).

Thus, all given instances obviously confirm that the parameters introduced in the present paper make essential impact on effectiveness of embedding process. Besides, the given parameters can be used as the additional key information.

### 3.3. The offered improved algorithm

Exhaustive search of every possible value of an index point and homogeneity threshold of blocks is inconvenient, since it requires a great number of calculations. Therefore in the present research differential evolution (DE) is used for the solution of the given problem. It is the known metaheuristics widely used for solving the problems of optimization in various application areas, including digital steganography [17]. It allows to optimize sets of real heterogeneous parameters.

Since DE is a well-known optimization method, it is not described in the present article. Let us only mention that the DE algorithm operates with the following parameters: the size of population  $N$ , mutation coefficient  $F$ , probability of crossing over  $CR$ , number of calculations of objective function  $K$ .

Objective function is defined by the following formula:

$$f = \begin{cases} \frac{PSNR - PSNR^{QTAR}}{PSNR^{QTAR}} + \frac{BPP - BPP^{QTAR}}{BPP^{QTAR}}, & \text{if } PSNR \geq PSNR^{QTAR} \text{ and } BPP \geq BPP^{QTAR}, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where  $PSNR^{QTAR}$  and  $BPP^{QTAR}$  are values of efficiency indexes at embedding according to the initial QTAR algorithm.

Then the new algorithm of high-capacity embedding of the information in the frequency domain of discrete cosine transform of digital images on the basis of algorithm QTAR can be represented as follows:

#### Input:

Square cover image  $I$ ; secret image  $S$ ; minimal block size  $m$ ; matrix of quantization of the size  $8 \times 8$   $Q$ ; scale factor  $k$ ; parameters of DE algorithm.

#### Output:

Cover image containing the secret image  $I'$ .

**Step 1.** To execute the scaling of the secret image pixels by formula  $\tilde{s}_i = k/255 s_i$ .

**Step 2.** To build in the secret image  $S$  into the cover image  $I$  being the QTAR algorithm. To record the received values of quality indexes and embedding capacity as  $PSNR^{QTAR}$  and  $BPP^{QTAR}$ . To calculate the value of objective function by formula (1) and to record it as  $f^{\max}$ .

**Step 3.** To generate  $N$  vectors of form  $\mathbf{x}^i = (x, y, Th_1, Th_2, Th_3)$ ,  $i = \overline{1, N}$ .

**Step 4.** For  $i = \overline{1, N}$  to execute the following:

**Step 4.1.** To represent the cover image in the form of a quad-tree consisting of  $M^i$  blocks of pixels, using the vector of parameters  $\mathbf{x}^i = (x, y, Th_1, Th_2, Th_3)$ .

**Step 4.2.** For  $j = \overline{1, M^i}$  to execute the following:

**Step 4.2.1.** To execute the two-dimensional DCT of the  $j$ -th block of pixels with the size  $m_j \times m_j$ .

**Step 4.2.2.** To expand the matrix  $\mathbf{Q}$  to the extent of  $m_j \times m_j$  using interpolation; and to divide the DCT-coefficients of a block into elements of the given matrix with the subsequent round-off.

**Step 4.2.3.** To select a square area of the greatest possible size  $n_j \times n_j$ , consisting only of nulls in the right lower angle of each block of quantized DCT-coefficients.

**Step 4.2.4.** In the initial block of DCT-coefficients (before quantization) to substitute the area of embedding  $n_j^2$  with pixels of the modified secret image  $\tilde{S}$ .

**Step 4.2.5.** To execute the inverse two-dimensional DCT.

**Step 4.3.** To calculate values of quality indexes and capacity of embedding  $\text{PSNR}^i$  and  $\text{BPP}^i$ , and to calculate the value of objective function  $f^i$  by formula (1).

**Step 4.4.** If  $f^i > f^{\max}$ , then to assign  $f^{\max} = f^i$  and to record the vector  $\mathbf{x}^i$  as the best solution  $\mathbf{x}^{\text{best}}$ .

**Step 5.** To renew the population by rules of differential evolution.

**Step 6.** If the amount of evaluations of objective function does not exceed  $K$ , then to pass to step 4. Otherwise to pass to step 7.

**Step 7.** To build in secret image  $S$  into cover image  $I$  using the vector of parameters  $\mathbf{x}^{\text{best}}$ , then return stego image  $I'$  and key sequence  $(n_1, n_2, \dots, n_M, x, y, Th_1, Th_2, Th_3)$  and complete the algorithm.

The extraction algorithm of the secret message is as follows.

**Input:**

stego image  $I'$ ; key sequence  $(n_1, n_2, \dots, n_M, x, y, Th_1, Th_2, Th_3)$ ; minimum block size  $m$ ; scale factor  $k$ .

**Output:**

extracted secret image  $S'$ .

**Step 1.** To represent the stego image in the form of a quad-tree out of  $M$  blocks of pixels with the size not less than  $m \times m$  pixels with the index point  $(x, y)$  and the threshold values  $Th_1, Th_2, Th_3$ .

**Step 2.** For  $j = \overline{1, M}$  to execute the following:

**Step 2.1.** To execute two-dimensional DCT of  $j$ -th block of pixels with the size of  $m_j \times m_j$ .

**Step 2.2.** To select in the right lower angle of the received block of DCT factors a square block of embedded data elements with the side  $n_j$ .

**Step 2.3.** To execute an inverse scaling of the elements of the selected block using the formula  $s'_p = \frac{255}{k} \tilde{s}_p$ ,  $p = \overline{1, n_j^2}$  in order to derive the block of pixels of the secret image.

**Step 3.** To restore secret image  $S'$  from separate blocks of pixels.

**Step 4.** To return extracted secret image  $S'$  and to complete the algorithm.

In the following section of the present article, the results of computing experiments with the given algorithm and its comparison to the QTAR algorithm are presented.

#### 4. Results of experiments and their discussion

Computing experiments with the QTAR algorithm and the offered algorithm were carried out on the test sampling including 19 grey-scale and 3 full-color images with the resolution of  $512 \times 512$  of pixels. The given sampling was formed from base of images [18]. The examples of test images are shown on Fig. 6.



Fig. 6. Examples of test images.

Fig. 7 shows the results of three experiments with the obtained algorithm for various values of embedding parameters. The images have the following order in each line:

- the cover image;
- the cover image divided into square homogeneous blocks;
- the stego image;
- the secret image extracted after embedding.

It is possible to see that stego image does not contain appreciable artefacts of embedding in each case. The secret image contains some distortions with the level comprehensible to human perception.

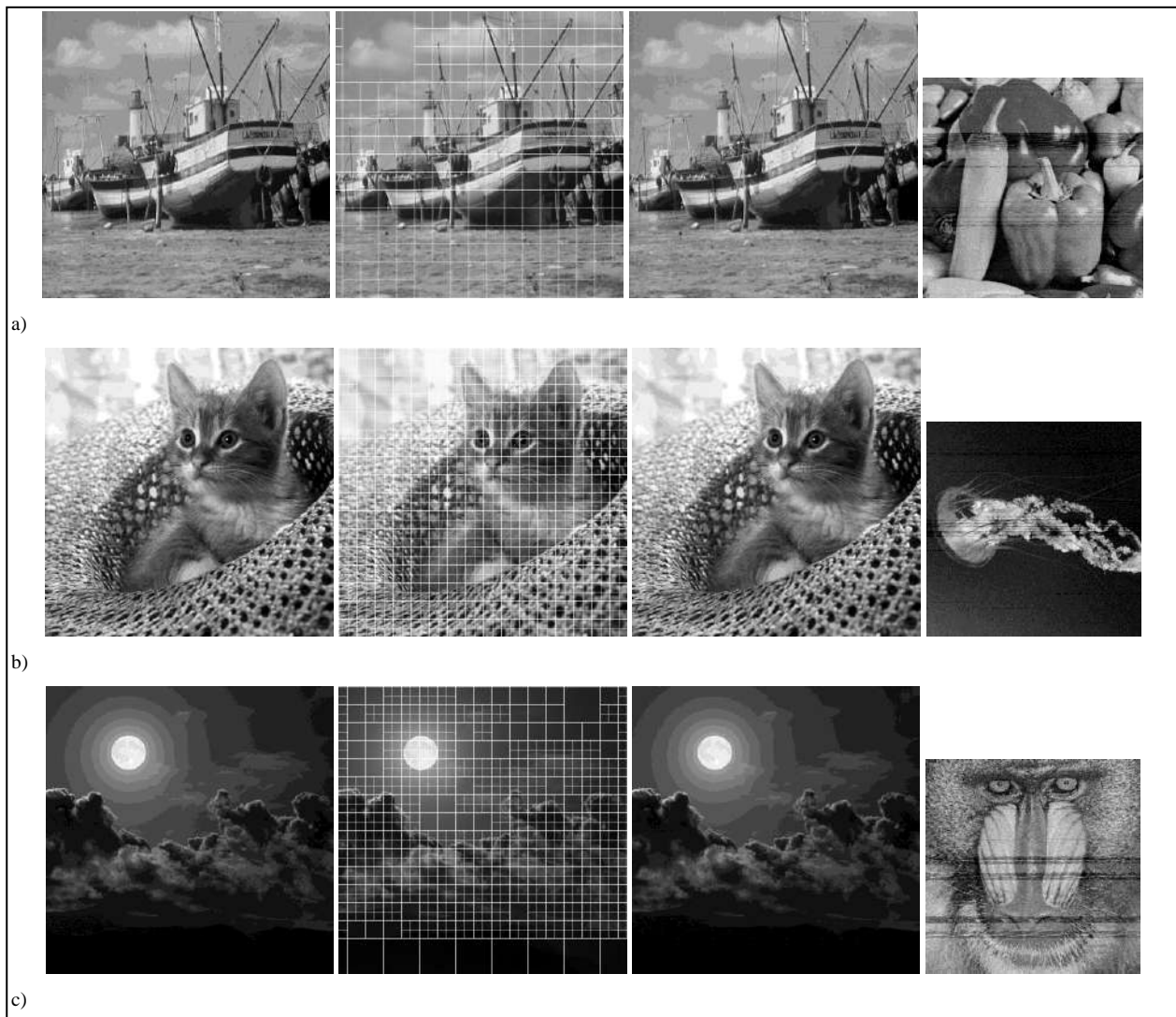


Fig. 7. Experiments: a) the image “Peppers” is embedded in the image “Boat”,  $m = 32$ ;  $k = 10$ ;  $(x, y) = (269, 0)$ ;  $Th_1 = 0,41$ ;  $Th_2 = 0,42$ ;  $Th_3 = 0,28$ ; b) the image “Jellyfish” is embedded in the image “Cat”,  $m = 16$ ;  $k = 10$ ;  $(x, y) = (96, 400)$ ;  $Th_1 = 0,02$ ;  $Th_2 = 0,11$ ;  $Th_3 = 0,26$ ; c) the image “Baboon” is embedded in the image “Clouds”,  $m = 16$ ;  $k = 4$ ;  $(x, y) = (17, 192)$ ;  $Th_1 = 0,07$ ;  $Th_2 = 0,13$ ;  $Th_3 = 0,38$ .

Table 1 shows the results of the efficiency estimation of our algorithm in comparison with the initial QTAR algorithm. For each image, the optimal parameters of embedding are specified, which were found by means of differential evolution. Parameters of differential evolution were set according to the guidelines presented in [19] for the optimization problem of dimensions 5. Last three table lines correspond to full-color images; the rest of the images are grey-scale.

One can see that in most cases our algorithm surpasses the QTAR algorithm and only on occasion shows comparable results. For example, it refers to images “Arctichare”, “Clouds”, “F15”. In those cases the proposed algorithm cannot reach substantial improvement. Also, if one parameter increases, other parameters decrease. It is possible to explain in the following way: the point by default  $(0, 0)$  for the given images gives the solution that belongs to Pareto-frontier. For images “Cat”, “Peppers”, “Baboon” the proposed algorithm noticeably surpasses QTAR in terms of BPP at the comparable value of PSNR. But for the majority of images the proposed algorithm surpasses the QTAR algorithm in terms of both considered indexes. As a result, the maximum advantage of PSNR over the best value of BPP is 1,07 dB, and the maximum advantage of BPP over comparable value of PSNR is 1,1 bits, which is significant improvement.

Regarding the stability against steganalysis, the offered algorithm also surpasses the QTAR algorithm, since the embedding operation in both cases is the same, but additional parameters used by the offered algorithm increase the private key size.

Table 1. Comparison of the proposed algorithm and the QTAR algorithm.

Image title	QTAR		Proposed algorithm					
	PSNR, dB	BPP	PSNR, dB	BPP	Index point	$Th_1$	$Th_2$	$Th_3$
Arctichare	41,63	<b>5,85</b>	<b>44,02</b>	4,02	(184, 40)	0,48	0,66	0,94
Baboon	32,99	3,02	<b>33,00</b>	<b>4,11</b>	(129, 88)	0,97	0,23	0,21
Barbara	26,86	5,03	<b>27,06</b>	<b>5,11</b>	(455, 448)	0,34	0,16	0,12
Boat	<b>33,11</b>	4,88	<b>33,11</b>	<b>4,98</b>	(448, 313)	0,23	0,18	0,84
Cameraman	37,58	5,53	<b>37,72</b>	<b>5,62</b>	(335, 402)	0,10	0,05	0,44
Cat	34,20	4,34	<b>35,27</b>	<b>4,38</b>	(192, 128)	0,16	0,89	0,34
Clouds	37,53	<b>5,64</b>	<b>38,46</b>	4,57	(432, 123)	0,78	0,84	0,07
Darkhair	37,16	5,94	<b>37,21</b>	<b>5,99</b>	(284, 256)	0,14	0,65	0,47
Fruits	34,59	5,16	<b>34,61</b>	<b>5,31</b>	(160, 178)	0,89	0,12	0,43
House	42,37	5,80	<b>42,76</b>	<b>5,80</b>	(384, 256)	0,90	0,40	0,20
Jellyfish	36,92	5,80	<b>37,31</b>	<b>6,26</b>	(480, 384)	0,07	0,79	0,94
Jetplane	36,15	5,03	<b>36,23</b>	<b>5,36</b>	(124, 208)	0,39	0,60	0,11
Lake	33,86	4,67	<b>33,92</b>	<b>4,83</b>	(5, 275)	0,00	0,54	0,14
Livingroom	34,51	4,52	<b>34,60</b>	<b>4,74</b>	(455, 185)	0,08	0,18	0,72
Peppers	33,99	4,33	<b>34,00</b>	<b>5,43</b>	(402, 234)	0,63	0,18	0,15
Pirate	33,24	4,81	<b>33,25</b>	<b>4,88</b>	(131, 112)	0,20	0,38	0,52
Sails	30,62	3,85	<b>30,64</b>	<b>3,89</b>	(208, 446)	0,14	0,35	0,69
Tiffany	32,01	5,21	<b>32,09</b>	<b>5,36</b>	(497, 317)	0,11	0,24	0,16
Walkbridge	31,16	3,85	<b>31,19</b>	<b>3,89</b>	(288, 152)	0,53	0,04	0,16
F15	38,59	<b>17,64</b>	<b>39,69</b>	14,65	(193, 353)	0,42	0,70	0,84
Lenna	33,37	16,00	<b>33,39</b>	<b>16,69</b>	(191, 384)	0,26	0,21	0,22
Tiger	36,26	17,01	<b>36,76</b>	<b>17,72</b>	(263, 304)	0,93	0,72	0,07

## 5. Conclusion

The given paper presents the new algorithm of high-capacity embedding of the information into the frequency domain of discrete cosine transform received on the basis of known QTAR algorithm [15]. The QTAR algorithm ensures high capacity of embedding at the expense of representation of the cover image in the form of a quad-tree of homogeneous blocks of pixels. The frequency spectrum of such blocks contains a small number of significant elements and high number of insignificant elements. Replacement of insignificant frequency coefficients with data elements of the secret image does not lead to appreciable distortions of the cover image.

A distinctive feature of the offered modification of the QTAR algorithm is a new approach to representation of the cover image in the form of a quad-tree of homogeneous blocks of pixels. New parameters are introduced into algorithm of quad-tree construction. The cover image is represented in the form of torus which allows one to arbitrarily choose an index point that corresponds to the left top angle in the initial algorithm. Besides, for blocks of pixels with various levels of brightness, the different threshold values defining the homogeneity criterion are set.

Deriving of optimal parameters for each concrete cover image is carried out by means of differential evolution.

Computing experiments have shown that the offered algorithm differs with greater effectiveness on quality and embedding capacity in comparison with QTAR algorithm.

Development of the given paper will consist in the search of new approaches to partition of the cover image into homogeneous blocks of pixels and synthesis of new algorithms of embedding.

Besides, the transfer of the initial approach to the achievement of high-capacity embedding on other transforms applied in digital image processing, except discrete cosine transform, is interesting.

## Acknowledgements

The given paper is completed with the support of the Ministry of Education and Science of the Russian Federation within the limits of the project part of the state assignment of TUSUR in 2017 and 2019 (project 2.3583.2017/4.6) and of the Russian Foundation for Basic Research (project 16-47-700350\_r\_a).

## References

- [1] Fridrich J. Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge: Cambridge University Press, 2010; 437 p.
- [2] Salomon D. Data Compression: the Complete Reference, 4th Edition. London: Springer-Verlag, 2007; 1092 p.
- [3] Cedillo-Hernandez M, Garcia-Ugalde F, Nakano-Miyatake M, Perez-Meana H. Robust watermarking method in DFT domain for effective management of medical imaging. Signal, Image and Video Processing 2015; 9(5): 1163–1178.
- [4] Poljicak A, Mandic L, Agic D. Discrete Fourier transform-based watermarking method with an optimal implementation radius. J Electron Imaging 2011; 20(3): 033008-1–033008-8.
- [5] Ridzon R, Levicky D. Content protection in grayscale and color images based on robust digital watermarking. Telecommun Syst. 2013; 52(3): 1617–1631.
- [6] Zhao J, Koch E. Embedding robust labels into images for copyright protection. Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies (KnowRight'95). Austria, Vienna, 1995: 242–251.

- [7] Chen B, Wornell GW. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Trans Inf Theory* 2001; 47(4): 1423–1443.
- [8] Ejaz N, Anwar M, Ishtiaq SW, Baik. Adaptive image data hiding using transformation and error replacement. *Multimed Tools Appl*. 2013; 73(2): 825–840.
- [9] Hassanien AE. Hiding iris data for authentication of digital images using wavelet theory. *Pattern Recognit Image Anal* 2006; 16(4): 637–643.
- [10] Benrouma O, Hermassi H, Belghith S. Tamper detection and self-recovery scheme by DWT watermarking. *Nonlinear Dyn* 2015; 79(3): 1817–1833.
- [11] Chen ST, Huang HN, Kung WM, Hsu CY. Optimization-based image watermarking with integrated quantization embedding in the wavelet-domain. *Multimed Tools Appl* 2016; 75(10): 5493–5511.
- [12] Pakdaman Z, Saryazdi S, Nezamabadi-pour H. A prediction based reversible image watermarking in Hadamard domain. *Multimed Tools Appl* 2016; 1–29.
- [13] Rabie T, Kamel I. On the embedding limits of the discrete cosine transform. *Multimed Tools Appl* 2016; 75(10): 5939–5957.
- [14] Rabie T, Kamel I. High-capacity steganography: a global-adaptive-region discrete cosine transform approach. *Multimed Tools Appl* 2016: 1–21.
- [15] Rabie T, Kamel I. Toward optimal embedding capacity for transform domain steganography: a quad-tree adaptive-region approach. *Multimed Tools Appl* 2016: 1–24.
- [16] Evsutin OO. Research of the discrete orthogonal transformation received with use the dynamics of cellular automata. *Computer Optics* 2014; 38(2): 314–321.
- [17] Huang HC, Chang FC, Chen YH, Chu SC. Survey of bio-inspired computing for information hiding. *Journal of Information Hiding and Multimedia Signal Processing* 2015; 6(3): 430–443.
- [18] Image Databases. URL: [http://www.imageprocessingplace.com/root\\_files\\_V3/image\\_databases.htm](http://www.imageprocessingplace.com/root_files_V3/image_databases.htm) (02.02.2017).
- [19] Pedersen MEH. Good parameters for differential evolution. Technical Report no. HL1002, Hvas Laboratories, 2010.

# A model for data hiding system description

Victor Fedoseev<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

The paper presents a new model for unified description of any information hiding systems which include both steganographic and watermarking systems. The model is based on considering three possible representations of information being embedded: a binary vector, a digital signal, and a feature matrix. Also we introduce a parametric description for information hiding systems according to the proposed model which completely defines all valuable algorithms used at the embedding and the extraction stages, as well as its parameters. Some examples of such descriptions a number of existing systems are presented.

*Keywords:* information hiding; data hiding; digital watermarking; watermarking system; steganography; steganographic system

---

## 1. Introduction

The paper is devoted to the development of a model for the unified description of information hiding systems (also called data hiding systems). Such systems embed secret or protective data into a digital signal (image, video, audio etc.) and include watermarking and steganographic systems [1]. Information hiding techniques were investigated by Ingemar Cox [1-3], Jessica Fridrich [3,4], Mauro Barni, Franco Bartolini [5], Fabien Petitcolas [6,7], Stefan Katzenbeisser [6], Eric Cole [8], Birgit Pfizmann [9] and others. They defined a common terminology, described the basic structural components and properties of information hiding systems. However, we can note the lack of a generally accepted model for a unified description of such systems. As a result, it is difficult to compare different systems and select the most appropriate system for a particular task.

Earlier, several models were proposed in papers [1, 3, 5, 6, 10-20] but each of them has a number of shortcomings, which do not allow them to be used for a complete description of information hiding system processes:

1. All the cited models except [16, 17] can describe either watermarking systems or steganographic systems. The more general case is not considered.
2. The models [1, 3, 6, 10-16, 18-19] do not determine such important details as analysis of the host asset, information encoding and decoding and others.
3. The models [1, 3, 5, 10-15, 17] do not take into account all possible inputs and outputs of the information hiding system.
4. The models [6, 10, 11, 15] are limited to the description of internal processes and do not allow to determine the external properties of systems.
5. Some models [11, 18-20] are intended only for media of a certain type (usually, for audio signals or images).
6. None of the above models have become a universally recognized standard.

In this paper, we propose a universal mathematical model for information hiding systems, which can describe all components of information hiding systems, and which is free of the shortcomings listed above.

## 2. The proposed model of information hiding system (MIHS)

### 2.1. Basic concepts

In the proposed model, we define *information hiding system* (IHS) as a set of data and processes (functions) of their processing. One of the most important concepts in this model is *internal information* that is the information embedded in the *host asset*.

In our model, we introduce three equivalent forms of internal information: *a binary vector*, *a digital signal*, and *a feature matrix*. The first form corresponds, for example, to a message transmitted via a steganographic channel, or to a digital code of a protective watermark. The second form coincides with the traditional form of the host asset (digital audio, image, video, etc.). The embedding itself proceeds in the third form, which is individual for each system. In each particular IHS, the internal information can be converted from one form to another.

We will use the following designations:

- $\mathbb{B}^n = \mathbb{N}_0 \cap [0.2^n - 1]$  is a set of n-bit nonnegative integers. A special case is a set  $\mathbb{B} = \mathbb{B}^1 = \{0,1\}$ .
- $\mathbb{S}_{[N_1 \times N_2 \times \dots \times N_m]}^m$  is an m-dimensional matrix of size  $N_1 \times N_2 \times \dots \times N_m$  composed of elements of a certain numerical set  $\mathbb{S}$ .
- $\mathbb{S}^m$  is an m-dimensional matrix of unknown size composed of elements of a certain numerical set  $\mathbb{S}$  (used when the matrix sizes are not important in the current context).

The introduced sets allow us to define the sets corresponding to the three above-mentioned forms of internal information. Thus, the first form of a binary vector corresponds to the set  $\mathbb{B}_{[N_b]}^1$ , where  $N_b$  is a vector length. Then, a multidimensional digital signal will be defined as  $X \in \mathbb{X}^m$  that is an m-dimensional matrix composed of elements of a set  $\mathbb{X} \subseteq \mathbb{R}$ . The set  $\mathbb{X}^m$



will be called as *the set of digital signals*. Finally, a feature matrix  $y \in \mathbb{Y}^1$  is an  $m$ -dimensional matrix composed of elements of a set  $\mathbb{Y} \subseteq \mathbb{C}$ .  $\mathbb{Y}^1$  will be called as *feature set*.

## 2.2. Main elements of the model

Let  $C \in \mathbb{X}^m$  be a host asset and  $C^W \in \mathbb{X}^m$  be an *information carrier* (an asset with embedded information). After its transmission, it can change due to distortions in the channel and possible attacks. Therefore, we will use another notation for the *received information carrier*  $\widetilde{C^W} \in \mathbb{X}^m$ .

The next important element of any system is the *composite key*  $\mathbf{k} \in K$ . It comprises the *secret key*  $k^s \in K^s \subseteq \mathbb{B}_{[N_k]}^1$ , which provides security of the system, and *public parameters*  $k^p \in K^p$  of functions and algorithms:  $\mathbf{k} = (k^s, k^p)$ . We will not specify the structure of the set  $K^p$  for the general model. It can be defined for particular systems.

For internal information, the following designations will be used:  $\mathbf{b}, \mathbf{b}^R \in \mathbb{B}_{[N_b]}^1$  (in the form of a binary vector);  $W, W^R \in \mathbb{X}^m$  (in the signal form);  $\Omega, \tilde{\Omega} \in \mathbb{Y}^1$  (in the form of feature matrix). The names of these and other structures are given in Table 1. Also, it is necessary to define the concept of *initial form* of internal information that is either  $\mathbb{B}_{[N_b]}^1$  or  $\mathbb{X}^m$  depending on the particular system.

Table 1. List of notations used in the model of information hiding system.

Data	Set	Name
$C$	$\mathbb{X}_0^m$	Host asset
$\mathbf{b}$	$\mathbb{B}_{[N_b]}^1$	Embedded information (internal information form)
$W$	$\mathbb{X}_0^m$	Embedded signal (internal information form)
$C^m$	$\mathbb{X}_0^m$	Information carrier (or filled asset)
$\widetilde{C^m}$	$\mathbb{X}_0^m$	Received information carrier
$k^s$	$K^s \subseteq \mathbb{B}_{[N_k]}^1$	Secret key
$k^p$	$K^p$	Public parameters
$\mathbf{k}$	$K = K^s \times K^p$	Composite key
$\mathbf{b}^R$	$\mathbb{B}_{[N_b]}^1$	Extracted information (internal information form)
$W^R$	$\mathbb{X}_0^m$	Extracted signal (internal information form)
$\xi$	$\mathbb{B}$	Detection result
$k^c$	$K^c$	Host asset parameters
$\widetilde{k^c}$	$K^c$	Estimated host asset parameters
$\Omega$	$\mathbb{Y}_0^1$	Embedded information feature matrix (internal information form)
$\tilde{\Omega}$	$\mathbb{Y}_0^1$	Extracted information feature matrix (internal information form)
$f$	$\mathbb{Y}_0^1$	Host asset feature matrix
$f^m$	$\mathbb{Y}_0^1$	Information carrier feature matrix
$\widetilde{f^m}$	$\mathbb{Y}_0^1$	Received information carrier feature matrix

We will use the three following functions to describe possible transformations of the internal information:

- *encoding function in signal space*

$$\mathcal{P} : \mathbb{B}_{[N_b]}^1 \times K \mapsto \mathbb{X}_0^m, \quad (1)$$

- *encoding function in feature space*

$$\mathcal{P}_f : \mathbb{B}_{[N_b]}^1 \times K \mapsto \mathbb{Y}_0^1, \quad (2)$$

- *signal-to-feature transformation function*, which most often has the form

$$\mathcal{F} : \mathbb{X}_0^m \mapsto \mathbb{Y}_0^1, \quad (3)$$

and rarely

$$\mathcal{F} : \mathbb{X}_0^m \mapsto \mathbb{Y}_0^1 \times \Psi, \quad (4)$$

along with the inverse functions  $\mathcal{P}^{-1}, \mathcal{P}_f^{-1}, \mathcal{F}^{-1}$ . The relationship between the various internal information forms is shown in Fig. 1.

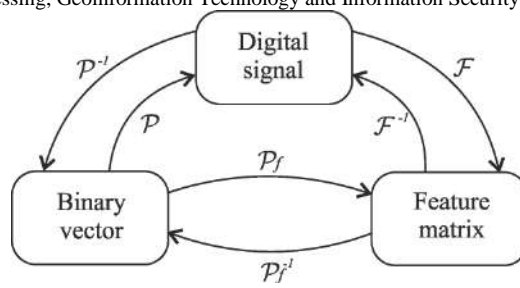


Fig. 1. The relationship between the various internal information forms.

Table 2 shows, which forms of internal information can be used at the particular stages of system operation. The presence of various options in some rows of Table 2 is explained by the differences in the systems. For one particular system, only one form is possible at each stage. It should be noted that in the last row one more option of the system output is possible: a binary value  $\xi \in \mathbb{B}$ , reflecting the result of internal information detection. We will consider this case later in more details.

Table 2. Possible forms of internal information.

Data processing stage	Input	Internal information form		
		Binary vector from $\mathbb{B}_{[N_b]}^1$	Digital signal from $X^m$	Feature matrix from $Y^l$
	Input	✓	✓	
	Internal information embedding			✓
	Information transmission within an asset		✓	
	Internal information detection	✓	✓	✓
	Output	✓	✓	

As noted above, the form of feature matrix is defined for all information hiding systems because it is used at the embedding stage. But this form is not used at the input. Therefore, at least one of two other forms should be determined. Some systems operate all three internal information forms. In order to define the used internal information forms, we use the following binary predicates:

$$\pi_{b_{in}} = \begin{cases} true, & \text{if the initial form is } \mathbb{B}_{[N_b]}^1, \\ false, & \text{if the initial form is } X_0^m. \end{cases} \quad (5)$$

$$\pi_P = \begin{cases} true, & \text{when coding to } X_0^m, \\ false, & \text{when coding to } Y_0^l. \end{cases} \quad (6)$$

The first one defines the initial form, while the other one defines the encoding method.

Fig. 2 shows the general flowchart of information hiding system according to the proposed model. The flowchart highlights the embedding and extraction subsystems, as well as the data transmission channel. Here and later (in Fig. 3-5), arrows indicate data streams, and rectangles indicate data processing processes. Solid arrows indicate mandatory data streams existing in all systems, and dashed – the optional ones. Circles mark merging data streams, while rhombuses mark branching ones. Rectangles with double borders mark processes consisting of several subprocesses.

Fig. 3 describes subprocesses of the composite embedding information process outlined in the general flowchart in Fig. 2. Similarly, Fig. 4 describes the contents of the composite information extraction process and Fig. 5 s the block of internal information processing.

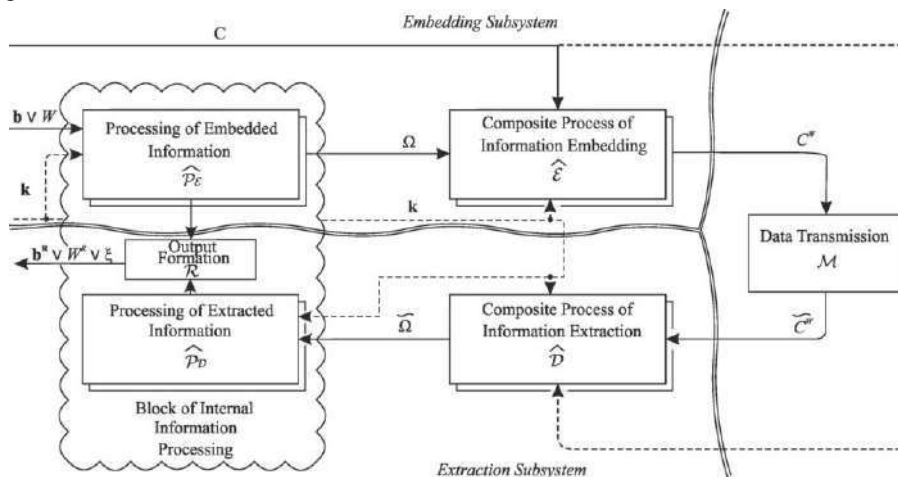


Fig. 2. The general information hiding system workflow.

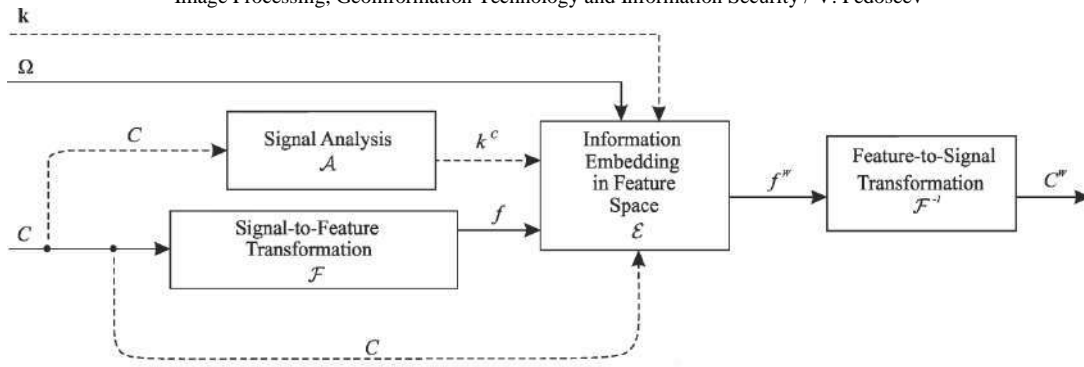


Fig. 3. Details of the composite process of information embedding.

Let us describe the general flowchart of IHS (Fig. 2). The input of any system includes a host asset  $C$ , an internal information in the form of  $\mathbf{b}$  or  $W$ , as well as a key  $\mathbf{k}$ . Then, at the preliminary stage (before embedding), the internal information is transformed into a feature matrix  $\Omega$ . The obtained matrix along with the host asset is fed to the input of the composite process of information embedding resulting in the information carrier  $C^W$ . Then, it is transferred to the extraction subsystem with possible distortions. Further, the received information carrier  $\widetilde{C}^W$  enters the input of the composite information extraction process (along with it, the original container transmitted through any closed channel can also be used in this block). The result of this stage is  $\widetilde{\Omega}$ . Finally, the system output is generated, which can be the extracted information  $\mathbf{b}^R$ , the extracted signal  $W^R$ , or the *detection result*  $\xi \in \mathbb{B}$ :

$$\xi = \begin{cases} 1, & \text{if } \widetilde{C}^W \text{ contains } \mathbf{b} \text{ (or } W), \\ 0, & \text{if } \widetilde{C}^W \text{ does not contain } \mathbf{b} \text{ (or } W). \end{cases} \quad (7)$$

The diagrams in Fig. 2-5 allow us to easily determine the form of the functions corresponding to individual processes. For example, according to the general flowchart (Fig. 2), the composite process of information embedding can be described by functions of the following types (depending on the use of the key):

$$\begin{aligned} \widehat{\mathcal{E}} &: \mathbb{X}_0^m \times \mathbb{Y}_0^l \times K \mapsto \mathbb{X}_0^m, \quad C^W = \widehat{\mathcal{E}}(C, \Omega, \mathbf{k}), \\ \widehat{\mathcal{E}} &: \mathbb{X}_0^m \times \mathbb{Y}_0^l \mapsto \mathbb{X}_0^m, \quad C^W = \widehat{\mathcal{E}}(C, \Omega). \end{aligned}$$

Similarly, there are four options for a composite information extraction process:

$$\begin{aligned} \widehat{\mathcal{D}} &: \mathbb{X}_0^m \times \mathbb{X}_0^m \times K \mapsto \mathbb{Y}_0^l, \quad \widetilde{\Omega} = \widehat{\mathcal{D}}(\widetilde{C}^W, C, \mathbf{k}), \\ \widehat{\mathcal{D}} &: \mathbb{X}_0^m \times \mathbb{X}_0^m \mapsto \mathbb{Y}_0^l, \quad \widetilde{\Omega} = \widehat{\mathcal{D}}(\widetilde{C}^W, C), \\ \widehat{\mathcal{D}} &: \mathbb{X}_0^m \times K \mapsto \mathbb{Y}_0^l, \quad \widetilde{\Omega} = \widehat{\mathcal{D}}(\widetilde{C}^W, \mathbf{k}), \\ \widehat{\mathcal{D}} &: \mathbb{X}_0^m \mapsto \mathbb{Y}_0^l, \quad \widetilde{\Omega} = \widehat{\mathcal{D}}(\widetilde{C}^W). \end{aligned}$$

### 2.3. Specification of the composite processes

As shown in Fig. 3, the composite process of information embedding includes the following subprocesses:

- Optional signal analysis function  $A$  aimed to estimate host asset parameters,
- Transformation function  $F$  and its inverse function  $F^{-1}$ ,
- Information embedding in feature space  $\mathcal{E}$ .

Signal analysis refers to the process of evaluating some numerical characteristics  $k^C$  of the host asset. For example, analysis of the image asset can consist in finding the coordinates of its feature points, carried out with a corner detector.

Processes  $F$  and  $F^{-1}$  mentioned above, are designed respectively to convert signals to feature matrices for reverse transformation. The peculiarity of these processes is the possible use of a value  $\psi \in \Psi$  that is a part of the function  $F$  result and an additional argument of the function  $F^{-1}$ . We will call this value as the *feature matrix complement*. It is not used for data embedding but allows to perform the inverse transformation. If  $F$  is reversible (i.e., it is DFT or DWT transform) than  $\psi$  is not defined.

The last process in Fig. 3  $\mathcal{E}$  involves the actual information embedding, that is the merging of the matrices  $f$  and  $\Omega$  in a single matrix  $f^W$ .

The details of the composite information extraction process are easily understood by Fig. 4. We only note that the signal analysis at the extraction stage can be performed either by the host asset (if it is known in a particular system) or by the received information carrier  $\widetilde{C}^W$ . In the latter case, it results in a vector of estimated characteristics  $\widetilde{k}^C$ . The actual information extraction is performed in the process  $D$  resulting in the feature matrix of extracted information  $\widetilde{\Omega}$ .

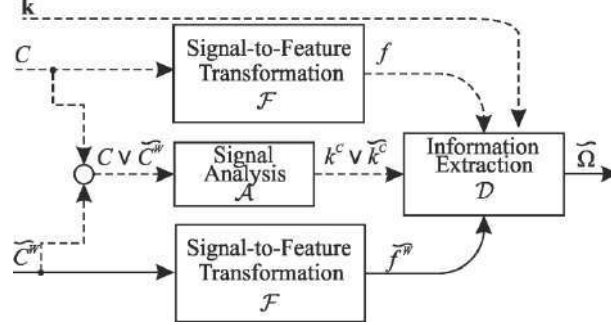


Fig. 4. Details of the composite process of information extraction.

Finally, the internal information processing block, shown in Fig. 5, includes the processes of its transformation from one form to another in both subsystems. For this, the previously introduced encoding-decoding functions  $P, P_f, P^{-1}, P_f^{-1}$  are used, and the particular configuration is determined by the two above mentioned predicates  $\pi_{bW}$  and  $\pi_p$ .

In addition to these processes, this block also includes a *detection function*  $\mathcal{R}$  operating in the extraction subsystem, which can have one of the following forms:

$$\mathcal{R} : \mathbb{B}_{[N_b]}^1 \times \mathbb{B}_{[N_b]}^1 \mapsto \mathbb{B}, \xi = \mathcal{R}(\mathbf{b}, \mathbf{b}^R), \quad (8)$$

$$\mathcal{R} : \mathbb{X}_0^m \times \mathbb{X}_0^m \mapsto \mathbb{B}, \xi = \mathcal{R}(W, W^R), \quad (9)$$

$$\mathcal{R} : \mathbb{Y}_0^l \times \mathbb{Y}_0^l \mapsto \mathbb{B}, \xi = \mathcal{R}(\Omega, \widetilde{\Omega}). \quad (10)$$

In all these cases,  $\mathcal{R}$  usually has the form of a threshold function:

$$\mathcal{R}(x, x^R) = \begin{cases} 1, & \rho(x, x^R) \geq T_p, \\ 0, & \rho(x, x^R) < T_p, \end{cases} \quad (11)$$

where  $x$  and  $x^R$  denote embedding and extracted information in the form used for the detection,  $T_p \in \mathbb{R}$  is the threshold, and  $\rho(x, x^R)$  is a function of the proximity of  $x$  and  $x^R$  determined individually for each particular system.

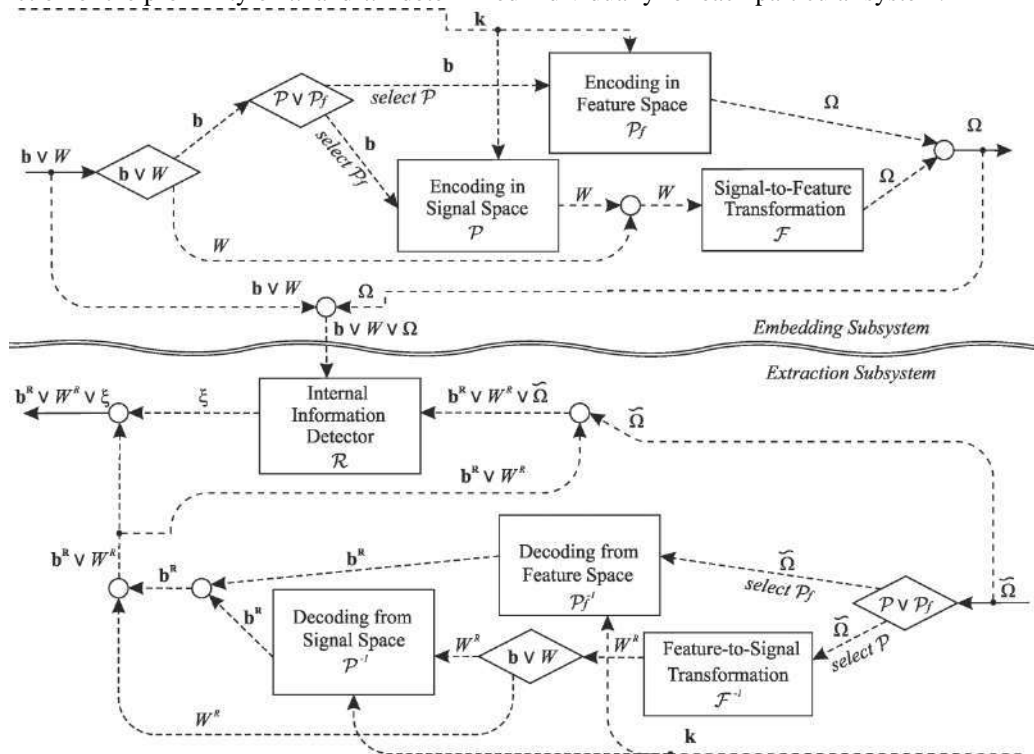


Fig. 5. Internal information processing.

The function  $r$  and the threshold  $T_p$  are determined at the system design. However, we can list general patterns:

- For systems resulting in  $\mathbf{b}^R$  or  $W^R$ , the detection form is the same as the initial form.
- For systems with the detection form  $\mathbb{B}_{[N_b]}^1$  the following function  $r$  is usually used:

$$\rho(\mathbf{b}, \mathbf{b}^R) = \frac{1}{N_b} \sum_{i=0}^{N_b-1} (1 - b_i \oplus b_i^R), \quad (12)$$

- For systems with the detection form  $\mathbb{X}^m$ , any conventional quality measure can be used as the function  $\rho$ . For example, for grayscale images belonging to the set  $\mathbb{X}^m = (\mathbb{B}^8)_{[N_1 \times N_2]}^2$ , PSNR values of two signals can be used [21]:

$$\rho(W, W^R) = PSNR(W, W^R) = 10 \lg \frac{255^2}{\varepsilon_{ke}^2(W, W^R)}, \quad (13)$$

where  $\varepsilon_{ke}^2(W, W^R)$  is a mean-square error.

- For systems with the detection form  $\mathbb{Y}^1$ ,  $r$  essentially depends on the structure of the set  $\mathbb{Y}^1$  itself. For instance, often the features reflect the energy characteristics, and therefore matrix elements with different indices can have different significance, in contrast to the pixels of digital signals.

### 3. Parametric description of information hiding systems

The developed model allows to make a unified description of any information hiding system by defining 14 parameters presented in Table 3. Moreover, this list can help for developing new systems by adopting some parameters from existing ones.

Also, in Table 3 we illustrate the ability of the proposed model to describe different systems. For that, we consider two examples of information hiding systems, which differ from each other in a number of components.

#### System 1: steganographic embedding into the least significant bits (LSB) of audio signals

In this system, a simple replacement of the lower bits of the signal is performed, according to the key and the bits of the secret message. For information extraction, the least significant bits are read at the specified positions. The system description is given in Table 3.

#### System 2: Phase image spectrum watermarking

In this simple system, the input data include a halftone host image and a watermark image with values  $\{0, \pm 1\}$  and the same size. Next, phase Fourier spectrum of the host image is calculated. Then, the phase components are replaced by non-zero values of the watermark pixels, previously mixed according to a secret key. For simplicity of the description, we define the mixing method as a cyclic shift to a vector  $\mathbf{k} = (k_1, k_2)$ . After the replacement, inverse Fourier transform is performed. When extracting information, the same transformations are performed to estimate the embedded watermark. Finally, the obtained estimation is compared with the initially embedded watermark in order answer the question of its presence in the given image. The description of this simple system is also provided in Table 3.

### 4. Conclusion

In this paper, we proposed a novel model designed for unified description of arbitrary information hiding systems, which include steganography systems and digital watermarking systems. It is based on the separation of the forms of internal information carried within the digital media. We described internal IHS processes, and also introduced a parametric description, which completely determines the existing watermarking and steganography algorithms, and also facilitates the synthesis of new systems. The applicability of this model is shown to describe two completely different information hiding systems.

### Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants 15-07-05576 and 16-41-630676) and by the Ministry of Education and Science of the Russian Federation by means of the Russian President's grant MK-1907.2017.9.

### References

- [1] Miller ML, Cox JJ, Linnartz J-PMG, Kalker T. A review of watermarking, principles and practices. Digital Signal Processing in Multimedia Systems 1999; 461–485.
- [2] Cox JJ, Miller ML, Bloom JA. Digital watermarking. Morgan Kaufmann Publishers, 2002; 568 p.
- [3] Cox JJ, Miller ML, Bloom JA, Fridrich J, Kalker T. Digital watermarking and steganography. USA: Elsevier, 2008; 587 p.
- [4] Fridrich J. Steganography in digital media: principles, algorithms, and applications. Cambridge University Press, 2010; 450 p.
- [5] Barni M, Bartolini F. Watermarking systems engineering. New-York: Marcel Dekker, 2004; 485 p.
- [6] Katzenbeisser S, Petitcolas FAP. Information hiding techniques for steganography and digital watermarking. Boston, London: Artech House, 2000; 237 p.
- [7] Petitcolas FAP, Anderson RJ, Kuhn MG. Information hiding – a survey. Proceedings of the IEEE 1999; 87(7): 1062–1078.
- [8] Cole E. Hiding in plain sight: steganography and the art of covert communication. Wiley Publishing, 2003; 362 p.
- [9] Pfitzmann B. Information hiding terminology: results of an informal plenary meeting and additional. Proceedings of the First International Workshop on Information Hiding 1996; 347–350.
- [10] Furht B, Muharemagic E, Socek D. Multimedia encryption and watermarking, Springer, 2006; 331 p.
- [11] Mohanty SP. Digital watermarking: a tutorial review. Bangalore, 1999.
- [12] Cohen AS, Lapidoth A. The gaussian watermarking game. IEEE Transactions on Information Theory 2002; 48(6): 1639–1667.



- [13] Zhao J, Koch E. A generic digital watermarking model. Computers and Graphics 1998; 22(4): 397–403.  
 [14] Cachin C. An information-theoretic model for steganography. Information and Computation 2004; 192(1): 41–56.  
 [15] Gribunin VG, Okov IN, Turintsev IV. Digital steganography. Moscow: Solon-Press, 2002; 272 p. (in Russian)  
 [16] Moulin P, O'Sullivan JA. Information-theoretic analysis of information hiding. IEEE Transactions on Information Theory 2003; 49(3): 563–593.  
 [17] Mittelholzer T. An information-theoretic approach to steganography and watermarking. LNCS 1999; 1768: 1–16.  
 [18] Nyeem H, Boles W, Boyd C. Developing a digital image watermarking model. 2011 International Conference on Digital Image Computing Techniques and Applications 2011; 468–473.  
 [19] Nyeem H, Boles W, Boyd C. Digital image watermarking: its formal model, fundamental properties and possible attacks. EURASIP Journal on Advances in Signal Processing 2014; 2014(1): 1–22.  
 [20] Ma L, Wu Z, Hu Y, Yang W. An Information-hiding model for secure communication. LNCS 2007; 4681: 1305–1314.  
 [21] Gonzalez RC, Woods REP. Digital Image Processing. 3 edition. New Jersey, Prentice Hall, 2007.

Table 3. List of notations used in the model of information hiding system.

#	Parameter	System 1: LSB audio steganography	System 2: Phase image spectrum watermarking
1	Host asset signal set $X_0^m$	$X_0^m = (\mathbb{B}^{16})_{ N_1 }^1$ (for one-channel audio)	$X_0^m = (\mathbb{B}^8)_{ N_1 \times N_2 }^2$
2	internal information initial form (given by the predicate $\pi_{bw}$ )	$\pi_{bw} = true$	$\pi_{bw} = false$
3	Binary vector length $N_b$ (if $\pi_{bw} = true$ )	$N_b \leq N$	-
4	Composite key set $K = K^s \times K^p$	$K = \mathbb{R}_{ N_1 }^1$	$K = \mathbb{Z} \cap [0, N_1 - 1] \times \mathbb{Z} \cap [0, N_2 - 1]$ ; $\Psi_0^t = \Psi = \mathbb{R}_{[N_1 \times N_2]}^2$ , $\mathcal{F}(x) = DFT(x)$ , $\mathcal{F}^{-1}(x) = DFT^{-1}(x)$ , $f_x = \arg \mathcal{F}(x)$ , $\psi_x = \arg \Psi(x)$ , where $f_x$ , $\psi_x$ are a feature matrix and a complement of a signal $x$ respectively
5	Pair of transformation functions $\mathcal{F}$ and $\mathcal{F}^{-1}$ , as well as sets $\Psi_0^t$ and $\Psi$	$\Psi_0^t = (\mathbb{B}^{16})_{ N_1 }^1$ ; $\mathcal{F}(x) = \mathcal{F}^{-1}(x) = x$ , no complement	-
6	Signal analysis function $\mathcal{A}$	-	-
7	Embedding function $\mathcal{E}$	$f^w(n) = \begin{cases} f(n), & n = k, i = 0..N_k - 1, \\ 2 \lfloor \frac{f(n)}{2} \rfloor + \Omega(n), & n = k, \end{cases}$	$f^w(n_1, n_2) = \begin{cases} \text{sign } f(n_1, n_2) \times \\ \times \max(1 + \varepsilon,  f(n_1, n_2) ), & \Omega(n_1, n_2) = 0, \\ \Omega(n_1, n_2), & \Omega(n_1, n_2) \neq 0, \end{cases}$ where $\varepsilon > 0$
8	Extraction function $\mathcal{D}$	$\tilde{\Omega}(n) = \tilde{f}^w(n) \pmod{2}$	$\Omega(n_1, n_2) = \begin{cases} 0, &  f^w(n_1, n_2)  > 1, \\ f^w(n_1, n_2), &  f^w(n_1, n_2)  \leq 1. \end{cases}$
9	Encoding method (given by the predicate $\pi_p$ , if $\pi_{bw} = true$ )	$\pi_p = true$ (but the value is not important because $X_0^m = \Psi_0^t$ )	-
10	Encoding function $\mathcal{P}$ (or $\mathcal{P}_f$ depending on $\pi_p$ , if $\pi_{bw} = true$ )	$\Omega(n) = \begin{cases} 0, & n = k, i = 0..N_k - 1, \\ b_i, & n = k, \end{cases}$	$\Omega(n_1, n_2) = \text{shift}(\Omega_1(n_1, n_2), \mathbf{k})$ , where $\text{shift}(x, \mathbf{a})$ is a cyclic shift of a matrix $x$ by the value defined by a vector $\mathbf{a}$
11	Output value: $\mathbf{b}^R$ , $\mathbb{W}^R$ , or $\xi$	$\mathbf{b}^R$	$\xi$
12	Detection form: $\mathbb{B}_{ N_b }^1$ , $X_0^m$ , or $\Psi_0^t$	$\mathbb{B}_{ N_b }^1$	$\Psi_0^t$
13	Inverse encoding function $\mathcal{P}^{-1}$ (or $\mathcal{P}_f^{-1}$ , depending on $\pi_p$ , if the detection form is $\mathbb{B}_{ N_b }^1$ or the output value is $\mathbf{b}^R$ )	$b_i^e = \tilde{\Omega}(k_i)$	Decoding is not performed
14	Detection function $\mathcal{R}$	Equations (11)-(12)	Equation (11) with the proximity function $\rho(\Omega, \tilde{\Omega}) = \frac{1}{N_1 N_2} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \eta(\Omega(n_1, n_2), \tilde{\Omega}(n_1, n_2))$ , where $\eta(x, y) = \begin{cases} 1, & x = y, \\ 0, & x \neq y. \end{cases}$

# DPCM with an adaptive extrapolator for image compression

M.V. Gashnikov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The method of image compression based on differential pulse-code modulation (DPCM) with an adaptive extrapolator is investigated. This extrapolator automatically adapts to local features of contours (boundaries) in the image. The issue of the negative effect of quantization on the result of optimizing the adaptive extrapolator is investigated. It is experimentally proved that, despite this effect, the adaptive extrapolator has the advantage over prototypes. Also, an experimental research of the considered method is carried out as a whole; a comparison is made with the JPEG method with respect to the maximum error in a half-tone Waterloo set of natural images.

*Keywords:* image compression; DPCM; extrapolation; quantization; entropy coding; compression ratio, maximum error

---

## 1. Introduction

Images have extremely large data sizes. The growth of data storage capacity does not solve this problem. First of all, this applies to multi- and hyperspectral images [1]. Due to the limited available resources, this problem is especially important for on-board image observation systems located on satellite and other aircraft, including atmospheric drones. So we have to use effective, often specialized, methods of image compression.

The number of approaches to image compression is very large [2-6]. Fractal compression methods [15] have the highest compression ratio, but they have not been widely used because of their computational complexity and unnatural distortion of the images. Methods based on wavelet transforms [8] are most preferable in the "efficiency/complexity" coordinates and have the widest possible scope. The method JPEG-2000 [9] is the most common of the wavelet compression methods. Methods based on two-dimensional discrete orthogonal transformations [10] have similar advantages. The method JPEG [11] is the most common of these compression methods. The JPEG method loses [12] a wavelet method in terms of efficiency, but JPEG significantly exceeds JPEG-2000 prevalence due to a lot of software and hardware in use.

However, all transformation-based compression methods require a lot of computing resources. Therefore, these methods are difficult to use in real-time systems, including on-board systems. In addition, such systems usually impose increased requirements for the control of the compressed data quality. But this is very problematic for the mentioned compression methods, because of the difficulties of controlling the compression error in the space of transformation coefficients.

Thus, in the situation of strictly quality control of compressed data and the restrictions on available resources, we need compression methods that do not use any spectral spaces. So these methods have to produce all the processing in the source brightness space. This allows providing low computational complexity and providing quality control of compressed information.

Such methods include differential compression methods [2-3], which decorrelate the signal by using the difference representation of this signal. This class of methods includes hierarchical methods of compression [13-14], which have a number of important advantages when used in on-ground image processing complexes. However, with on-board processing the hierarchical methods have no special advantages, but they have a high structural complexity. Therefore, in the opinion of the author, difference methods based on differential pulse code modulation (DPCM) [2-3] are the most preferable for on-board systems. In addition, the scope of DPCM method is not limited to real-time systems. For example, DPCM is included in other compression methods, such as JPEG. Thus, the task of research and increasing the effectiveness of compression methods based on DPCM is still relevant.

DPCM compression is based on extrapolation (prediction) of image pixels and coding of the difference between the original and extrapolated values of pixels. In [15], a compression method based on DPCM with an adaptive extrapolator is proposed, which is optimized on the basis of the minimum absolute value of extrapolation error. However, the paper [15] does not close a number of important issues related to the investigation of this compression method.

First of all, the effect of quantization on the optimizing the parameters of the adaptive extrapolator is not taken into account. In addition, the results of the extrapolator research are not given, and the conclusion about its effectiveness is made on the basis of research of the compression method as a whole. The present work aims to close these gaps in the research of this compression method, to draw substantiated conclusions about its effectiveness and to develop recommendations for its use.

## 2. Image compression based on differential pulse code modulation

Here is a brief simplified description of the image compression method based on DPCM. Let the non-negative integer pixels  $x(m, n)$  of the original digital image be processed line by line. Let's designate  $\hat{x}(m, n)$  the pixels of the decompressed (restored after compression) image. We calculate these values already at the stage of compression when processing the corresponding pixels for organizing the feedback. For each pixel  $x(m, n)$ , we calculate its extrapolated value  $\hat{x}(m, n)$  using an extrapolator  $P(\dots)$  based on the restored values  $\hat{x}(m, n)$  of the already processed pixels:

$$\hat{x}(m,n) = P\{\bar{x}(i,j) : i < m \text{ or } i = m \text{ and } j < n\}. \quad (1)$$

Then, an extrapolated value  $\hat{x}(m,n)$  is subtracted from the original pixel value  $x(m,n)$  to calculate the difference signal  $f(m,n)$ . Then, the difference signal  $f(m,n)$  is quantized  $Q(f)$ , the result of this quantization is a quantized difference signal  $\bar{f}(m,n)$ . This signal is encoded and transmitted through a communication channel or sent to an archive file. The quantized signal  $\bar{f}(m,n)$  is immediately used to calculate the corresponding recovered pixel value  $\bar{x}(m,n)$  :

$$f(m,n) = x(m,n) - \hat{x}(m,n), \quad \bar{f}(m,n) = Q(f(m,n)), \quad \bar{x}(m,n) = \bar{f}(m,n) + \hat{x}(m,n). \quad (2)$$

The restored pixel value  $\bar{x}(m,n)$  is used to extrapolate (1) the next pixel.

### 3. Extrapolation for image compression based on DPCM

The requirement of low computational complexity makes it necessary to apply in DIKM only the simplest [16] extrapolators of the form:

$$\hat{x}^{(0)}(m,n) = \bar{x}(m-1,n), \quad (3)$$

$$\hat{x}^{(1)}(m,n) = \frac{1}{2}(\bar{x}(m-1,n) + \bar{x}(m,n-1)), \quad (4)$$

$$\hat{x}^{(2)}(m,n) = \bar{x}(m,n-1). \quad (5)$$

These linear extrapolators work worst on contours (object boundaries). As an answer to this problem, extrapolators that are invariant to contours are considered, for example Greham's extrapolator [3], which is invariant to vertical and horizontal contours:

$$\hat{x}^G(m,n) = \begin{cases} \hat{x}^{(0)}(m,n), & \text{if } \lambda_m(m,n) < \lambda_n(m,n); \\ \hat{x}^{(2)}(m,n), & \text{if } \lambda_m(m,n) \geq \lambda_n(m,n), \end{cases} \quad (6)$$

где

$$\lambda_m(m,n) = |\bar{x}(m,n-1) - \bar{x}(m-1,n-1)|, \quad \lambda_n(m,n) = |\bar{x}(m-1,n) - \bar{x}(m-1,n-1)|. \quad (7)$$

The smaller difference from the differences (7) gives the direction of the contour in a small neighborhood of the current image pixel. The relation (6) provides extrapolation "along" this direction. Such an extrapolator is more accurate on the contours, but on flat sections it loses to the extrapolator (4), which is more stable to noise due to averaging.

The advantages of the "contour" extrapolator (6) and the "averaging" extrapolator (4) are combined by an adaptive extrapolator:

$$\hat{x}^A(m,n) = \begin{cases} \hat{x}^{(0)}(m,n), & \text{if } \lambda(m,n) < \lambda^{(-)}; \\ \hat{x}^{(1)}(m,n), & \text{if } \lambda^{(-)} \leq \lambda(m,n) \leq \lambda^{(+)}; \\ \hat{x}^{(2)}(m,n), & \text{if } \lambda^{(+)} < \lambda(m,n), \end{cases} \quad (8)$$

where  $\lambda(m,n)$  is a «contour direction feature»:

$$\lambda(m,n) = \lambda_m(m,n) - \lambda_n(m,n), \quad (9)$$

$\lambda^{(-)}, \lambda^{(+)}$  are parameters of the adaptive extrapolator, which are selected in the ranges:

$$-x_{\max} \leq \lambda^{(-)} \leq 0 \leq \lambda^{(+)} \leq x_{\max}, \quad (10)$$

where  $x_{\max}$  is the maximum brightness in the image. If the feature (9) is close to zero (the current pixel is in a flat image area), then the averaging extrapolator (4) is used, but if the feature (9) has a large value (positive or negative), then



extrapolation (6) "along the contour" occurs. The extrapolator (8) adapts to each particular image: if the contours in the image are small, then the parameters  $\lambda^{(-)}$ ,  $\lambda^{(+)}$  are set far from zero, but if there are many vertical and/or horizontal contours, the corresponding parameter must have a large absolute value.

The parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$  are automatically calculated before the actual DPCM processing for each particular image. A special optimization procedure for the adaptive extrapolator is used for this (see below). Then the parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$  are placed in the archive, because they are also necessary for decompression.

#### 4. Optimization of the adaptive extrapolator for image compression based on DPCM

Optimization of the adaptive extrapolator (search of parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$ ) is based on minimizing the sum of the absolute values of extrapolation errors over the set  $\omega = \{(m, n)\}$  of coordinates of all image pixels:

$$\delta(\lambda^{(+)}, \lambda^{(-)}) = \sum_{(m,n) \in \omega} |x(m, n) - \hat{x}(m, n)| \rightarrow \min_{\lambda^{(+)}, \lambda^{(-)}}. \quad (11)$$

The error (11) can be divided into three component parts corresponding to different ranges of the "contour direction feature" (9):

$$\delta(\lambda^{(+)}, \lambda^{(-)}) = \delta^{(-)}(\lambda^{(-)}) + \delta^{(0)} + \delta^{(+)}(\lambda^{(+)}) \quad . \quad (12)$$

where

$$\delta^{(-)}(\lambda^{(-)}) = \sum_{(m,n) \in \omega^{(-)}} |x(m, n) - \hat{x}(m, n)|, \quad \delta^{(0)} = \sum_{(m,n) \in \omega^{(0)}} |x(m, n) - \hat{x}(m, n)|, \quad \delta^{(+)}(\lambda^{(+)}) = \sum_{(m,n) \in \omega^{(+)}} |x(m, n) - \hat{x}(m, n)|,$$

$$\omega = \omega^{(-)} \cup \omega^{(0)} \cup \omega^{(+)}, \quad \omega^{(-)} = \{(m, n) : \lambda(m, n) < 0\}, \quad \omega^{(0)} = \{(m, n) : \lambda(m, n) = 0\}, \quad \omega^{(+)} = \{(m, n) : \lambda(m, n) > 0\}.$$

As a result, the two-parameter optimization problem (11) is decomposed into two one-parameter problems that are solved independently of each other:

$$\lambda^{(+)} = \arg \min_{\lambda} \delta^{(+)}(\lambda), \quad \lambda^{(-)} = \arg \min_{\lambda} \delta^{(-)}(\lambda). \quad (13)$$

To solve these problems, a special matrix is filled in the preliminary pass through the image

$$\Delta_{i, \lambda} = \sum_{(m,n) \in \omega(\lambda)} |x(m, n) - \hat{x}^{(i)}(m, n)|, \quad 0 \leq i \leq 2, \quad -x_{\max} \leq \lambda \leq x_{\max}. \quad (14)$$

Each element  $\Delta_{i, \lambda}$  of this matrix contains the sum of extrapolation errors of extrapolator number  $i$  (3-5) for all pixels for which the value of feature (9) is equal  $\lambda$ . For the filled matrix (14), a one-dimensional array of extrapolation error values  $\delta^{(+)}$  is filled using a recurrence procedure:

$$\delta^{(+)}(x_{\max}) = \sum_{\lambda=0}^{x_{\max}} \Delta_{1, x_{\max}}, \quad \delta^{(+)}(\lambda) = \delta^{(+)}(\lambda + 1) + \Delta_{2, \lambda} - \Delta_{1, \lambda}, \quad 0 \leq \lambda < x_{\max}. \quad (15)$$

The computational complexity of this procedure does not depend on the image size. Optimal value  $\lambda^{(+)}$  can be found in this array  $\delta^{(+)}(\lambda^{(+)})$  by an exhaustive search (the length of this array is only  $x_{\max} + 1$ ). Analogously  $\lambda^{(-)}$  is calculated.

#### 5. Quantization for image compression based on DPCM

For quantization in DPCM, the Max scale [5-6] is usually used, which provides the minimum relative root-mean-square error

$$\varepsilon_{rel}^2 = \frac{\varepsilon^2}{D_x} = \frac{1}{MND_x} \sum_{(m,n) \in \omega} (x(m, n) - \bar{x}(m, n))^2, \quad (16)$$

where  $x(m, n)$  is the original image,  $\bar{x}(m, n)$  is the restored (decompressed) image,  $D_x$  is the image variance,  $M \times N$  are image sizes.

For unique data, more strong error control [17] is necessary. For example, it is necessary for compression of hyperspectral images [18-20]. In this case, a quantizer with a uniform scale [3] can be used for DPCM. This quantizer provides maximum error control:

$$\varepsilon_{\max} = \max_{(m,n) \in \Omega} |x(m,n) - \bar{x}(m,n)|. \quad (17)$$

## 6. The effect of quantization on the adaptive extrapolator optimization

It should be noted an important nuance that occurs when optimizing an adaptive extrapolator. The optimization procedure described above for the adaptive extrapolator can be performed only in the absence of quantization (when the quantizer is "switched off"). In this case, the original  $x(m,n)$  and recovered  $\hat{x}(m,n)$  values of the image pixels are equal, the difference  $f(m,n)$  and quantized difference  $\hat{f}(m,n)$  signals are also equal. The "activation" of the quantizer at the stage of extrapolator optimization would lead to the impossibility of calculating the restored values of the pixels, since they, through the chain of transformations, depend on the parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$  of the extrapolator, which at this stage are still unknown.

Thus, the general scheme of DPCM compression with an adaptive extrapolator is as follows. First, to optimize the extrapolator, a preliminary pass through the image with the "switched off" quantizer (i.e., zero-error) is performed. In this case, the extrapolator parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$  are calculated. After that, the DPCM-compression itself is made, in which the quantizer is "switched on" again, and the parameters  $\lambda^{(+)}$ ,  $\lambda^{(-)}$  found on the preliminary pass are used for extrapolation.

As a result, the parameters of the extrapolator, optimal by criterion (11) with zero error, are used in compression with nonzero error. In this situation, these parameters are no longer optimal, and the question of how far from the optimum they are, requires additional research, which can only be experimental. Computational experiments were carried out on the so-called set of halftone images "Waterloo" [21], which is traditionally used for the research of compression methods. Typical results are shown in Fig. 1-2.

The investigation was carried out for one-parameter problems (13). The quantities  $\delta^{(+)}$ ,  $\delta^{(-)}$ , introduced by the relation (12), which are calculated with zero quantization error, will be referred to below as the "estimative total error". The quantities  $\delta^{(+)}$ ,  $\delta^{(-)}$ , calculated with a non-zero quantization error, will be referred to below as the "true total error". Thus, it is necessary to answer the question of how closely the minimum of the true total error and the minimum of the estimative total error are located.

The dependence of the total error  $\delta^{(+)}$  on the extrapolator parameter  $\lambda^{(+)}$  is shown in Fig. 1. The value of the second parameter  $\lambda^{(-)}$  was fixed (it was chosen in the optimal way). Then the same research was performed for total error  $\delta^{(-)}$ . The dependence of the total error  $\delta^{(-)}$  on the extrapolator parameter  $\lambda^{(-)}$  is shown in Fig. 2. Thus, accordingly, the value of the parameter  $\lambda^{(+)}$  was fixed.

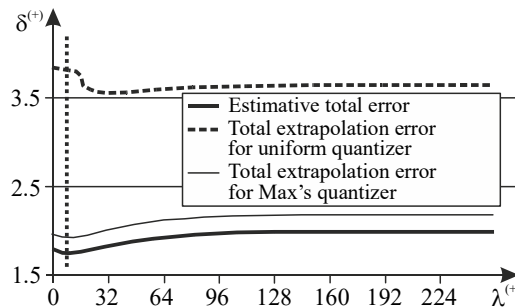


Fig. 1. Dependence of the extrapolation total error  $\delta^{(+)}$ , corresponding to positive values of the contour direction feature (9), from the extrapolation parameter  $\lambda^{(+)}$  for a fixed  $\lambda^{(-)} = -19$  (the vertical line shows the position of the minimum of the estimative total error).

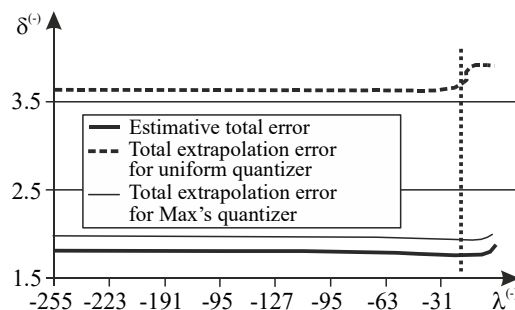


Fig. 2. Dependence of the extrapolation total error  $\delta^{(-)}$ , corresponding to negative values of the contour direction feature (9), from the extrapolation parameter  $\lambda^{(-)}$  for a fixed  $\lambda^{(+)} = 7$  (the vertical line shows the position of the minimum of the estimative total error).

These researches allow drawing the following conclusions:

1. Quantization affects the results of optimization of the adaptive extrapolator, since the minimum of the true total error and the minimum of the estimative total error may not be equal
2. When using the Max quantizer, the values found for the parameters of the adaptive extrapolator are closer to the optimum.
3. It is necessary to carry out a research of the extrapolator's efficiency for estimating the effect of a mismatch between the optimums of the true total error and estimative total error.

## 7. Research of the effectiveness of the adaptive extrapolation algorithm

To evaluate the effectiveness of the adaptive extrapolator for compression, this extrapolator was compared with other extrapolators. The comparison was produced by the entropy  $H_q$  of the quantized difference signal. This entropy is a good estimate of compressed data size. The results are shown in Fig. 3-4.

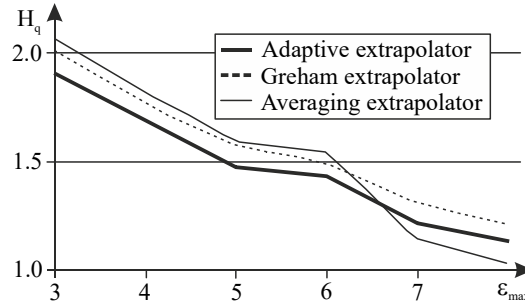


Fig. 3. Dependence of the entropy  $H_q$  of the quantized difference signal on the maximum error  $\epsilon_{\max}$  when using a uniform quantization scale.

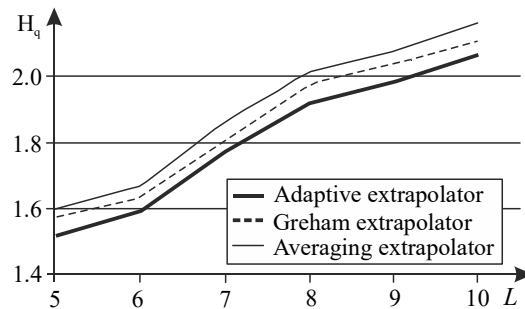


Fig. 4. Dependence of the entropy  $H_q$  of the quantized difference signal on the number  $L$  of quantized levels when using Max's quantization scale.

### Conclusions:

1. The adaptive extrapolator has the advantage over prototypes over the entropy of the quantized signal.
2. Consequently, the negative influence of the quantizer described in the previous section does not have a determining value (at least for small extrapolation errors).
3. With increasing maximum error, the negative influence of the quantizer on the efficiency of the adaptive extrapolator increases. With a maximum error of more than six, adaptive extrapolator loses the advantage.

## 8. Experimental research of DPCM with an adaptive extrapolator for image compression

To evaluate the efficiency of the image compression method based on DPCM with the adaptive extrapolator, it was compared with method JPEG in the coordinates "error-compression ratio" on the set of images "Waterloo" [21], which is traditionally used for comparison of compression methods. The results obtained, averaged over all images of the set, are shown in Fig. 5. The results of the experiments demonstrate a significant gain (up to two times) of DPCM with an adaptive extrapolator for the JPEG method with respect to the maximum error.

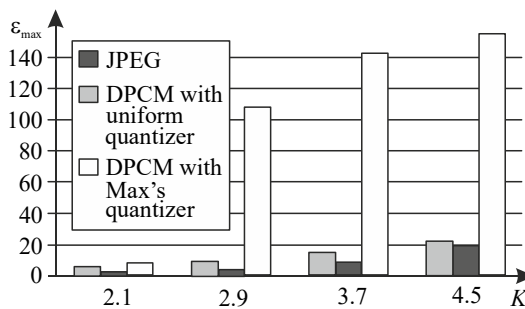


Fig. 5. The dependence of the maximum error  $\epsilon_{\max}$  on the compression coefficient  $K_c$  when comparing DPCM with an adaptive extrapolator versus the JPEG compression method.

## 9. Conclusion

The significant influence of quantization on the optimization of the parameters of the adaptive extrapolator is described and experimentally confirmed. It has been shown experimentally that, despite this negative effect of quantization, the adaptive extrapolator within the DPCM compression method still has an advantage over prototypes with a small error. Also, computational experiments were carried out to research the efficiency of the DPCM compression method with an adaptive extrapolator and showed its advantage over the JPEG compression method with respect to the maximum error. Based on the obtained results, it can be concluded that the considered method is promising for image compression systems and image transmission systems.

Further research will be aimed at eliminating the need for a preliminary pass through the image when optimizing the adaptive extrapolator, which is necessary to simplify the use of the method of image compression for real-time systems, including on-board systems. The possibility of such an improvement is based on the admissibility of estimating the distribution of extrapolation errors during the actual DPCM processing. The question of the quality of such a "consistently refined" assessment requires additional investigation.

## Acknowledgements

The work was funded by the Russian Science Foundation, grant No. 14-31-00014.

## References

- [1] Chang C. *Hyperspectral Data Processing: Algorithm Design and Analysis*. Wiley Press, 2013; 1164 p.
- [2] Sayood K. *Introduction to Data Compression*. The Morgan Kaufmann Series in Multimedia Information and Systems, 4ed., 2012; 743 p.
- [3] Salomon D. *Data Compression. The Complete Reference*. Springer-Verlag, 4ed, 2007; 1118 p.
- [4] Vatolin D, Smirnov M, Yukin V. *Data compression methods. Archive program architecture, image and video compression*. Moscow: "Dialog-MIFI" Publisher 2002; 384 p. (in Russian)
- [5] Woods E, Gonzalez R. *Digital Image Processing*. Prentice Hall, 3ed, 2007; 976 p.
- [6] Pratt W. *Digital image processing*. Wiley, 4ed, 2007; 807 p.
- [7] Woon WM, Ho ATS, Yu T, Tam SC, Tan SC, Yap LT. Achieving high data compression of self-similar satellite images using fractal. *Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2000*: 609–611.
- [8] Gupta V, Sharma V, Kumar A. Enhanced Image Compression Using Wavelets. *International Journal of Research in Engineering and Science (IJRES) 2014*; 2(5): 55–62.
- [9] Li J. *Image Compression: The Mathematics of JPEG-2000*. Modern Signal Processing. MSRI Publications 2003; 46: 185–221.
- [10] Plonka G, Tasche M. Fast and numerically stable algorithms for discrete cosine transforms. *Linear Algebra and its Applications 2005*; 394(1): 309–345.
- [11] Wallace G. The JPEG Still Picture Compression Standard. *Communications of the ACM 1991*; 34(4): 30–44.
- [12] Ebrahimi F, Chamik M, Winkler S. JPEG vs. JPEG2000: An Objective Comparison of Image Encoding Quality. *Proceedings of SPIE Applications of Digital Image Processing XXVII 2004*; 5558: 300–308.
- [13] Gashnikov M. Interpolation for hyperspectral images compression. *CEUR Workshop Proceedings 2016*; 1638: 327–333.
- [14] Gashnikov M, Glumov NI. Development and Investigation of a Hierarchical Compression Algorithm for Storing Hyperspectral Images. *Optical Memory and Neural Networks*. Allerton Press 2016; 25(3): 168–179.
- [15] Gashnikov M, Mullina SS. Adaptive parameterized predictor for differential image compression. *Proceedings of the International Conference "Information Technologies and Nanotechnologies"*. Samara 2015: 64–67. (in Russian)
- [16] Efimov V, Kolesnikov AN. Effectiveness estimation of the hierarchical and line-by-line lossless compression algorithms. *Proceedings of the III conference "Pattern recognition and image analysis"*. Nijni Novgorod 1997; 1: 157–161. (in Russian)
- [17] Lin S, Costello D. *Error Control Coding: Fundamentals and Applications*, second edition. New Jersey: Prentice-Hall, inc. Englewood Cliffs, 2004; 1260 p.
- [18] Chang C. *Hyperspectral data exploitation: theory and applications*. Wiley-Interscience, 2007; 440 p.
- [19] Gashnikov MV, Glumov NI. Hierarchical GRID Interpolation under Hyperspectral Images Compression. *Optical Memory and Neural Networks (Information Optics)*. Allerton Press 2014; 23(4): 246–253.
- [20] Borengasser M, Hungate W, Watkins R. *Hyperspectral Remote Sensing – Principles and Applications*. CRC Press, 2004; 128 p.
- [21] Waterloo Grey Set. University of Waterloo Fractal coding and analysis group: Mayer Gregory Image Repository. URL: <http://links.uwaterloo.ca/Repository.htm> (19.12.2016).
- [22] Gashnikov MV, Glumov NI. Onboard processing of hyperspectral data in the remote sensing systems based on hierarchical compression. *Computer Optics 2016*; 40(4): 543–551. (in Russian)

# Intelligent geographic information platform for transport process analysis

O. Golovnin<sup>1</sup>, A. Fedoseev<sup>1,2</sup>, T. Mikheeva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Space-Rocket Centre Progress, 18, Zemetsa str., 443009, Samara, Russia

---

## Abstract

We developed an intelligent geoinformation platform for transport process analysis. The paper describes a purpose and functions of the synthesized platform as well as its structure including components and tools. We used the intelligent transport geoinformation platform for solving problems of acquisition, storage, processing and analysis of objects, processes and phenomena of urban transport infrastructure. The paper presents results of simulation and full-scale experiments.

*Keywords:* intelligent transport system; geoinformation system; traffic flow; transport infrastructure, ITSGIS; GIS; ITS

---

## 1. Introduction

Improvement of economy, comfort and traffic safety is promoted by the use of modern systems for transport processes analysis, including [1,2]: information systems of Vehicle-to-Infrastructure class, automated traffic control systems and intelligent transport systems. These systems are based on modern achievements and innovations in the field of transport management [3,4]: information support for traffic participants, transport detectors and radars, meteorological information systems and video surveillance systems, traffic lights, which include intelligent neural networks [5] and genetic [6] algorithms, intelligent pedestrian crossings and information displays on road forks.

In terms of complex intelligent transport systems development, information sources determine the heterogeneity of transport processes data. These information sources also cause heterogeneity of both hardware and software platforms, which leads to significant reduction in the efficiency and relevance of the data and decision-making procedures [7, 8]. All the factors mentioned above ensure consistency and relevance of developing special means for transport process analysis that makes it possible to compensate the heterogeneity of information spaces.

High efficiency in the study of transport processes can't be achieved without the spatial reference of numerous static and dynamic objects that compose the transport infrastructure [9,10]. Storage and manipulation of geospatial and attribute data that describe the objects, processes and phenomena of the transport infrastructure can be implemented with a high degree of efficiency in the environment of the geoinformation system [11,12]. Geoinformation system allows constructing a geoinformation model of the transport network in an urbanized area reflecting all changes in the real world transport infrastructure [14]:

- changes in the infrastructure component of the street-road network: overlapping lanes, narrowing the roadway when performing road works, reconstruction, construction of new residential areas, shopping and entertainment centers and other points of attraction;
- changes in the deployment of traffic management facilities: implementation of temporary schemes of traffic management, modernization of traffic lights, installation of new traffic lights and road signs, introduction of dedicated lanes for public transport.

Thus, the goal of this work is to develop intelligent algorithms and software that form an integrating platform based on the geoinformation system that allows solving heterogeneous tasks of transport process analysis in a single operating environment:

- monitoring the characteristics of traffic flows, street-road network, traffic management facilities, environment;
- organization of freight and passenger transportation: analysis of transport demand, construction of traffic routes, reduction of resources spent, increase in profits;
- optimal management (local and global): development of the transport network of regions and megacities, interaction of various modes of transport, reducing transport delays, necessary and sufficient deployment of traffic management facilities;
- safety improvement: pre-crash restraint, risk of accidents reduction, accident incidence rate decrease.

## 2. Purpose, capabilities and structure of the platform

Functional capabilities of the transport analysis platform cover the whole range of tasks posed in the management of transport processes [15,16]:

- monitoring of objects, processes and phenomena of transport infrastructure;
- organization of freight and passenger transportation;
- optimal management of transport infrastructure and traffic flows;
- safety.

The platform is built on a modular principle: each task or part of it is implemented as a subsystem (module). All subsystems are implemented on the basis of a single tool environment – the core of the system. The core is a universal software infrastructure that includes a set of software components, modules and a georeferencing database.

Functional specification of the core includes:

- formation, processing and storage of a database;
- multithreaded processing of data by algorithms;
- ensuring interaction with the geographic information system;
- provision of interaction with subsystems and between them, other systems;
- provision of platform integration capabilities in various public and network services (internet portals);
- import of static data from other formats and systems;
- export data to external formats;
- providing information to the operator in a graphic form for decision making [17].

Functional specification of the monitoring subsystem includes:

- obtaining information from assessment, measuring, meteorological information systems and processing of survey data;
- transfer of operational attribute information to the georeferencing database;
- definition of parameters of fundamental diagrams for sections of the road network, characteristics and composition of traffic flows;
- calculation of the forecast of the development of the situation [18];
- response to the results of self-diagnosis of technical monitoring and management tools;
- formation and transmission of events to the control subsystem.

Functional specification of the control subsystem includes:

- reception and processing of events coming from the monitoring subsystem;
- the definition of the management and information scenario;
- transformation of the control actions generated by the control subsystem into the format of actuators (traffic light controllers, variable information signs);
- transmission of control signals to actuators.

Functional specification of the subsystem of information support of the traffic participants includes:

- conversion of the control actions developed by the control subsystem to the format of the information support facilities for the traffic participants (web service, SMS-distribution);
- providing access to public elements of visualization and information services platform through the website;
- ensuring the functioning of the Internet services of the platform on mobile devices.

The territorial remoteness of platform users, the interest in different functional components, the use of their own data warehouses determine the distributed architecture of the platform [19, 20], in which the interaction is carried out by means of local computer networks or the Internet using a central server (Fig. 1). Access of interested persons (users) to data is limited by their spheres of influence [21].

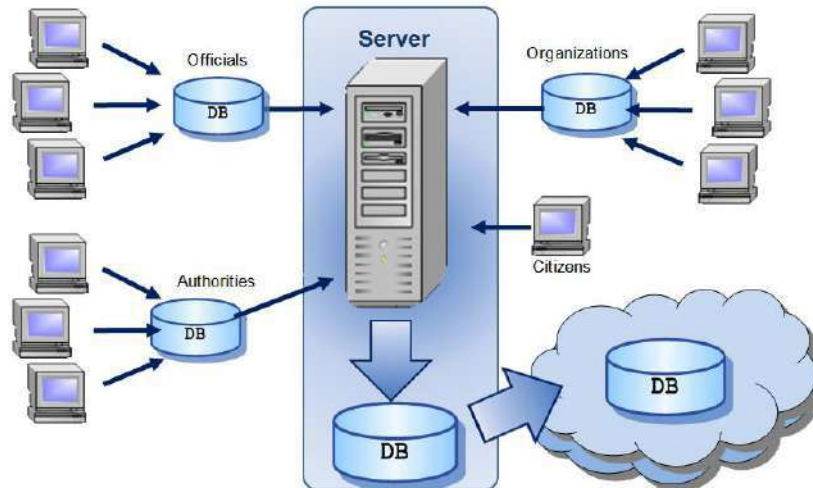


Fig. 1. Scheme of interaction between platform users.

The platform includes:

- database server with a database management system supporting geospatial objects;
- application server;
- client applications.

The database management system provides secure storage and manipulation of data, maintaining their integrity, replicating. The server is a host for hosting platform services and their business logic, provides multi-user work, authorization and delineation of access rights for clients and subsystems, compression and encryption of data, maintaining connections with customers. Client applications provide end-user access to platform services. Subsystems that perform a task are embedded in the database, application server, and client applications using the developed methods for securely connecting / disconnecting subsystems.

The georeferencing database is built on the basis of the “Virtual Database” pattern and is an integrated repository of attribute and geospatial data, including:

- electronic basis of the map of the urbanized territory;
- directories and registers;
- geobjects of transport infrastructure, implemented by connected subsystems.

The virtual database of the platform supports topological relationships in queries, ensures execution of both SQL and LINQ. The requirement of interaction with third-party subsystems and data sources dictates the need to apply open standards in the database used for storage, processing and transmission of geospatial information. As such standards, the platform uses OGC specifications and standards [22].

The PostgreSQL database management system with the PostGIS geospatial extension used to store geospatial and attribute information. The extension defines a special type of spatial data – a full set of functions and indexes for working with them. The introduction of additional relational tables into the database is provided by the version subsystem of database structure migration: when adding tables, a transaction is created that makes changes to the existing data schema and updates the version of the schema; when you remove the tables, the reverse transaction is executed.

One of the functional components of the platform is the ITSGIS geoinformation system. ITSGIS provides a platform for analyzing transport process with tools for viewing an electronic map with applied objects with the ability to connect / disconnect map layers, scale, and select geodata.

The reference of transport infrastructure objects to an electronic map is determined by the “Geometry” entity that migrated from the ITSGIS core as an OGC geometry in the Well-known Binary format. Geometry is equipped with a description of the visualization style and has a reference

to the layer of the electronic map (for example, for the “Road Sign” object, this layer is the “Posts” layer).

The transport infrastructure objects reference is defined by the “Geometry” entity, migrated from ITSGIS’ core and represented by OGC-geometry in the format of “Well-known Binary”. Geometries possess visualization style descriptions and are bound to an electronic map layer (for example, a “Road sign” object is bound to the “Posts” layer).

Objects can possess a description of geometry’s address ing, describing such address components as “Country”, “Region”, “City”, “District”, “Street”, “Highway”, “Landmark”. A “Landmark” is commonly represented by a house number, a picketage or a verbal description.

Requirements for security, reliability and interoperability of the application server platform lead to the need to build a cross-platform unified transaction model. In the developed platform, this model provides a set of technologies Windows Communication Foundation. The application server delineates the access rights of users of the system based on the geo-role subsystem: the right to view / modify information is determined both by taking into account the layer of the electronic map and the polygonal area on the map. Additional services and their business logic are connected to the application server during the configuration of the server using the IoC container. The container registers all additional services, models, data access layers, domain and auxiliary objects. When the application server is started, dependencies between objects are automatically resolved, additional services are created and launched.

Client applications are implemented as “thick” (Microsoft .NET application with WinForms and WPF interfaces for the CLR runtime) and “thin” (web application on the Yii2 platform with an interface implemented in the browser environment using HTML, CSS and JavaScript) clients. All client applications provide the connection of additional modules (subsystems) and flexible configuration of the user interface. The client consists of freely connected modules that are dynamically detected and compiled into a single unit at runtime. Modules contain both visual and non-visual components, representing different vertical layers of the platform.

### 3. Toolware of the platform

Based on the tasks to be solved, the software modules of the platform for transport process analysis, built using the “Embedded extension” pattern, are implemented for the C # and XAML runtime CLR.

Modules (subsystems) are aimed to solve monitoring tasks:

- PluginYamgis – module “Monitoring”;
- PluginIntensity – module “Characteristics of traffic flows”.

Modules (subsystems) are aimed to solve the problems of freight and passenger transportation:

- PluginUds – the module “Street-road network”;
- PluginRoute – the module “Routes”.

Modules (subsystems) are aimed to solve management problems:

- PluginPassport – module “Management of transport infrastructure”;
- PluginSimulation – module “Modeling”.

Modules (subsystems) are aimed at solving security tasks:

- PluginDTP – module “Road accidents”;
- PluginFirecenter – module “Accident clusters”.

The following modules (subsystems) have been developed for storage, processing and analysis of data on transport infrastructure objects:

- PluginInfo – module “Organizations on the map”;
- PluginPetrolstation – module “Gas station”;
- PluginRWC – module “Railway crossings”;
- PluginPost – module “Road signs and traffic lights”;

- PluginRoadmarking – module “Road marking”;
- PluginBarricado – module “Fences”;
- PluginBusStop – module “Public transport stops”;
- PluginCabenetwork – module “Artificial lighting”.

To obtain primary cartographic information about the transport infrastructure, the ITSGIS geoinformation system [23] and the cartographic service Open Street Map [24] are used. Data on the transport infrastructure in ITSGIS are presented in the form of detailed layers of the electronic map with exact geometric parameters of the street-road network and are used to form the model of the graph of the street-road network. Geobjects are stored and processed in a format and according to requirements that conform to the OGC specifications [22]. Each class of objects that have a spatial reference has its own layer of the electronic map. The Open Street Map data is stored as OSM format files with a description of the road segments [24] that are used to compile the cartographic base in ITSGIS. ITSGIS includes the program module ITS.MapConverter, which converts data of various formats (including OSM) to the form used by ITSGIS, and vice versa.

The web application for the provision of traffic information for the Yii2 platform with an interface implemented in a browser environment using HTML, CSS and JavaScript was developed. ITSGIS provides access to the geodata for the Web application in the form of XML documents and tiles, which are displayed by the OpenLayers 2 module.

Features of the physical implementation of the system are described by the component diagram shown in Fig. 2.

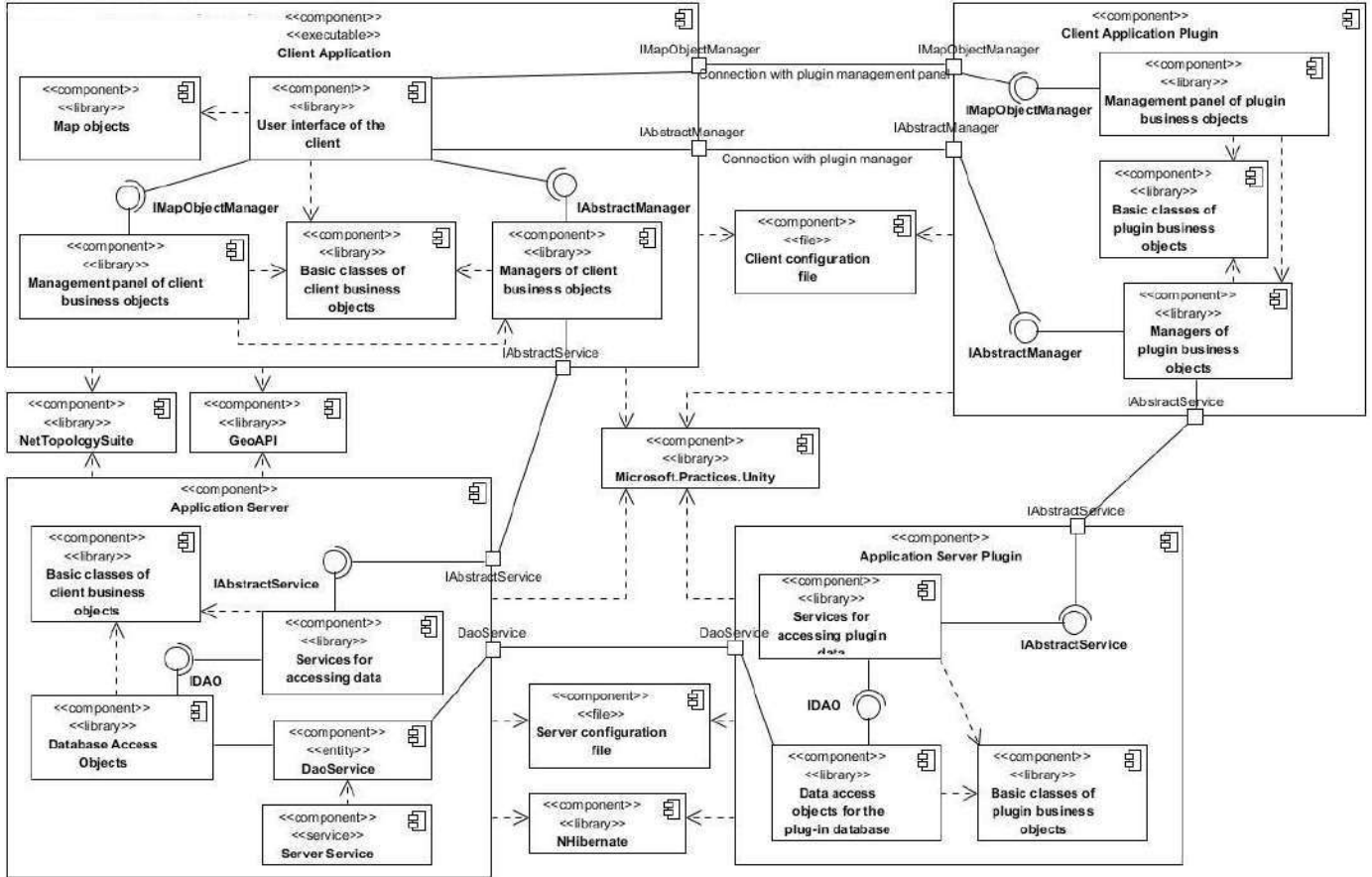


Fig. 2. Component diagram of the platform.

To represent the topology of the developed transport analysis platform, a deployment diagram was constructed (generally, Fig. 3). At the heart of the system is a multi-level architecture in which levels are distinguished: data processing (database server), business logic (application server based on the WCF technology) and presentation to the user (clients).

Modified deployment method provides access to the system through the site-geoportal from the web browser. In this case, the system nodes are:

- “app.itsgis server” – the server on which the application server and the geoserver are deployed;
- “web-portal server” – the geoportal application server on which Apache and MySQL components are deployed;
- “client” – the end-user computer on which the web browser is installed.

#### 4. Deployment of the platform for solving problems of transport process analysis

##### 4.1. Transport management tasks

The task of studying local control in conjunction with a point zone (Fig. 4) is associated with the method of local control of transport flows at a separate intersection.



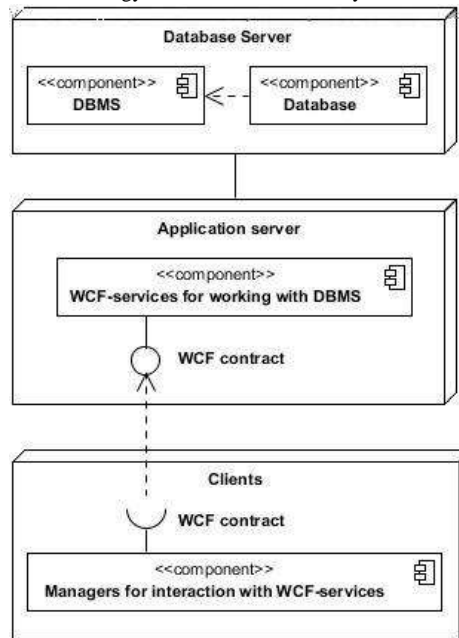


Fig. 3. Deployment diagram of the platform.

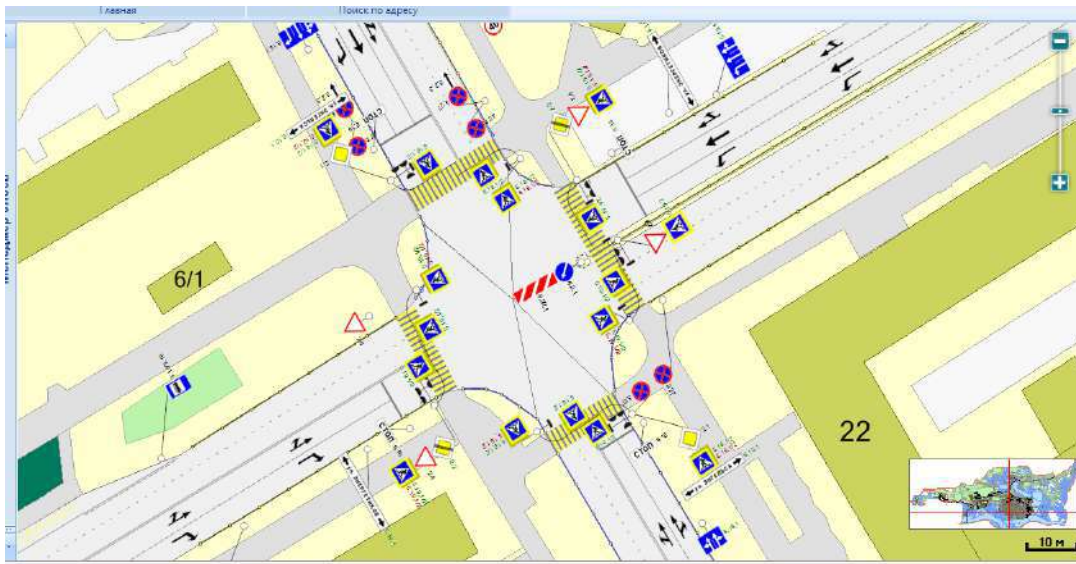


Fig. 4. Functional zone of local control.

Co-ordinated control methods involve the optimization of transport processes on the highway, the control zone will be defined in this case by linear decomposition.

System zonal management affects several classes of transport infrastructure objects, united by the task being solved, and therefore spatial zoning will be defined by the polygon (Fig. 5).

A lot of sections of the street-road network contain subsets of distances, intersections, pedestrian crossings, railway crossings, overpasses and tunnels for solving the problems of studying traffic flows in these sections when describing them by a planar graph. Fig. 6 shows a multi-level transport interchange.

#### 4.2. Traffic flow intensity research

The layer of the electronic map showing the intensity of traffic flows contains the mean annual daily intensity at intersections (Fig. 7).

The graph of the street-road network of the local intersection is characterized by the intensity of traffic flows (Figure 8). The color indicates the power intensity on this arc of the graph of the street-road network. The platform uses the “Divergent color scheme” pattern to visualize the intensity of traffic flows. For the central value of the intensity of the traffic flow, the capacity of the section of the road network is assumed. The intensity values above the center value are displayed in shades of red, values below the central level are displayed in shades of green. The use of a non-standard color scheme for geoinformation systems is conditioned by the specifics of the subject area and the following analogies: the green signal of the traffic light allows movement – low intensity will allow the road network to pass freely, the red signal of the traffic light prohibits traffic – a high intensity value will not allow the passage without forced stops.

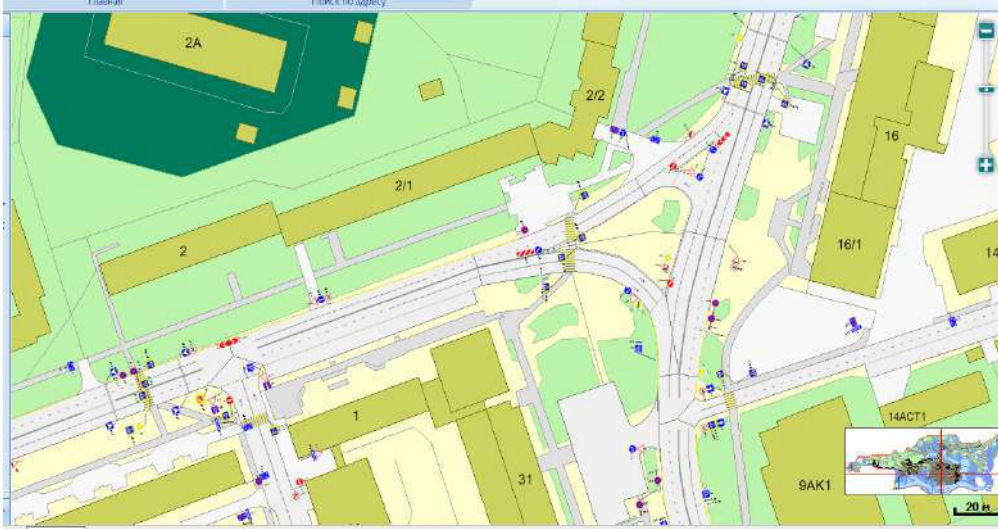


Fig. 5. Functional zone of system control.

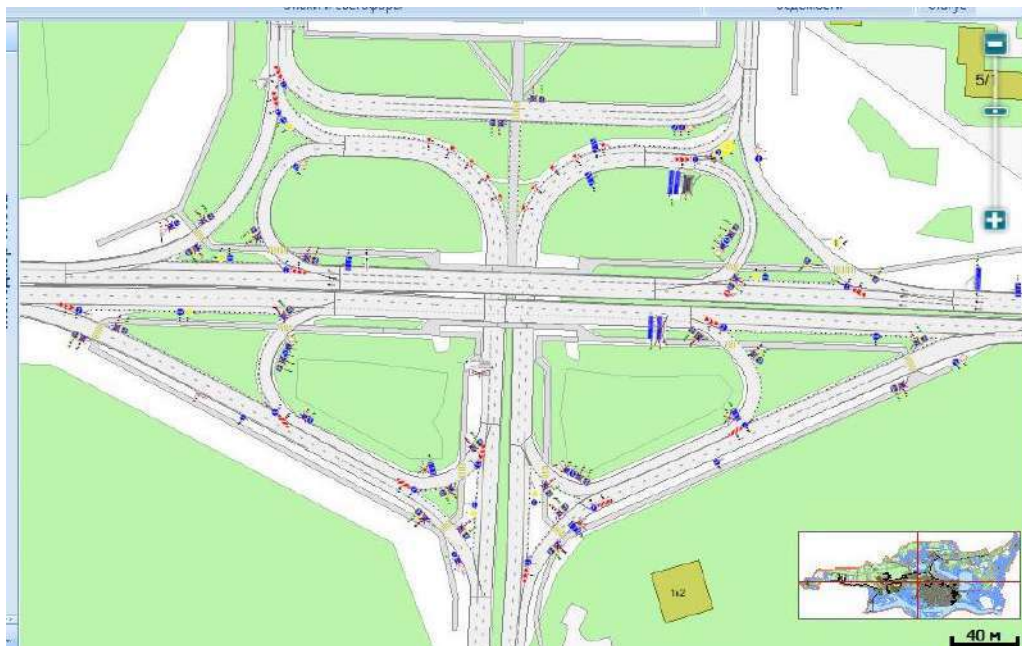


Fig. 6. Multi-level road junction.



Fig. 7. Annual average daily intensity of transport flows.



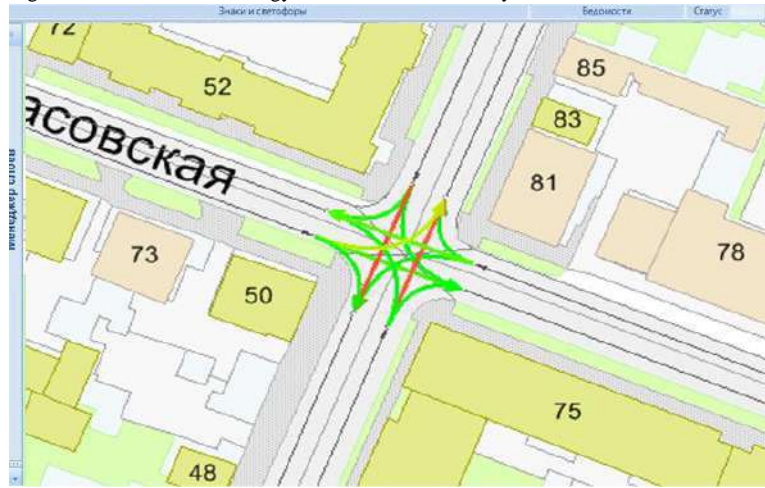


Fig. 8. Intensity of transport flows of a local crossroad.

### 5. Results and Discussion

The expected results from the implementation of the platform for transport process analysis for the cities of Surgut, Trehgorny, Oktyabrsk and the rural settlement of Kinel-Cherkasy (Russia) are shown in Fig. 9 and Fig. 10.

The number of accident clusters (Fig. 9) is calculated using the developed PluginDTP and PluginFirecenter subsystems.

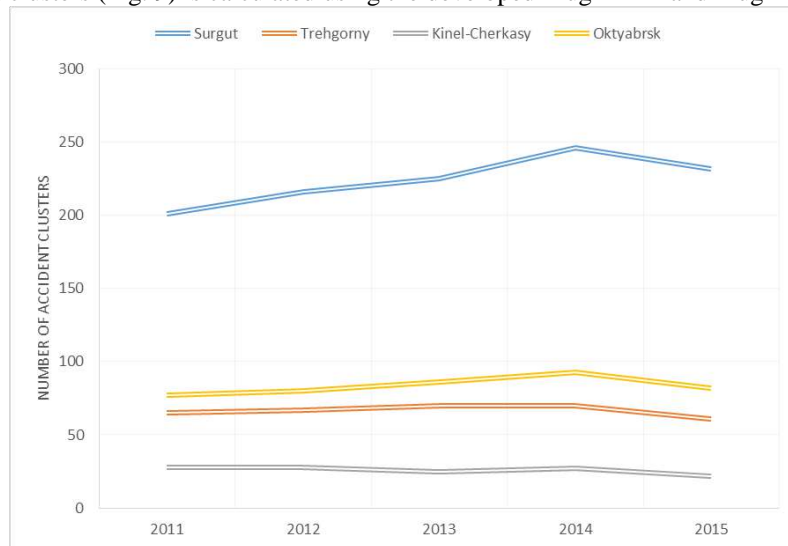


Fig. 9. Number of accident clusters.

The effect of reducing the transport delay will be obtained by reducing the time of movement of vehicles along sections of the street-road network. The data were obtained as a result of modeling the transport situation in the developed mesomodeling subsystem in the ITSGIS environment (Fig. 10).

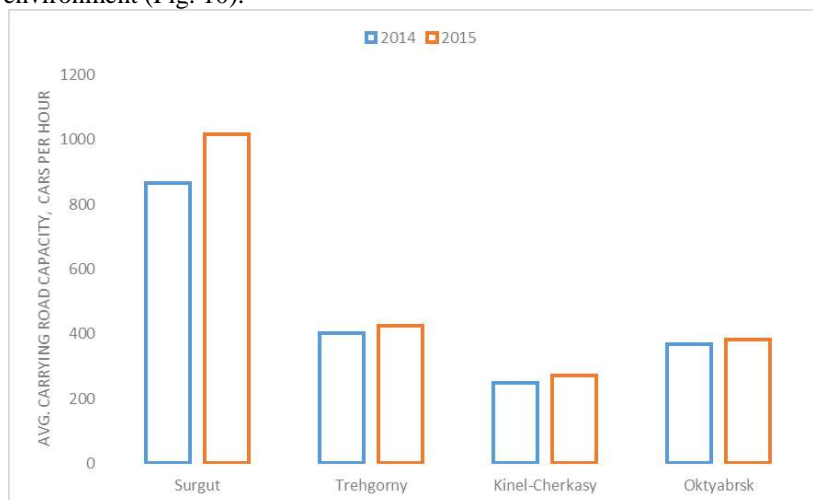


Fig. 10. Average carrying road capacity.

The main result of the research is the introduction of the developed algorithms, tools and software into the practice of transport process management on an urbanized territory. Theoretical and practical results of the work related to the creation of the intelligent geoinformation platform for transport process analysis were applied in the following works at the science and production center “Intelligent Transport Systems” in 2011–2016: “Establishment of the municipal geoinformation system of the Samara City Administration in the part of creating an applied GIS of the Department of Improvement and Ecology”, “Development of the road traffic management project in Oktyabrsk”, “Preparation of initial data for creating an electronic transport model of Samara”, “Preparation of initial data for calibration of the transport model of Samara”, “Adjustment of the road traffic management project on the roads of Surgut”, “Development of road traffic management project in Trekhgorny (Chelyabinsk region)”, “Development of an integrated scheme for traffic management in the rural settlement of Kinel-Cherkasy”.

The validity of the work results is confirmed by the correct use of theoretical and experimental methods, basing on the fundamental works of domestic and foreign scientists, approbation of research results in practice, deployment in the science and production center “Intelligent Transport Systems” and in the Samara University.

## 6. Conclusion

We developed the software and algorithms based on intelligent models and modern approaches to the distributed geoinformation systems creation in order to improve efficiency of transport process analysis. Adequacy of the developed algorithms is confirmed by the results of applying an intelligent geoinformation platform for the investigation of transport processes in Surgut (Khanty-Mansiisk autonomous district) State Traffic Safety Inspectorate of the Ministry of Internal Affairs, in Trekhgorny (Chelyabinsk region), Oktyabrsk (Samara region), Kinel-Cherkasy (Samara region) administrations.

## Acknowledgements

Special thanks to science and production center “Intelligent transport systems” for providing technical base for research.

## References

- [1] Burkov SM, Markelov GYa, Pugachev IN. Problems of system analysis and methodology for the formation of an intelligent management system for the city's transport complex. *Vestnik TSU* 2013; 4(31): 83–90.
- [2] Izyumsky AA, Kotenkova IN. Problems and perspective directions of development of intelligent transport systems in Russia. *Modernization and scientific research in the transport sector* 2013; 2: 206–211.
- [3] Kolosz B, Grant-Muller S, Djemame K. Modelling uncertainty in the sustainability of Intelligent Transport Systems for highways using probabilistic data fusion. *Environmental Modelling & Software* 2013; 49: 78–97.
- [4] Gatiyatullin Kh, Zagidullin RR. Intelligent transport system for large cities. *Vestnik NCBZhD* 2010; 5: 76–82.
- [5] Chong Y, Quek C, Loh P. A novel neuro-cognitive approach to modeling traffic control and flow based on fuzzy neural techniques. *Expert Systems with Applications* 2009; 36(3): 4788–4803.
- [6] Posmitny EV, Medovshchikov MI. The method of adaptive control of high-intensity traffic flows in city conditions on the basis of the meso model of dynamics using genetic algorithms. *Scientific journal of KubSAU* 2012; 84(10): 1–11.
- [7] Robinson VB. On fuzzy sets and the management of uncertainty in an intelligent geographic information system. *Recent Issues on Fuzzy Databases*. Physica-Verlag HD, 2000; 109–127.
- [8] Skvortsov AV, Boykov VN. Common data environment as a key element of the information modeling of highways. *CAD and GIS of highways* 2015; 2(5): 37–41.
- [9] Mikheeva TI. Construction of mathematical models of the city road network with the use of geoinformation technologies. *Information Technologies* 2006; 1: 69–75.
- [10] Skvortsov AV, Pospelov PI, Kotov AA. *Geoinformatics in the road sector*. Moscow: MADI, 2005; 250 p.
- [11] Christodoulo OI. A joint description of spatial and attributive data based on multidimensional information objects. *Software products and systems* 2011; 3(95): 48–54.
- [12] Kotsab M, Raday K. Integration of cartographic data into a single information system. *Geodesy, cartography and aerial photography* 2013; 78:127–131.
- [13] Kotikov YuG. ArcGIS in models of transport systems of megacities. *ArcReview, Data+, ESRI* 2013; 64: 18–19.
- [14] Pavlov CV, Efremova OA, Sokolova AV. The formalized description of spatial information in the composition of three-dimensional models of potentially dangerous objects on the basis of the set-theoretic approach. *Electrical and information systems and systems* 2014; 10(1): 66–72.
- [15] Mikheeva TI, Klyuchnikov VA, Golovnin OK. Methods and procedures of transport infrastructure survey. *Modern problems of science and education* 2014; 6. URL: <http://www.science-education.ru/120-16656>.
- [16] Mikheeva TI, Mikheev SV, Golovnin OK. Method of synthesis of zonal network-centric transport management system. *Proceedings of the Samara Scientific Center of the RAS* 2016; 4(4): 799–807.
- [17] Golovnin OK, Sidorov AV, Mikhailov DA. Support of decision-making of the automatic dislocation of transport infrastructure geobjects. *Proceedings of the Samara Scientific Center of the RAS* 2014; 4(2): 413–418.
- [18] Fedoseev AA, Mikheev SV, Golovnin OK. Data mining in problems of transport infrastructure development forecasting. *Modern problems of science and education* 2013; 1. URL: <http://www.science-education.ru/107-8153>.
- [19] Vasiliev SN, Oparin GA, Feoktistov AG, Sidorov IA. Intelligent technologies and tools for creating computing infrastructure in the Internet. *Computational technologies* 2006; S8(11): 34–44.
- [20] Belyakov SL, Bozhenyuk AV, Rosenberg IN. Adaptation of the procedure for visualization of spatial data by geoinformation services. *Izvestia of SFU* 2015; 3(164): 248–265.
- [21] Mikheev SV, Sidorov AV, Golovnin OK, Mikhailov DA. Architecture of geoinformation reference system of urban infrastructure objects. *Modern problems of science and education* 2013; 3. URL: <http://www.science-education.ru/109-9608>.
- [22] Open Geospatial Consortium Web Site. URL: <http://www.opengeospatial.org> (16.01.2017).
- [23] Geographic information system ITSGIS Web Site. URL: <http://itsgis.ru/> (16.01.2017).
- [24] Open Street Map Web Site. URL: <http://www.openstreetmap.org> (16.01.2017).

# Feature Selection Methods for Remote Sensing Images Classification

E. Goncharova<sup>1</sup>, A. Gaidel<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

Different methods of feature selection are used to improve the performance of remote sensing images classification. In this work two methods of feature selection are examined. The first one is based on the discriminant analysis, and the second one rests on building the regression model. Histogram and textural features are considered as characteristics of an image. The experiments on the remote sensing dataset UC Merced Land Use show the effectiveness of these methods. As the result, the largest fraction of correctly classified images accounts for the 95%. Dimension of the initial feature space consisting of 18 features has been reduced to 3 features.

*Keywords:* Feature selection; classification; remote sensing images; discriminant analysis; regression analysis

---

## 1. Introduction

Remote sensing images are a huge storage of data, which have become readily available lately. The analysis of such images allows us not only to enrich human's knowledge about the Earth but also to solve large number of applied problems. For example, to control the cultivation of croplands, trace the spread of crop pests, prevent forest fires, etc. To solve the outlined problems the high-level and effective methods of image processing should be developed.

The dimension reduction, or feature selection, is a crucial step in performing the classification task. This fact may be explained by the following reasons.

1. An image is described by various features, however their extraction requires large amount of resources. The more features are extracted, the more challenging the task is. Therefore, choosing the most informative features makes the classification cheaper and faster.

2. Each feature influences the object discrimination differently. Moreover, the classifier is not ideal, therefore it includes some error, which depends on the quality of feature space. Thus, uninformative and noise descriptors may complicate the process of building a prediction model.

There are a large number of feature extraction methods, which guarantee good performance. For instance, in [1] the combination of various descriptors was used to divide images into 19 classes. The mean portion of the correctly classified objects was 93.6%, in some classes it peaked at 100%. The problem of reducing the number of features for the purpose of pattern recognition was investigated in [2]. The feature space included several hundred thousand characteristics (pixels of the initial images), and its dimension was reduced to several dozens of features.

Various approaches for feature selection are widely used in the analysis of biomedical images. In [3] the group of 5 significant features was extracted from the set of 169 properties, which characterize the progress of the chronic obstructive pulmonary disease (COPD). The classification error rate of 0.11 was obtained using this reduced feature space.

In this work it is proposed to examine the histogram and textural features. The images for classification were received from the available UC Merced Land Use dataset, including aerial optical images, belonging to different classes (agricultural field, forest, beach, etc.). The two approaches of feature selection were proposed. The former was based on the discriminant analysis, the latter – on the regression model. To assess the performance of the proposed methods the nearest neighbor algorithm of object classification was applied.

## 2. The object of the study

The object of the study is the set of features, characterizing an image, and methods of selection the most informative subset of features, which has the strongest discriminatory power.

The histogram and textural image characteristics and a degree of their influence on the performance of dividing images into two classes are analyzed.

The first method of feature selection is based on the maximization of the discriminant analysis criterion and a greedy strategy of adding a feature to the informative subset. In the second method we propose to assess the importance of a feature according to its coefficient in the regression model. The greedy strategy of removing a feature with the minimal coefficient from the informative subset is used in the implementation of this method.

The set of image characteristics that should be considered to get accurate classification results was extracted via the use of these two methods. The k-nearest neighbors algorithm was implemented to perform the classification task.

### 3. Methods

#### 3.1. Feature extraction

An image is represented by its intensity matrix  $I^{(M \times N)}$ , where  $M \times N$  is an image size. The intensity of each pixel of image (RGB color space) is defined as follows:

$$I(m, n) = \frac{R(m, n) + G(m, n) + B(m, n)}{3}, \quad m = \overline{1, M}, \quad n = \overline{1, N},$$

where  $R, G, B$  is an intensity of red, green, and blue component of the image resolution cell having coordinates  $(m, n)$  respectively.

$I(m, n)$  ranges in value from 0 to  $L - 1$ , where  $L$  is a maximum gray level.

There are a large number of different features, which can characterize an image. In this work we use the histogram features that describe the spatial distribution of gray values. If the discrete image is considered as a two-dimensional stochastic process, we can estimate its spatial distribution of gray values and, therefore, raw (2) and central moments (3).

$$\nu_k = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N I^k(i, j). \quad (2)$$

$$\mu_k = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I(i, j) - \nu_1)^k. \quad (3)$$

The calculated features are:

– mean intensity:

$$\bar{I} = \nu_1, \text{ and also } (I_R, I_G, I_B - \text{mean intensity of red, green, and blue component respectively});$$

– second raw moment (mean energy):

$$s = \nu_2;$$

– standard deviation:

$$\sigma = \sqrt{\mu_2};$$

– skewness:

$$\gamma_1 = \frac{\mu_3}{\sigma^3};$$

– kurtosis (a measure of the “tailedness” of the probability distribution):

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3.$$

The autocorrelation matrix (4) describes dependence among the pixels of an image [4].

$$R(m, n) = \frac{\frac{1}{(M - |m|)(N - |n|)} \sum_i \sum_j I(i, j) I(i + m, j + n)}{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N I^2(i, j)}. \quad (4)$$

Two textural features are presented by the average of four values of the function (4) for two distances:

$$- r_1 = \frac{1}{4} (R(0, -1) + R(0, 1) + R(1, 0) + R(-1, 0));$$

$$- r_5 = \frac{1}{4} (R(0, -5) + R(0, 5) + R(5, 0) + R(-5, 0)).$$

Another type of textural characteristics is the widely known Haralick's features. Let  $P_{d, \theta}(i, j)$  be a frequency with which two pixels of image, separated by distance  $d_1$  in direction  $\theta$ , occur on the image with the intensity  $i$  and  $j$  respectively. Then the gray-level spatial dependence matrix can be build according to the following rule [5]:

$$P_{d_1, d_2}(i, j) = \{(m, n) \in \{1, 2, \dots, M\} \times \{1, 2, \dots, N\} \mid I(m, n) = i, I(m + d_1, n + d_2) = j\}, i, j = \overline{0, L-1}.$$

Textural features are extracted from the spatial dependence matrices, which are calculated for eight different distances  $(d_1, d_2)$ :  $(1, 0)$ ,  $(0, 1)$ ,  $(1, \pm 1)$ ,  $(2, 0)$ ,  $(0, 2)$ ,  $(2, \pm 2)$ . To get the invariant under rotation features, they are extracted from the average matrices. Thus, eight more textural features can be defined as follows:

– angular second moment:

$$f_1 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \left( \frac{|P(i, j)|}{R} \right)^2;$$

– contrast:

$$f_2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 \frac{|P(i, j)|}{R};$$

– entropy:

$$f_3 = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{|P(i, j)|}{R} \log_2 \left( \frac{|P(i, j)|}{R} \right);$$

– correlation:

$$f_4 = \frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} ij \frac{|P(i, j)|}{R} - M_x M_y}{\sqrt{D_x D_y}},$$

where  $P(i, j)$  – an element of averaged over the four dimensions  $(1, 0), (0, 1), (1, \pm 1)$  and  $((2, 0), (0, 2), (2, \pm 2))$ .

$R$  – a number of neighboring pixel pairs;

$M_x, M_y$  – the row and column means;

$D_x, D_y$  – the row and column variance.

### 3.2. Feature selection methods

Let  $\Omega$  be a set of objects for recognition. In this work a feature vector  $\mathbf{x}_k \subseteq \mathbf{R}^K$ , where  $K$  is a number of features, is considered as the element of this set. The set is divided into two classes  $\Lambda = \{\Omega_j\}_{j=1}^2$  with the following properties:

$$1) \Omega_0 \cup \Omega_1 = \Omega;$$

$$2) \Omega_0 \cap \Omega_1 = \emptyset.$$

Let  $\Phi(\mathbf{x}_k) : \Omega \rightarrow \Lambda$  be the ideal operator that puts an object in correspondence with its class. As long as the ideal operator is unknown, another operator  $\Phi(\mathbf{x}_k) : \Omega \rightarrow \Lambda$  can be created.  $\Phi(\mathbf{x}_k)$  tries to predict a class of input object, according to the information got from a training set of data  $U \subseteq \Omega$ , in which the outcome of object is observable.

As the features can be measured in varied units, firstly, they should be standardized to get zero mean and unit variance. For this purpose the expected value:

$$M(i) = \frac{1}{|U|} \sum_{k=1}^{|U|} x_k(i), i = \overline{1, K}, M \in \mathbf{R}^K$$

and variance:

$$R(i, i) = \frac{1}{|U|} \sum_{k=1}^{|U|} (x_k(i) - M(i))^2, i = \overline{1, K}, R \in \mathbf{R}^{K \times K}$$

should be estimated for each feature.

Therefore, the feature vectors can be standardized by applying the formula (5).

$$x_k(i) = \frac{x_k(i) - M(i)}{\sqrt{R(i, i)}}, k = \overline{1, |U|}, i = \overline{1, K}. \quad (5)$$

To extract the subset of informative features two methods were examined. The former belongs to the discriminant analysis theory. According to this method, we choose the set of features that provides the largest value of the criterion  $J(Q)$  [6]:

$$J(Q) = \frac{\text{tr } R}{\sum_{j=1}^2 P(\Omega_j) \text{tr } R_j},$$

where  $Q$  – current set of features;

$R$  – mixture covariance matrix;

$R_j$  – within-class covariance matrix;

$P(\Omega_j)$  – prior probability of class  $\Omega_j$ , there  $P(\Omega_j) = \frac{1}{2}$ .

Thus, the stronger the scattering between two classes exceeds the average within-class scattering, the better selected set of features is.

To form the set of the most informative descriptors a greedy strategy of adding a feature was applied. Let the initial feature set be empty –  $Q_{(0)} = \emptyset$ . In step  $i$  we consider all the sets, like  $Q_{(i,j)} = Q_{(i-1)} \cup \{j\}$ , and calculate the criterion  $J_{i,j} = J(Q_{(i,j)})$ .

Then choose the set that maximizes the criterion:

$$Q_{(i)} = Q_{(i-1)} \cup \left\{ \arg \max_{j \in [1;K] \cap \mathcal{Z} \setminus Q_{(i-1)}} J_{i,j} \right\} = Q_{(i-1)} \cup \left\{ \arg \max_{j \in [1;K] \cap \mathcal{Z} \setminus Q_{(i-1)}} J(Q_{(i-1)} \cup \{j\}) \right\}.$$

These steps are iterated until a required number of features are obtained.

The second approach is based on the regression analysis. The regression analysis estimates the relationships among the dependent variable and one, or more, independent variables.

We propose that the number of class, which  $\mathbf{x}_k$  can belongs to, is an independent variable  $y(\mathbf{x}_k)$ . This implies that the feature vector  $\mathbf{x}_k$  influences  $y(\mathbf{x}_k)$ , and the regression model (6) can be built as follows:

$$y = X\theta + \xi, \quad (6)$$

where  $y = (y_1 \quad y_2 \quad \dots \quad y_n)^T$  – output vector;

$X$  – feature matrix;

$\theta = (\theta_0 \quad \theta_1 \quad \dots \quad \theta_{|Q|})^T$  – regression weights;

$\xi = (\xi_1 \quad \xi_2 \quad \dots \quad \xi_n)^T$  – error vector.

The unknown coefficients belonging to the vector  $\theta$  are determined from the training set data via the ordinary least squares method:

$$(y - X\theta)^T (y - X\theta) \rightarrow \min_{\theta}.$$

The value of each feature is directly related to its weight in the regression equation (6). According to this proposal, the greedy strategy of removing a feature can be applied to forming the set of the informative descriptors.

Let the initial feature set  $Q_{(0)} = Q$  contain all the analyzed features. In each step  $i$  the linear regression model  $y_{(i)} = X_{(i)}\theta_{(i)}$  is built in the corresponding feature space. Then a feature with the minimal coefficient is removed from the set according to the following rule:

$$Q_{(i+1)} = Q_{(i)} \setminus \left\{ \arg \min_{j \in [1;K] \cap \mathcal{Z} \setminus Q_{(i)}} |\theta_{(i)}(j)| \right\}.$$

As in the previous case these steps are iterated until a required number of features are obtained.

To estimate the classification power of the obtained feature subsets the nearest-neighbor classification is carried out. The Euclidean distance in feature space is defined as follows:

$$\rho(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^K (x(i) - y(i))^2}.$$

The classifier assigns the class of the vector  $\mathbf{x}$  to the class of its closest point in the training set. In terms of the computational complexity, this method is rather simple in comparison with others. Since this classifier is memory-based, if the number of objects in the training set becomes large, this computational requirement may become excessive. The nearest-neighbor misclassification rate is no more than twice larger than the Bayes error rate [7].



The nearest-neighbor error rate is assessed as follows:

$$\varepsilon = \frac{|\{\mathbf{x}_k \in \mathbf{U} \mid \Phi(\mathbf{x}_k) \neq \Phi(\mathbf{x}_k)\}|}{|\mathbf{U}|}, \quad k = 1, \overline{|\mathbf{U}|},$$

where  $|\mathbf{U}|$  – test set.

#### 4. Results and Discussion

To assess the performance of the proposed approaches two image sets from the remote-sensing UC Merced Land Use dataset were used. This dataset includes aerial optical images, belonging to different classes (agricultural field, forest, beach, etc.), 100 for each class. Each image measures 256×256 pixels (RGB color space). There are two classes of images (agricultural fields and forest) being examined in this work. Figure 1 illustrates sample images belonging to the two classes.

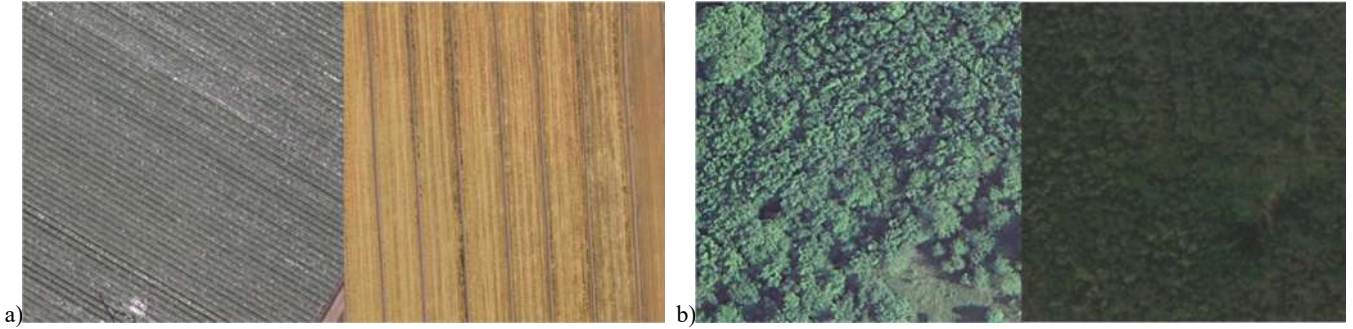


Fig. 1. Sample images from UC Merced Land Use dataset (a – agricultural field, b - forest).

To carry out the experiments we used 5-fold cross-validation. The results obtained with the discriminant and regression analysis methods are shown in tables 1 and 2 respectively.

Table 1. Groups of the first 8 informative features selected with the discriminant analysis.

Features	$\varepsilon$
$I_R$	0.5
$I_R, \bar{I}$	0.075
$I_R, \bar{I}, S$	0.05
$I_R, \bar{I}, S, I_G$	0.075
$I_R, \bar{I}, S, I_G, I_B$	0.225
$I_R, \bar{I}, S, I_G, I_B, r_1$	0.175
$I_R, \bar{I}, S, I_G, I_B, r_1, r_5$	0.175
$I_R, \bar{I}, S, I_G, I_B, r_1, r_5, \gamma_1$	0.2

Table 2. Groups of the first 8 informative features selected with the regression analysis.

Features	$\varepsilon$
$I_R$	0.5
$I_R, I_G$	0.075
$I_R, I_G, \bar{I}$	0.2
$I_R, I_G, \bar{I}, I_B$	0.175
$I_R, I_G, \bar{I}, I_B, r_5$	0.075
$I_R, I_G, \bar{I}, I_B, r_5, r_1$	0.1
$I_R, I_G, \bar{I}, I_B, r_5, r_1, S$	0.1
$I_R, I_G, \bar{I}, I_B, r_5, r_1, S, f_{22}$	0.275

Table 3 shows a so called confusion matrix for the group of three features, extracted by the discriminant analysis method and performed best on this task. Table rows show the real classes of objects, while the columns indicate the predicted ones. The fraction of objects that were predicted correctly is represented by the diagonal cells.

Having analyzed the results, we can conclude that the discriminant analysis method performed best on this classification task. The lowest classification error rate of 0.05 was achieved in three-dimensional feature space, consisting of  $I_R, \bar{I}, s$ . The studied textural features have no significant effect on the quality of this classification. The inclusion of more textural characteristics, considering the correlation of features on various distances, may provide a better performance of this feature group.

Table 3. Confusion matrix.

True class	Predicted class		
	agricultural	forest	
agricultural	100%	0%	
forest	10%	90%	
			95%

## 5. Conclusion

Thus, for the task of the remote sensing images classification the subset of informative features was extracted. On the images from the UC Merced Land Use dataset, the histogram features produced the best outcome. It should be mentioned that the images were represented in RGB color space; hence the mean intensity of these three components appeared to have considerable impact on the discriminatory power.

The feature vector, selected with the discriminant analysis method, produced the best classification performance (using the nearest-neighbor classification method) on the images from the UC Merced Land Use dataset. The minimal classification error rate made up 0.05, therefore the proportion of the correctly classified images was 95%. This rate was achieved in the reduced three-dimensional feature space, consisting of the descriptors  $I_R, \bar{I}, s$ .

Thus, applying the feature selection methods leads to improving the image classification performance. In this study, the combination of three of the 18 initial descriptors appeared to be informative, while the other features increased the misclassification rate.

The method based on the discriminant analysis criterion provided good results and can be applied to fulfill the task of feature selection. Overall, in the future work we are interested in considering more features, which can characterize an image, and multiclass classification that can enable us to get more universal results.

## Acknowledgements

The work was partially supported by the Russian Foundation of Basic Research (grant 16-41-630761 p\_a), the Russian Federation Ministry of Education and Science as a part of Samara University's competitiveness enhancement program in 2013-2020 and the RAS based research program "Bioinformatics, modern information technologies and mathematical methods in medicine".

## References

- [1] Guofeng Sheng, Wen Yang, Tao Xu, Hong Sun. Guofeng Sheng. High-resolutionsatellite scene classification using a sparse coding based multiple featurecombination. International Journal of Remote Sensing 2012; 33(8): 2395–2412.
- [2] Glumov NI, Myasnikov EV. Method of the informative features selection on the digital images. Computer Optics 2007; 31( 3): 73–76. (in Russian)
- [3] Gaidel AV, Zelter PM, Kapishnikov AV, Khramov AG. Computed tomography texture analysis capabilities in diagnosing a chronic obstructive pulmonary disease. Computer Optics 2014; 38(4): 843–850.
- [4] Gaidel AV, Pervushkin SS. Research of the textural features for the bony tissue diseases diagnostics using the roentgenograms. Computer Optics 2013; 37(1): 113–119. (in Russian)
- [5] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 1973; 3: 610–621.
- [6] Goncharova EF, Gaidel AV, Khramov AG. Statistical study of the factors affecting the cardiovascular disease. Information Technology and Nanotechnology 2016; 1020–1025. (in Russian)
- [7] Fukunaga K. Introduction to statistical pattern recognition . San Diego: Academic Press, 1990; 592 p.

# Development and study of methods for estimating retinal vessel parameters using a modified local fan transform

N.Yu. Ilyasova<sup>1,2</sup>, A.S. Baisova<sup>1</sup>, A.V. Kupriyanov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

Estimation of the geometric parameters of blood vessels is an important stage in the diagnosis of many cardiovascular diseases. In this work, we describe a method for estimating the diameter of blood vessels based on a modified local fan transform. We present experimental results that show in which way the accuracy of blood vessel estimation is affected by the noise-to-signal ratio in the image under analysis, vessel curvature radius, and the number of points and angles over which the averaging is done. The method is experimentally shown to be immune to various types of noise, structural complexity of the object, and variations in the vessel curvature radius.

*Keywords:* local fan transform; eye fundus; vascular image processing; local parameter estimation

---

## 1. Introduction

Early detection, analysis, and timely treatment of eye pathologies are critical in the prevention of eye-sight loss. Automatic detection and classification of eye diseases have recently become an important area of research, showing a tremendous potential for the early treatment of eye diseases [1, 2].

Glaucoma is a severe pathology of the eye which can be diagnosed by analyzing the state of the optic disk, or more precisely, the relationship between vascular parameters of different areas. By studying the derived information, the eye image can be classified as normal or containing a pathology. In a most quick way, the pathology can be detected by estimating the blood vessel width [3, 4]. According to modern data, the deviation of the diameter of a pathological vessel from a normal one is just about 20%. This fact is a major motivation behind the development of most precise methods for estimating the retinal arterioles.

The vascular disease is diagnosed using diagnostic features based on the geometric parameters of vessels. Among most important parameters are the vessel's diameter and direction.

Recent years have seen the development of many image processing algorithms aimed at analyzing both the vascular system in general and the optic disk, in particular. The algorithms use different approaches and have their benefits and shortcomings [5-12]. The algorithms based on a complex continuous wavelet-transform [12] are best suited for describing the structure of lines in different directions. A method for estimating vascular parameters proposed in [13] relies on mathematical morphology as an instrument for extracting image fragments suitable for the description of boundaries and skeleton of a vessel. The method utilizes a sparse representation of signals. In [13], each signal is assumed to be composed of a linear combination of several morphologically different components. The final map of a vessel is constructed using an adaptive threshold method. The method was shown to perform well in detecting anomalies and pathologies in the retina image. When used on their own, a major disadvantage of the morphological techniques is that they disregard the information on the vessel profile shape. Besides, while searching just for elongated structures, heavily twisted vessels can be missed out.

A method for estimating the vessel diameter using an algorithm based on a parametric model of an arbitrarily complex-shaped vessel was reported in Ref. [14]. The automatic algorithm is capable of segmenting the entire vascular tree, calculating vessel's diameter and direction in a digital ophthalmologic image. An algorithm that utilizes a new parametric surface model of the vessel intensity profile was described in Ref. [15]. Compared to other methods, the said method offers an advantage of robustness, whereas, as the authors mention, on the negative side is the dependence of the results on the test data and experimental conditions. If vessel diameters in the test data vary in a wide range, the measurement accuracy significantly deteriorates. Other approaches to measuring the diameter of vessels rely upon the approximation of the vessel brightness profile. With the vessel brightness normally having a Gaussian profile, a Gaussian curve is often used to approximate the vessel cross-section [16, 17]. However, with the diseased vessels tending to have well pronounced boundaries, their profile looks like a combination of two Gaussians, making it difficult to measure their diameter in an automatic mode and creating a possibility to falsely recognize a single vessel as two ones.

In this work, we propose a method for estimating local parameters of a fundus vascular system that exploits a modified fan transform, which is more noise-immune thanks to the additional filtering and noise averaging. Another advantage is that the method can perform a high-efficiency analysis of bifurcations, crossovers, and termination of vessels in the presence of interfering factors, such as close vessels.

## 2. A modified local fan transform

We introduce a modified local fan transform (MLFT) intended to operate with real-life low-quality vascular network imagery, which is well suited for processing images characterized by spots and closely spaced vessels.

With this method, the brightness distribution function is analyzed sector-wise depending on the sector's radius, size, and rotation angle (Fig. 1a) [2]. For each sector's position, the following local parameters are calculated: the average value and

variance of the brightness function  $f(x, y)$ . By analyzing these parameters as a function of angle, it is possible to detect the vessel in a given point, estimating its width and direction.

A modified LFT is defined by the following formulae:

$$F(x_0, y_0, a, q, r) = \frac{1}{S_q} \int_0^{\alpha} \int_0^r f(x_0 + t \cos j, y_0 + t \sin j) dt dj, \tag{1}$$

$$D(x_0, y_0, a, q, r) = \frac{1}{S_q} \int_0^{\alpha} \int_0^r [f(x_0 + t \cos j, y_0 + t \sin j) - F(x_0, y_0, a, q, r)]^2 dt dj, \tag{2}$$

where  $(x_0, y_0)$  is a point of measurement,  $a$  is a polar angle,  $q$  is a solid angle of the sector,  $r$  is the radius, and  $S_\theta = \theta R^2/2$  is the sector's area.

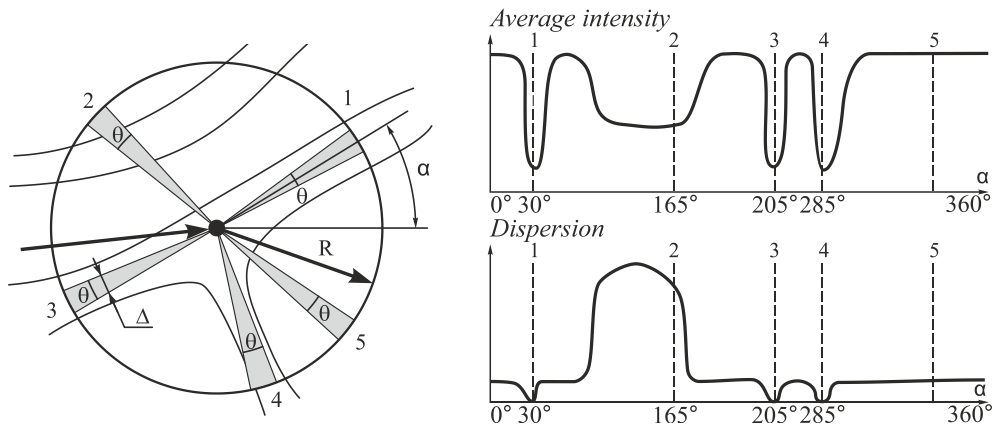


Fig. 1. A circular region with differently oriented sector.

### 3. Estimating the local direction of a vessel

The MLFT-aided method enables a local vessel direction to be evaluated. When compared with the LFT-aided approach, the MLFT method enables the radial profile with more pronounced minima to be obtained, with the local minima corresponding to bifurcation directions (Fig. 2).

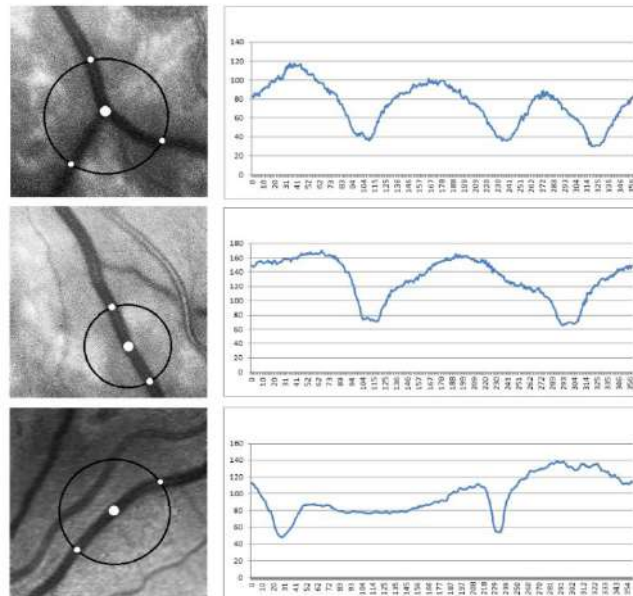


Fig. 2. Vessel fragments with corresponding radial profiles.

Thus, for a vessel direction to be identified, an optimization problem of searching for minima needs to be solved for each angle-dependent radial profile for a specific radius [2]. The algorithm operates by analyzing a list of directions, which represents a vector for each radius, with its length being equal to the number of directions and its magnitude taking a unit value if a bifurcation is detected and being zero otherwise. The list of directions may contain regions of constant values equal to unity.

This is the indication of detecting a bifurcation with larger-than-pixel width. In this case, the unit value is taken just at the central region's pixel.

#### 4. Estimating the vessel width

The vessel width will be evaluated using an LFT and on the assumption that the vessel is in parallel with the  $Ox$ -axis (Fig. 3).

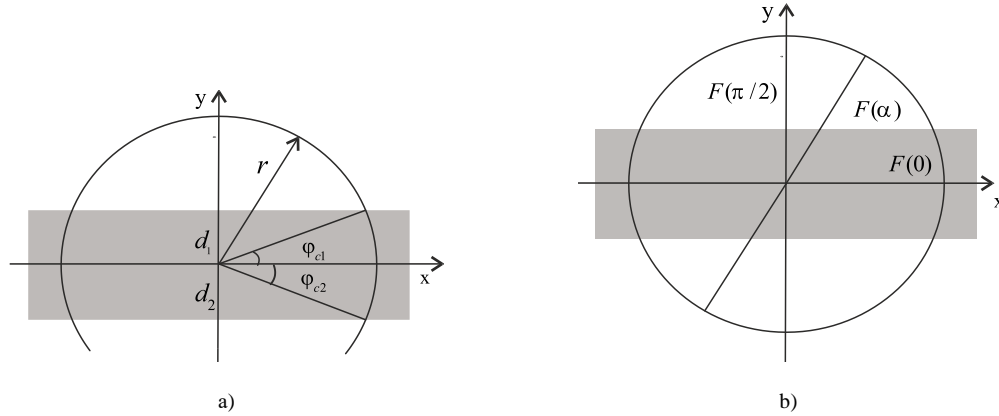


Fig. 3. Horizontal vessel section.

The fan transform of Eq. (1) can be analytically calculated at point  $(x_0, y_0)$  at  $\theta=0$  (Fig. 3a):

$$F_f(\alpha, r) = \left( r - \frac{d_1}{\sin \alpha} \right) f_0 + \frac{d_1}{\sin \alpha} f_1,$$

where  $f_0$  is the background brightness,  $f_1$  is the vessel brightness,  $d_1, d_2$  are width components with respect to the axis,  $\alpha \in (\varphi_{c1}, \pi - \varphi_{c2})$ ,  $\varphi_{c1} = \arcsin \frac{d_1}{r}$ ,  $\varphi_{c2} = \arcsin \frac{d_2}{r}$ . In the general case:

$$F(\alpha, r) = \begin{cases} (r - d_1/\sin \alpha) f_0 + f_1 d_1/\sin \alpha, & \alpha \in (\varphi_{c1}, \pi - \varphi_{c1}) \\ (r - d_2/\sin(\alpha - \pi)) f_0 + f_1 d_2/\sin(\alpha - \pi), & \alpha \in (\varphi_{c2}, \pi - \varphi_{c2}) \\ r f_1, & \alpha \in (0, \varphi_{c1}) \cup (\pi - \varphi_{c1}, \pi + \varphi_{c2}) \cup (2\pi - \varphi_{c2}, 2\pi) \end{cases}$$

Let us analyze a set of equations (Fig. 3b):

$$\begin{cases} F(0, r) = 2r f_1 \\ F(\pi/2, r) = (2r - d) f_0 + d f_1 \\ F(\alpha, r) = (2r - d/\sin \alpha) f_0 + f_1 d/\sin \alpha, \alpha \in (\varphi_{c \max}, \pi - \varphi_{c \max}), \varphi_{c \max} = \max(\varphi_{c1}, \varphi_{c2}) \end{cases}$$

Here, the angle  $\alpha$  is such that the interval of integration intersects both boundaries of the vessel. Hence, the diameter can be evaluated as

$$d = 2r \frac{F_r(\alpha, r) - F_r(\pi/2, r)}{\frac{F_r(0, r) - F_r(\pi/2, r)}{\sin \alpha} - F_r(0, r) + F_r(\alpha, r)} \quad (3)$$

In practice, diameter components  $d_1, d_2$  often need to be evaluated. Given a circular window of radius  $r$  and a horizontal vessel, the terms  $d_1$  and  $d_2$  can be described by similar sets of equations:

$$\begin{cases} F(0, r) = r f_1 \\ F(\pi/2, r) = (r - d_1) f_0 + d_1 f_1 \\ F(\alpha_1, r) = (r - d_1/\sin \alpha_1) f_0 + f_1 d_1/\sin \alpha_1, \alpha_1 \in (\varphi_{c1}, \pi - \varphi_{c1}) \end{cases},$$

$$\begin{cases} F(0, r) = r f_1 \\ F(3\pi/2, r) = (r - d_2) f_0 + d_2 f_1 \\ F(\alpha_2, r) = (r - d_2/\sin(\alpha_2 - \pi)) f_0 + f_1 d_2/\sin(\alpha_2 - \pi), \alpha_2 \in (\varphi_{c2} + \pi, 2\pi - \varphi_{c2}) \end{cases}$$

With a real vessel having an arbitrary direction, the algorithm for width estimation is, at first, given the vessel direction defined by an angle  $a_s$ , which is preliminarily calculated by the direction algorithm. With a specific value prescribed to the angle, the diameter components are evaluated as

$$d_1 = r \frac{F(a_1 + a_s, r) - F(p/2 + a_s, r)}{\frac{F(0 + a_s, r) - F(p/2 + a_s, r)}{\sin a_1} - F(0 + a_s, r) + F(a_1 + a_s, r)} \quad (4)$$

$$d_2 = r \frac{F(a_2 + a_s, r) - F(3p/2 + a_s, r)}{\frac{F(0 + a_s, r) - F(3p/2 + a_s, r)}{\sin(a_2 - p)} - F(0 + a_s, r) + F(a_2 + a_s, r)} \quad (5)$$

To enhance the accuracy of estimating the diameter components, a value averaged over  $N_\varphi$  different angles will be considered. Below, an estimate for a single diameter component is given:

$$d_1 = \frac{1}{N_\varphi} \sum_{k=1, \varphi_k \in (\varphi_{c1}, \pi - \varphi_{c1})}^{N_\varphi} \left( r \frac{F_f(\varphi_k + \alpha_s, r) - F_f(\pi/2 + \alpha_s, r)}{\frac{F_f(0 + \alpha_s, r) - F_f(\pi/2 + \alpha_s, r)}{\sin \varphi_k} - F_f(0 + \alpha_s, r) + F_f(\varphi_k + \alpha_s, r)} \right)$$

If Eqs. (3) and (5) are employed in a straightforward manner, the components of the ray and fan transforms  $F(a, r)$  [2,7,9] taken for a heavily noised vascular image can be calculated with an error, resulting in an incorrectly evaluated vessel's diameter. To avoid this in this work, the values of the LFT are averaged over points located uniformly on the perpendicular line of integration. Figure 4 illustrates the process of averaging three ray transform components.

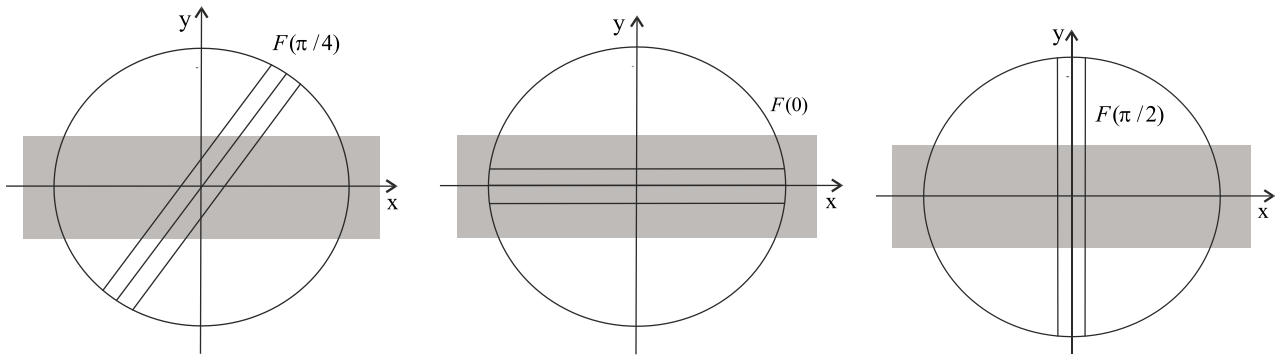


Fig. 4. An example of calculating a LFT using point-wise averaging.

### 5. Experimental study

We analyzed in which way the accuracy of estimating the vessel diameter depends on the noise-to-signal ratio for additive white noise. Figure 5 shows the relative diameter estimation error as a function of the noise-to-signal ratio:  $\varepsilon^2 = \frac{1}{N} \sum_{i=1}^N (\hat{d}_i - d)^2 / d^2$ , where  $\hat{d}$  is the evaluated diameter,  $d$  is the real diameter measured on a test image, and  $r$  is supposed to be not smaller than  $d/\sin(\pi/3)$ .

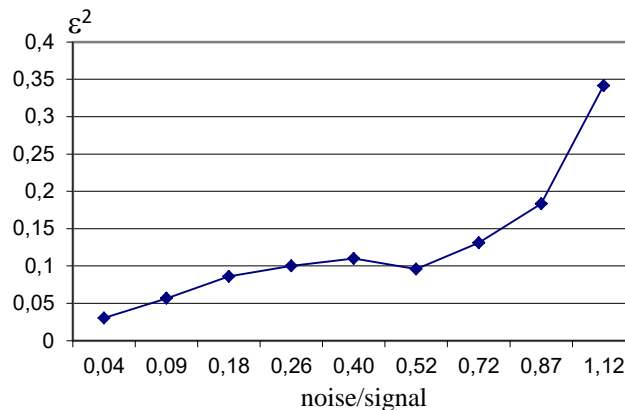


Fig. 5. Error of width estimation vs. the noise-to-signal ratio.

The study conducted on synthetic images has shown the method for estimating local parameters to be immune against the additive noise. For instance, the error of estimating the local diameter was found to be not larger than 8% given the noise-to-signal ratio under 0.25.

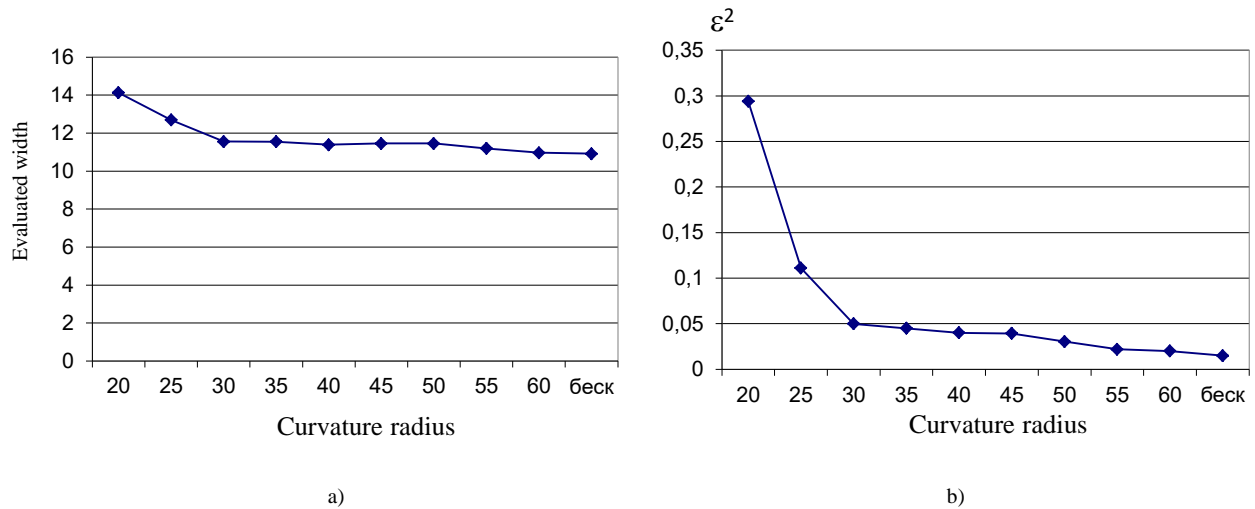


Fig. 6. (a) Evaluated width and (b) estimation error against the curvature radius.

The average vessel diameter estimate as a function of vessel curvature is shown in Fig. 6a. As test vascular images, we utilized the images of an annular sector 11 pixels in width and inner circle radius ranging from 20 to 60 pixels. A same-width straight-line segment (of infinite curvature radius) was also analyzed in order to determine an "estimated width". From the above plots, the error of width determination is seen to increase with increasing vessel curvature. Figure 7 depicts in which way the average vessel diameter estimate depends on the number of averaging points (see section 4, Fig. 4), given the signal-to-noise ratio equal to 25. The experimental results have shown that with the number of averaging points increasing to 13, the error falls from 0.1 till 0.02.

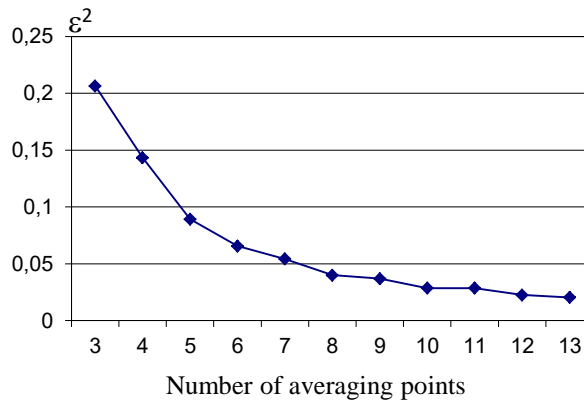


Fig. 7. The error of width estimation against the number of averaging points for the signal-to-noise ratio of 25.

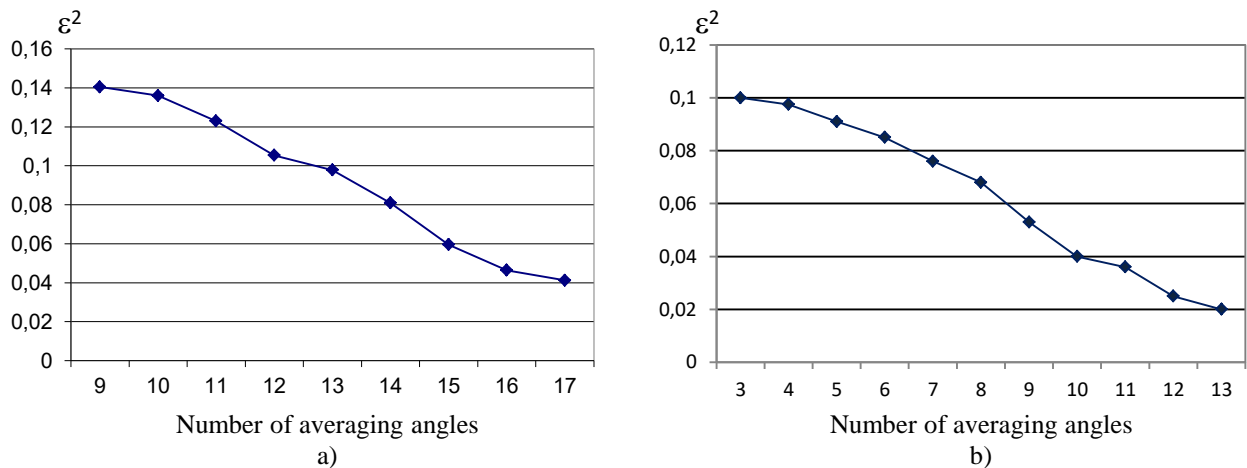


Fig. 8. The error of width estimation against the number of averaging angles given the signal-to-noise ratio of (a) 25 and (b) 15.

Figure 8 shows the diameter estimation error against the number of averaging angles (scanning sector size) with the r.m.s. signal-to-noise ratio, respectively, equal to 15 and 25. The experimental study showed that the greater is the volume of data used for averaging, the more reliably is the width estimation due to additional filtering of noise.

The experimental study showed that the estimation error can be essentially reduced by performing the averaging over a designated circumference sector of a local fan transform. As a disadvantage, we can mention that this approach is sensible to the highly curved vessels and the inability of the algorithm to be adjusted to the vessel and background brightness.

The directions were evaluated in a noisy 1024 x 1024 image (Fig. 9) with bifurcations uniformly located along the x-and y-axes. The general number of objects was 2,304.

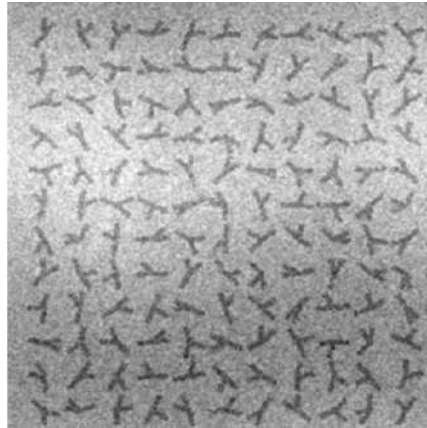


Fig. 9. Noisy test image.

A comparative study of the following methods for vessel direction and bifurcation identification was conducted: a direct method for direction identification (KM method), a method of a local discrete Radon transform, a method of a local discrete fan transform (LDFT) and a modified LFT [2]. The worst results were demonstrated by the algorithm based on the Radon transform variance estimation, because a classical Radon transform is unable to discern two opposite directions, thus leading to numerous cases of false recognition and the increased number of missed-out objects.

Table 1. Results of bifurcation detection using different techniques.

Method	Number of correctly recognized objects	Number of missed objects	Number of falsely recognized objects	General number of falsely recognized objects
KM method	2221	83	403	486
Discrete LRT	2016	288	192	480
LRT variance estimate	1812	492	235	727
MLFT method	2257	47	65	112
DLFT method	2240	64	222	286

The analysis showed that compared to other methods, the MLFT provides the least error of bifurcation angle estimation and the least error of false bifurcation recognition, is able to recognize correctly a larger proportion of bifurcations and the most stable to white noise.

## 6. Conclusion

Estimation of local blood vessel parameters used to form diagnostic features is a major problem of modern medicine, enabling an early diagnosis of various vascular pathologies. In this work, we have proposed a method for vessel diameter measurement that exploits a local fan transform. The method is based on the Radon transform, which is modified in such a way that bifurcations, crossovers, and terminations of vessels can be efficiently analyzed in the presence of interfering factors, including spots and close vessels. By analyzing the average brightness and variance of the radial function against the angle, vessel's width and direction can also be estimated and bifurcation points identified. To enhance the robustness, the transform is performed for a range of radii. The developed algorithm is stable to noise and disturbances, enabling bifurcations, crossovers, and terminations of vessels to be analyzed with high efficiency in the presence of interfering factors. The results of experiments have been discussed, showing in which way the accuracy of vessel width estimation is affected by the noise-to-signal ratio in the image under analysis, the vessel curvature, and the number of points and angles of averaging. The proposed method has been experimentally confirmed to be stable to various types of image noise.

The worst results have been shown by the estimation technique based on the Radon transform variance estimation because a classical Radon transform is unable to discern between two opposite directions, leading to a large number of false identifications and increased number of missed objects.



## Acknowledgements

This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world's leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29-03823, # 15-29-07077, # 16-41-630761; # 16-29-11698); by the ONIT RAS program # 6 "Bioinformatics, modern information technologies and mathematical methods in medicine" 2016 -2017.

## References

- [1] Partha Sarathi M, Dutta Malay Kishore, Singh Anushikha, Travieso CM. Blood vessel inpainting based technique for efficient localization and segmentation of optic disc in digital fundus images. *Biomedical Signal Processing and Control* 2016; 108–117.
- [2] Ilyasova NYu, Kupriyanov AV, Khramov AG. Information technologies of image analysis in medical diagnostics. M.: Radio I Svyazj, 2012; 424 p.
- [3] Astakhov YS, Krasavina MI, Grigoryeva NN. Modern approaches to the treatment of a diabetic macular edema. *Ophthalmologic sheets* 2009; 59–69.
- [4] Pedersen L, Grunkin M, Ersbøll B, Madsen K, Larsen M, Christoffersen N, Skands U. Quantitative measurement of changes in retinal vessel diameter in ocular fundus images. *Pattern Recognition Letters* 2000; 1215–1223.
- [5] Ilyasova N. Methods for digital analysis of human vascular system. literature review. *Computer Optics* 2013; 37: 517–541.
- [6] Ilyasova N. Computer Systems for Geometrical Analysis of Blood Vessels Diagnostic Images. *Optical Memory and Neural Networks (Information Optics)* 2014; 23(4): 278–286.
- [7] Kupriyanov AV, Ilyasova NYu, Ananin MA, Malapheev AM, Ustinov AV. Evaluation of the geometric parameters of the optic nerve region on the images of the fundus. *Computer Optics* 2005; 28: 136–139.
- [8] Ilyasova N. Estimation of Geometric Characteristics of the Spatial Structure of Vessels. *Pattern Recognition and Image Analysis* 2015; 25(4): 621–625.
- [9] Ilyasova NYu. Evaluation of geometric features of the spatial structure of blood vessels. *Computer Optics* 2014; 38(3): 529–538.
- [10] Ilyasova NYu, Kupriyanov AV, Paringer RA. The Discriminant Analysis Application to Refine the Diagnostic Features of Blood Vessels Images. *Optical Memory & Neural Networks (Information Optics)* 2015; 24(4): 309–313.
- [11] Kupriyanov AV, Ilyasova NYu. Development of information technology for estimation of geometrical parameters of the image of the fundus. *Bulletin of the Samara State Aerospace University. Academician of SP Korolev (National Research University)* 2008; 2.
- [12] Fathi A, Naghsh-Nilchi AR. Automatic wavelet-based retinal blood vessels segmentation and vessel diameter estimation. *Biomedical Signal Processing and Control* 2013; 71–80.
- [13] Elaheh I, Malihe J, Hamid-Reza P. Improvement of retinal blood vessel detection using morphological component analysis. *Computer Methods and Programs in Biomedicine* 2015; 263–279.
- [14] Konstantinos K, Aristides I, Tsonos C, Assimakis N. Automatic model-based tracing algorithm for vessel segmentation and diameter estimation. *Computer Methods and Programs in Biomedicine* 2010; 108–122.
- [15] Lupas A, Tegolo D, Trucco E. Accurate estimation of retinal vessel width using bagged decision trees and an extended multiresolution Hermite model. *Medical Image Analysis* 2013; 1164–1180.
- [16] Gao X, Bharath A, Stanton A, Hughes A, Chapman N, Thom S. Measurement of vessel diameters on retinal images for cardiovascular studies. *On-line Conference Proceedings: Medical Image Understanding and Analysis* 2001; 123–135.
- [17] Gao XW, Bharath A, Stanton A, Hughes A, Chapman N, Thom S. Quantification characterisation of arteries in retinal images. *Computer Methods and Programs in Biomedicine* 2000; 63(2): 133–146.
- [18] Cai Menga, Jun Zhang, Fugen Zhou, Tianmiao Wang. New method for geometric calibration and distortion correction of conventional C-arm. *Computers in Biology and Medicine* 2014; 52: 49–56.
- [19] Ilyasova NYu, Kazanskiy NL, Korepanov AO, Kupriyanov AV, Ustinov AV, Khramov AG. Computer technology for reconstructing the 3D structure of coronary arteries from angiographic projections. *Computer Optics* 2009; 33(3): 281–318.
- [20] Moravec J, Hub M. Automatic correction of barrel distorted images using a cascaded evolutionary estimator. *Information Sciences* 2016; 366: 70–98.

# Concerning the possibilities of successional changes revealing in anthropogenically transformed ecosystems on the base of remote sensing and ground-based survey data integration

L.M. Kavelenova<sup>1</sup>, N.V. Prokhorova<sup>1</sup>, E.S. Korchikov<sup>1</sup>, A.Yu. Denisova<sup>1</sup>, D.A. Terentyeva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The districts of Samara region are characterized by specific combination of orographic structure, hydrological regimes, soil and vegetation cover features, combined with a high level of anthropogenic pressure. The revealing of negative changes associated with the anthropogenic exploitation regimes including salinization and waterlogging after irrigation, soil erosion, transformation of non-cultivated fields into deposits, overgrowing of old quarries etc. seems to be a difficult task when carrying out by ground-based studies related to a large-scale land resources of the region. The use of remote sensing data, resulted by a time series of images for the same territory, opens up wide opportunities, on condition that the regional ground-based standardization is carried out.

*Keywords:* anthropogenic exploitation of ecosystems; overgrowing of non-cultivated fields and quarries; ground-based ecosystem survey; remote sensing data

---

## 1. Introduction

The Samara region is characterized by a complex combination of orographic, hydrological, soil and vegetation features, in combination with a high level of anthropogenic pressure. The share of agricultural land in the region approaches to the level of 77% whereas for some administrative regions it exceeds 90% (Alekseevsky, Bolshechernigovskiy, Bolsheglushitsky, Krasnoarmeisky, Pestravsky districts, where the share of tillage lands is more than 70%) and has the minimum levels in Syzransky, Shigonsky, Stavropolsky administrative regions where more than 40...50% belong to agricultural lands [1]. The result of plowing was the loss of most of the steppe lands, their plots could be preserved in point unusable for plowing (steppe yards, steep slopes of hills) [2, 3]., These components of the landscape often are characterized by a high degree of erosion hazard. The territory of the region demonstrates a tendency of the water logging processes intensification, especially for lands located in the zone of influence of the Kuibyshevskoye and Saratovskoye reservoirs and such large irrigation systems as Kutuluk, Vetlyanskaya, Spasskaya.

The total area of waterlogged agricultural lands is 127.1 thousand hectares or 3.3%, of which not in river valleys are 70.5 thousand hectares, including 53.4 thousand hectares of arable land, where water logging is mainly due to anthropogenic impact - [4]. Wetlands, mainly forage lands, occupy 25.7 thousand hectares or 0.7%, of which 0.4% are bogged up in an average degree.

In recent decades, the formation of deposits (not cultivated for a long period of arable land plots) has been observed in the region. On the non-cultivated fields (deposits) the successive changes in the vegetation cover take place including stages as: in the first 2-3 years of idleness, the arable grows with annual and biennial plants, in the next 5-7 years rhizome plants dominate it, further vegetation develops characteristic of the steppe conditions. Further unused arable land in the forest-steppe zone overgrows with bushes and trees. The abandoned land is the source of the spread of weeds to active arable land [5]. The area of lands overgrown by shrubs and trees on formerly agricultural land in 2015 in the Samara region was 18.7 thousand hectares. The total area of agricultural lands with saline soils is 110.1 thousand hectares or 2.9%, including arable land - 57.1 thousand hectares or 1.9%. According to the degree of salinity in the soil profile, slightly saline soils predominate with easily soluble salts exceeding the toxicity thresholds. As a result of the irrigation regime disturbance, for example, drainage lack in conditions of close saline groundwater bedding, 11.2 thousand hectares of secondary saline arable land were identified [4, 6]. Agricultural lands with solonchaks soils and solonchaks were found on an area of 156.1 thousand hectares, including arable land - 65.0 thousand hectares (or 4.1 and 2.2% respectively).

The identification of negative changes related to anthropogenic exploitation (soil erosion, salinization and waterlogging as a result of irrigation) and its cessation (conversion of non-cultivated fields into deposits, overgrowing of decommissioned quarries) presents a difficult task in carrying out ground-based research related to a large-scale land resources. The integrated terrestrial ecosystem surveys fulfillment, providing primary data for the subsequent remote sensing materials processing. Such materials seems to be presented by a temporary series of the same territory images, what opens up wide possibilities for monitoring the territory in aims of wide range of applied problems solving.

The specialized satellite imagery of resource assignment (Terra, Aqua, Landsat, etc.) makes it possible to receive information for a certain time period and with a certain spatial resolution. For retrospective monitoring elaboration, it is necessary to select and catalog satellite data, as well as to fulfill their processing, with the formation of spectral indices reflecting the state of different natural environment components. The methods of estimating recreational resources adjusted in this way allow to study and indicate the most threatened territories where intensification of unfavorable natural processes occurs, also as landscapes undergoing destruction, areas of phyto- and biodiversity reduction, as well as areas of increased man-caused and anthropogenic load [7]. As for our country, the practice of use time series images in the analysis of changes already finds its application in the

monitoring of saline lands, technogenically violated forests, wetlands [8-10]. The experience of such researches abroad is also quite rich (see, for example, [11-15]).

As applied to the tasks set by us, three objects of anthropogenically transformed ecosystems were proposed as control polygons for the detection of various stages of negative changes in time. For them, we had information on their state about 30 ... 40 years ago). Specificity of selected polygons allowed us to assess among the negative changes in the dynamics: 1) the formation of pseudo-forest formations with different species composition on the deposited fields; 2) development of salinized plots (solonchaks and solonchaks) under the influence of a closely located system of ponds (reservoirs); 3) the visualization of overgrowth stages on the bottom of the spent limestone quarry, which can be considered as model of the natural revitalization of the plant cover fully destructed by the open method extraction of mineral raw materials.

## 2. Methods

Three polygons were chosen as models for the detection of various stages of negative changes in time, in which 2 - 5 reference plots were allocated (the total number of reference areas was 11):

**1. "Neighborhoods of the Nizhnenikolsky village"** (rural settlement Domashka, Kinel municipal district of the Samara region). The reference plots studied in 2016 are located 1.5 km to the north-north-west (reference plot 1) 2 km to the south-south-west (reference plots 2-5) from the village of Nizhnenikolsky.

**2. "Neighborhoods of the Pekilyansky Reservoir"** (in the valley of the Gusikha river, 3 km to the south-west from the village of Pekilyanka, Bolshekhernigovsky district, Samara region). Within the boundaries of the polygon, the solonchak meadow and solonchak with yields of salt crystals on the soil surface were studied (reference plots 6, 7).

**3. The Ust-Soksky (Soksky, or Western) quarry** is located on the northern macroslope of the western part of the Sokoliye Mountains, a few kilometers from the confluence of the river Sock in the Volga river (Saratov reservoir), in the Krasnoglinsky district of the Samara city (reference plots 8-11).

For these objects, we had information relating to a retrospective assessment of their state (in particular, for Nizhnenikolsky and Ust-Soksky quarry - 30 ... 40 years ago).

During the ground-based complex survey, the coordinates of the central points were determined using the GPS-binding using the GarminEtrex navigator, the soil and vegetation cover survey including projective covering of plants, the primary plant species lists were compiled. The data obtained were used to prepare brief ecological characteristics of the reference plots.

The following images were used to identify the objects in polygon areas:

**1) "Geoton"**, spacecraft Resource-P, spatial resolution 0.8 m. To create a georeferenced image, panchromatic and multispectral images were used. The panchromatic image (black and white) has a resolution of 0.8 meters, the multispectral image is 2.4 meters and contains four channels: red, green, blue and infrared. To create a color image with a resolution of 0.8 meters, a panchromatic image was used to enhance the spatial resolution of the multispectral image, after processing which resulted in a complex image. The image was then re-calculated from the standard WGS-84 coordinate system into the local system of the Samara region and superimposed on the region map. The image was tied with a precision of 2 pixels, that is, up to 1.6 meters. For binding, a second-order polynomial transformation with support points and an accurate relief model was used.

**2) Spot-7**, spacecraft SPOT-7, spatial resolution of 1.5 m. To create a multispectral image with a resolution of 1.5 meters, software was used to process data from SPOT6 / 7 satellites. The image was then re-calculated from the standard WGS-84 coordinate system into the local system of the Samara region and superimposed on the region map. The image was attracted to within 2 pixels, that is up to 3 meters. For binding, a second-order polynomial transformation with support points and an accurate relief model was used.

**3) GoogleEarth.** GoogleEarth data are freely available on the Internet. These images have a high resolution, but inaccurate binding (sometimes error of up to 50 meters). Before processing the image data, they were saved in the tiff format, then manually bound using the Raster Manager in the InGeo GIS. It was used to bind the projective transformation using five or more control points.

Using the method of regression tree [16], classifiers were constructed for each polygons, which make it possible to extract territories with similar characteristics of soil cover in space images. As attributes for the classifier, brightness, red, green, blue, and near infrared spectral channels were used; the normalized difference The vegetative index (NDVI) based on NIR and G channels [17], the chlorophyll coefficient [18], the local average in the  $3 \times 3$  window, variance, correlation coefficients and entropy, textural signs of Haralick [19] and Gabor [20]. The resulting set of 41 traits using the principal component transformation led to reduced representation with a smaller number of features used for classification.

The experiments included the training of the classifier and its verification on remote sensing images with different spatial resolution. Since the data in the infrared channel was available only for images of Spot-7 and Geoton spacecraft, it was for these photographs that experimental studies were carried out. The spatial resolution of the images was 1.6 m and 0.8 m, respectively. The purpose of the experiments was to determine the best number of main components N to construct a classifier. The evaluation of the classification error was performed by cross-validation for 100 launches of training and classification procedures. In all experiments, a lot of data with known results of ground surveys were divided into training and control samples in the ratio 75:25 respectively.

### 3. Results and Discussion

The results of ground surveys allowed us to characterize the state of the parts of the polygon reference plots taking into account the specificity of the succession changes connected with the changes in anthropogenic exploitation state in past years.

Polygon "Neighborhoods of Nizhnenikolsky", located in the valley of the Samara river, demonstrates a leveled mesorelief with an absolute height of 45 ... 47 m above sea level. In general, the vicinity of Nizhnenikolsky village is confined to the valley of the Samara river (left bank), from the floodplain to the floodplain terraces and watershed. Steppe, meadow-steppe and ruderal-steppe plant communities, also as pseudo-forest plant communities at different succession stages, monocenoses of agricultural crops (in 2016 - sunflower fields, winter cereal crops, etc.) are represented here. The relief is mostly heterogeneous, with frequent depressions rounding lakes, altered with plane elevations. Until the 1990s, plains were almost completely used for agriculture, forests with the participation of various willows and poplars and an admixture of other tree species filled the depressions along the channels of temporary watercourses and formed along the lakes shores. Later, the plowing of some of the lands ceased, and the development of deposits began on them, in recent years some fallow lands have been reintroduced into agricultural land. The reference plots laid in 2016 are represented by fallow areas (cessation of agricultural cultivation for more than 20 years), which are in the process of development of the steppe (reference plot 4), semi-shrub (reference section 3) and pseudo-forest communities (reference plots 1, 2 And 5).

Reference plot 1 is a non-cultivated field on which a steppe community (mixed-grass-grass associations) was formed with *Ulmus pumila* L. (karagach), undergrowth of trees and shrubs. The occurrence of such communities with the participation of the *U. pumila* is observed in the steppe regions of the Samara region on elevated leveled relief forms in the absence of excessive moisture. The tree component of the non-cultivated field is characterized by a height of up to 8 m, a crown density up to 0.2 (20%). The share of the open soil surface was less than 1%. The site shows fresh porpoises of wild boars. The reference plot 2 is a non-cultivated field transformed to a pseudo-forest community formed by *Elaeagnus angustifolia* L., aged over 15 years, with a height of 6-8 m, crown diameter from 3 to 5 m, an average crown density up 0.5 (50%). The grass cover is formed by herbage-grass association. The open surface of the soil is not expressed. Harvesting of farmland by *E. angustifolia* is typical for the southern parts of the region in relief depressions with partial soil salinization. Reference plot 3 is a non-cultivated field, now transformed to a steppe community in the form of a wormwood-grass association with wormwood dominating. The height of the grass reaches 1 m with a projective covering is more than 80%. The open surface of the soil is not expressed. In the places where the cattle are run, the ruderal nature of plant cover is presented clearly. Reference plot 4 is a non-cultivated field, now turned into a steppe herbage-grass association with dominance of grasses. The height of the grass is on the average 30 cm, the projective covering is more than 80%. The open surface of the soil is not expressed. Reference plot 5 is a non-cultivated field, now is a steppe community in the form of a herbage-grass association on the initial stage of *E. angustifolia* overgrowth. The projective coverage of the grass stand is more than 80%. Young *E. angustifolia* trees have a height of up to 1.5 m with a crown density sometimes up to 0.6 (60%), and an average of 0.35 (35%). The open surface of the soil is also not expressed.

Experiments on the classification of the Nizhnenikolsky areas showed that the lowest classification error is achieved with the use of the 9 main components of the system of characteristics under consideration, while the probability of correct classification for images of the Spot-7 was 78% and 74% for the "Geoton" data. The reduction of detection efficiency in comparison with a more coarse resolution image should be attributed to the use of small spatial windows to calculate texture attributes ( $3 \times 3$  pixels windows were used in the experiment). As a result, in a higher resolution image, a larger number of points were required for a single surface object of the same size, so larger windows should be used to characterize the intensity of brightness changes within a certain window. However, increasing the window size with a larger image size will significantly increase the processing time, therefore it is recommended to use the images of the Spot-7 with a resolution of 1.6 m. The study shows that the best imade identification was achieved for pseudo-forest communities. Steppe communities were classified the worst. Pseudo-forest communities according to subspecies (dominated by *Ulmis pumila* or *Elaeagnus angustifola*) can not be classified relative to each other.

Table 1. Formalized results of classifier training and its verification on remote sensing images of reference plots.

Criteria	The results of image processing	
	SC Spot-7	SC "Geoton"
"Nizhnenikolsky"		
Optimum number of main components N	9	9
Probability of correct detection	78%	74%
"Neighborhoods of the Pekilyansky Reservoir"		
Optimum number of main components N	3 (from 3 to 12)	9
Probability of correct detection	99%	98%
Ust-Soksky quarry		
Optimum number of main components N	9	9 (from 6 to 9)
Probability of correct detection	92%	85%

The classifier constructed for the plots 2-5 of the polygon "Nizhnenikolsky" was used for the plot 1. The classification results (N = 9) showed that the site number 1 is classified in the same group as the site number 2, that is, refers to the pseudo-forest

community. However, as was mentioned above, it was not possible to establish differences in the species composition of the tree layer of these pseudo-forest communities (overgrowing with *Ulmis pumila* or *Elaeagnus angustifolia*) during the analysis of the plot images.

The polygon "Neighborhoods of the Pekilyansky Reservoir" is confined to the southern part of the Samara Region, which is characterized by a high degree of plowing of the indigenous steppes. The reduction of the cultivated lands area in the late 1990s led to the formation of non-cultivated fields with various forms of ruderalized, steppe and meadow plant communities. Cattle grazing, in some places haymaking are carried out.

The formation of ponds and reservoirs in the damming of watercourses (for the considered landfill, the Gusikha River), which are used for water supply, irrigation and fish farming, is a characteristic feature of territory. The secondary salinization of the soil cover occurs in the adjacent areas of these ponds. The Gusikha river valley possessing solonets soils complex is used for hayfields and pastures. Reference plot 6 is located to the north-east of Pekilyansky reservoir, the creation of which resulted in soil cover secondary salinization. The vegetation is represented by a solonetz meadow, in the herbaceous cover with domination of *Artemisia pauciflora* Web. and *Atriplex verrucifera* Bieb. The average projective coverage of the grass is 85%, the open soil surface is about 5%. The degree of salinity can be characterized as moderate, what is confirmed by the nature of the formed ecosystems. The reference plot 7 is also located to the north-east of the Pekilyansky reservoir. Its soil cover is characterized by a pronounced secondary salinity, that was clearly indicated by changing the color of the grass stand, manifested from the middle of the vegetative period. The soil is represented by a solonchak, with typical halophytes vegetation, where *Suaeda corniculata* (C.A. Mey.) Bunge and *Salicornia europaea* L. dominate. The average projective coverage of the grass is 45%, and areas with open surface soil - 55%. On the surface of the open soil there is a whitish-yellowish crust of salt deposits, extending from the surface to a depth of 5 mm.

When working with Spot-7 images of the polygon "Neighborhoods of the Pekilyansky Reservoir", the previously described procedure was fulfilled and showed that 3 to 12 main components of the characteristics system can be used to classify the etalon areas, the classifier provides a correct classification probability of about 99% (See table 1). It should be noted that it is reasonable to select as few signs as possible to accelerate the calculations, therefore the optimal choice in this case can be considered 3 main components. The training and classification results for "Neighborhoods of the Pekilyansky Reservoir" according to the data of the "Geoton" showed that the application of 9 main components led the best result, with the classifier providing the correct detection probability of the level of 98%. The images combination showed that the classifier error should be attributed to the displacement of the mask relative to the snapshot, which indicates the need for a very accurate snapping for the images that the classifier is trained on. The study showed that strong soil salinity points with 55% bare soil surface (plot 7 of the polygon "Neighborhood of the Pekilyansky Reservoir"), is well diagnosed from images.

Ust-Soksky (Soksky, or Western) quarry is located on the northern macro-slope of the western part of the Sokoliye Mountains, a few kilometers from the confluence of the river. Sok in the Saratov Reservoir in the Krasnoglinsky district of Samara city. It represents the oldest site of the Soksky carbonate deposit, where industrial extraction of carbonate rocks was carried out for the production of building materials (crushed stone, quarry stone, construction mixtures). As a result on the northern slope of the Sokol'y Mountains, a large trough-shaped technogenic excavation arose having maximum length along the bottom from north to south about 1 km and from west to east more than 2 km. The relative height of the steep sides of the man-made trench reaches tens of meters, in some cases - 100-150 m. Many times after the operation of the quarry the new dumps of household and construction debris (no more than 5% of the floor area) were created. This time, the sub-eastern side of the quarry is storing off-grade rock from the Central and Eastern sections of the Soksky deposit, which does not affect the bottom of the quarry, but is distributed only along its terraces. The recent years trend has been the use of the central and western parts of the bottom under the shooting range, as well as the organization of walking tours and quadrocycles to the local lake and entrances to the tunnel. Since the early 70's. XX century the industrial extraction of building materials in the Ust-Soksky quarry was stopped and the processes of natural self-growth and primary soil formation began to develop. The reference plot 8 (in the eastern part of the Ust-Soksky quarry) presents the forest amphicycnosis formed during 40 years overgrowth of trees of *Populus nigra* L., *Betula pendula* Roth. and *Pinus sylvestris* L. (1: 1: 1) on bottom of the quarry. The grass cover under the canopy of tree vegetation and on open positions practically is not expressed. The height of the tree layer is up to 8-10 m, the crown density is 0.6 (60%), the share of the open rocky surface (dolomites, limestones) is 0.4 (40%). Reference plot 9 (in the central part of the quarry near the lake) can be described as forest amphicycnosis, formed in the process of 40-year overgrowing by *Populus nigra* L. and *Salix caprea* L. in the ratio of 9: 1. The grass cover is almost indistinguishable, but on the stony substrate there are separate clumps of green moss. The height of the tree storey is 6-9 m, the crown density is 0.8 (80%), the share of the open rocky surface (dolomite, limestone) is 0.1 (10%). Reference plot 10 (in the central part of the quarry opposite the galleries) presents forest amphitocenosis formed in the process of 35 years overgrowing by *Populus nigra* L., *Pinus sylvestris* L. and *Betula pendula* Roth. in a ratio of 6: 2: 1 with an admixture of *Populus tremula* L. and various willow species. The grass cover is almost not expressed. The height of the tree layer is 3-5 m, the crown density is 0.3 (30%), the share of the open stony surface (dolomites, limestones) is 0.7 (70%). The reference plot 11 (in the western part of the quarry) was formed during 35 years of overgrowing as forest amphitocenosis with a stand represented by a *Populus nigra* L., *Pinus sylvestris* L. and *Betula pendula* Roth. in a ratio of 4: 4: 2. The grass cover is almost not expressed. The height of the tree layer is 1.5-3 m, the crown density is 0.2 (20%), the share of the open rocky surface (dolomites, limestones) is 0.8 (80%).

For the polygon Ust-Soksky quarry plots, according to the previously described methodology, it was obtained that 9 or more major components of the feature system provide a high quality of classification, while the probability of correct classification is approximately 92% (see Table 1). The results of training and classification of instrumentation and instrumentation "Ust-Soksky quarry" according to the GA "Geoton" allowed to conclude that here for the identification of plant communities can be used

from 6 to 9 main components of the characteristics system. The study showed that the parts of the Ust-Sok quarry can be classified using remote sensing data, while a smaller classification error is ensured by using images of the Spot-7 spacecraft with 9 main components in the system of signs.

#### 4. Conclusion

Thus, the results of complex field surveys on three test sites (polygons) in different parts of Samara region, integrated with the selection of RS data Spot-7, Geoton, GoogleEarth for this territory, were used for the development of in the assessing territorial resources methodology with geodata classifiers. The use of regression tree method based on textural features number showed the undoubted promise of this approach in identifying of regionally significant negative symptoms of land cover state. With the smallest error in succession changes revealing in anthropogenically transformed ecosystems, it is advisable to use images of the SPOT-7 spacecraft with a spatial resolution of 1.5 m. It is important to develop this methodology for achieving in future a more detailed classification of regional vegetation types in photographs, in particular – different types of forest communities.

#### Acknowledgements

The authors are grateful to the Autonomous State Institution of the Samara Region "Center for Innovative Development and Cluster Initiatives" for the opportunity to carry out this study.

#### References

- [1] Poroshina LN. The Atlas of the Lands of the Samara Region. Samara, 2002; 101 p. (in Russian)
- [2] Kavelenova LM, Rozno SA, Pomogaybin AV, Ruzaeva IV, Zhavkina TM, Soboleva MN, Pomogaibin EA, Demenina LG. Some aspects of the preservation of phytodiversity in anthropogenically transformed environment (by the example of the Samara region). *Izvestiya of the Samara Scientific Center of the Russian Academy of Sciences* 2012; 14: 1(9): 2233–2236. (in Russian)
- [3] Kavelenova LM, Prokhorova NV, Golovlyov AA, Rozno SA. Preservation of phytodiversity as an integral part of sustainable development strategy in the Samara region. *Volga Ecological Journal* 2014; 1: 12–20. (in Russian)
- [4] State report on the state of the environment and natural resources of the Samara region for 2014. Samara, 2015; 25: 298 p. (in Russian).
- [5] Ledovskikh AA, Kalashnikova EB. The problem of "abandoned" lands in the Samara region. *Approbation* 2015; 6(33): 107–109. (in Russian)
- [6] State report on the state of the environment and natural resources of the Samara Region for 2015. Samara, 2016; 26: 296 p. (in Russian)
- [7] Shevryev SL, Antsiferova GA, Shevryev MZh. On the satellite monitoring of mining enterprises of Primorsky Krai (on the example of the Pavlovsky-2 section). *Bulletin of the VSU. Ser. Geology* 2015;2: 128–133. (in Russian)
- [8] Lyamina VA, Glushkova NV, Smolentseva EN, Zolnikov ID. Use of GIS and ERS methods for monitoring the area of lakes and solonchaks in the south of Western Siberia. *Interexpo Geoiberia 2010*; 2: 3–7. (in Russian)
- [9] Cherosov MM, Ammosova EV, Savvin TI, Vinokurov EN, Tarasov IM. Experience in the use of GIS technologies and remote sensing for assessing the impact of anthropogenic factor on the vegetation of individual territories of Yakutia. *Advances in modern natural science* 2012; 11(1): 63–65. (in Russian)
- [10] Glushkova NV, Chupina DA, Kotler S.A. Analysis of the dynamics of saline complexes for assessing the degree of aridization of the territory of western Siberia on the basis of GIS and DZ. *Interexpo Geo-Siberia 2013*: 2: 77–81. (in Russian)
- [11] El-Asmar HM, Hereher ME. Change detection of the coastal zone east of the Nile Delta using remote sensing. *Environ. Earth Sci.* 2011; 62(4): 769–777.
- [12] Rebelo LM, Finlayson CM, Nagabhatla N. Remote sensing and GIS for wetland inventory, mapping and change analysis. *J. Environ. Manage* 2009; 90: 2144–2153.
- [13] AlaviPanah SK, Goossens R. Relationship between the Landsat TM, MSS data and soil salinity. *J. Agric. Sci. Technol.* 2001; 3: 21–31.
- [14] Eldeiry AA, Garcia LA. Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. *Soil Sci. Soc. Am. J.* 2008; 72(1): 201–211.
- [15] Kalra NK, Joshi DC Potentiality of Landsat, SPOT and IRS satellite imagery, for recognition of salt affected soils in Indian Arid Zone. *J. mote Sens* 1996; 17(15): 3001–3014.
- [16] Kuznetsov AV, Myasnikov VV. Comparison of algorithms for controlled element-wise classification of hyperspectral images. 10; 494–502. (in Russian)
- [17] Pettorelli N, Mysterud AGaillard J-M. Tucker CJ, Stenseth NC. Using the satellite-derived NDVI to assess ecological responses to environmental. *Trends in ecology & evolution.* 20; 9: 503–510.
- [18] Marx A. Erkennung von borkenkäferbefall in fichtenreinbeständen mit multi-temporalen rapideye-satellitenbildern und datamining-techniken. Mysterud, J-Photogrammetrie, Fernerkundung, Geoinformation 4: 243–252.
- [19] Haralick RM. Textural features for image classification. *Environ. Manage* 2009 RJ; 90. *J. Environ. Manage* 2009; 90. *IEEE Transactions on systems, man, and cybernetics*, 1973; 6: 610–621.
- [20] Fogel I. Gabor filters as texture discriminator. *Biological cybernetics* 1989; 61(2): 103–113.

# Method of automated epileptiform seizures and sleep spindles detection in the wavelet spectrogram of rats' EEG

I.A. Kershner<sup>1</sup>, Yu.V. Obukhov<sup>1</sup>, I.G. Komoltsev<sup>2</sup>

<sup>1</sup>*Kotel'nikov Institute of Radio Engineering and Electronics of RAS, Mokhovaya 11-7, 125009, Moscow, Russia*

<sup>2</sup>*Institute of Higher Nervous Activity and Neurophysiology of RAS, Butlerova 5A, 117485, Moscow, Russia*

---

## Abstract

A method and algorithm for automatic detection of epileptiform seizures, sleep spindles, and other high voltage rhythmic activity were developed. They based on the analysis of the ridges of EEG wavelet-transformation. The uninformative points of the ridge are removed adaptively on the basis of power spectral density histograms analysis.

*Keywords:* Traumatic brain injury; EEG; Wavelet; Spectrogram; Ridges; Epileptiform seizures; Sleep spindles; Event detection

---

## 1. Introduction

The study of long-term electroencephalographic (EEG) signals of patients who have suffered from traumatic brain injury (TBI) to detect markers of posttraumatic epilepsy (PE) [1] is an unsolved issue. Immediate and early seizures within the first week after PTI are important risk factors for appearance of late convulsive seizures, which are a manifestation of PE. Early seizures are associated with brain damage, while late ones are associated with the processes of restructuring the neuronal connections and many other changes called epileptogenesis. Late convulsive seizures can develop months or even years after TBI, as epileptogenesis proceeds extremely slowly and asymptotically. At the moment there are no clear EEG criteria for this pathological process. Therefore, the detection of biomarkers of PE in the acute period of TBI is of great importance for timely diagnosis, as well as researches of new methods of preventing epilepsy.

In the study of neurobiological mechanisms of epileptogenesis, animal models (rats) are widely used. The most adequate of these is the model of posttraumatic epilepsy caused by injury, resulting from lateral fluid percussion (LFP) [2,3]. The first unprovoked seizures in rats occur months after the injury. The appearance of epileptiform activity in the EEG signals in the early post-traumatic period (first week) can serve as a predictor of the PE development. For the detailed characterization of the early post-traumatic period, it becomes necessary to automatically detect epileptiform discharges (ED) in long-term (day, week) EEG records.

In recent years, much attention has been paid to the automatic detection of epileptiform discharges in patients with confirmed diagnosis of epilepsy. This is due to several reasons. Among them: the need to predict the emergence of seizures in order to prevent them by use electrostimulation, difficulties in processing long-term EEG recordings, the need to classify the forms of epilepsy, and a number of other reasons [4].

A review of the works on automatic epileptic seizures detection is presented in [5, 6]. In [5], the EEG signal is decomposed into empirical modes, for which the standard deviation, the asymmetry coefficients and the kurtosis are calculated. These parameters were entered in the learning machine. Then the part of the signal is classified as a seizure activity or background activity. This method does not give knowledge about the threshold criteria by which epileptic discharges differ both from background activity and from other high-amplitude activity in the EEG signal (hereinafter we will call such activity an event).

In [6], the parameters for the seizure classification are skewness and kurtosis coefficients, the Fourier peak frequency of the spectrum, the median of the frequency, entropy, correlation dimension, and variance of the EEG signals. However, as in [5], epileptic discharge is distinguished from the background activity, but other high-amplitude events present in the signal were not considered.

In [7] studied sleep spindles (SS). To detect them, a method based on the analysis of the wavelet transform was used. The calculation of the mean value over the time-frequency rectangles of the instantaneous energy of the wavelet transform was carried out. After that, in comparison with the parameters inherent in EEG signal background activity, the conclusion was made whether this event is a sleep spindle or not. This method does not consider the presence in the EEG signals of such high-energy activity as epileptiform discharges. Although in works [8-10] it is said about the possibility of transformation the sleep spindle into the peak wave's discharge. In long-term EEG records (day, week), in addition to epileptiform discharges, there are other high-energy activities that differ from the background EEG signal, such as sleep spindles. SS, as well as ED, belong to the group of high-amplitude brain rhythmic electrical activity. In humans and animals with absence epilepsy, the frequency range of SS and ED ranges from one to fifteen Hz [7,11-16].

Automatic detection of sleep spindles and epileptiform discharges in the early post-traumatic period, in which the mechanisms of the occurrence of epileptiform activity differ from those that occur in epilepsy, is an unresolved task.

As in [13-16], in order to investigate the time-frequency dynamics of the EEG, we use the ridges of the Morlet wavelet transform. However, in contrast to these works, when the beginning of epileptiform discharges was set by the expert manually, in this article we describe the method of automatically finding the beginning and end of high-amplitude activity, and calculating its parameters.

## 2. The method of automatic detection of events containing high-amplitude rhythmic activity

Long-term EEG records represent a large array of  $\sim 10^8$  numeric data. Typically, the EEG is measured at a sampling rate of 250 Hz. EEG signals were divided into 10-minute intervals, as there is a limit to the amount of data that can be processed in the Matlab.

To remove linear trends, power supply noise and low-frequency noise, daily fragments of EEG records were filtered by a 16th-order Butterworth discrete filter with a bandwidth ranging from 2 Hz to 30 Hz. The bandwidth of the filter exceeds the frequency range typical for ED and SS. The signal is filtered in two stages. At the first stage, synthesis of 8th order discrete bandpass filter with a bandwidth ranging from 2 Hz to 30 Hz was realized by using function "butter". As a result, the transfer function  $H$  in decreasing order of powers of the variable  $z$  was obtained:

$$H(z) = \frac{\sum_{i=1}^{n+1} b(i) * z^{1-i}}{1 + \sum_{i=2}^{n+1} * z^{1-i}} \quad (1)$$

Where  $n = 8$  is the order of the filter.

In the second stage, the phase shift was compensated. By means of the "filtfilt" function, discrete filtering using the Fast Fourier Transform (FFT) is implemented in conjunction with the division of the signal into blocks. The signal is filtered from the beginning of record to its end, then obtained signal is filtered a second time - from the end to the beginning. Thus, the phase shifts were compensated, and the resulting filter order was doubled:  $n = 16$ .

The result of filtration of 10-minute signal fragment is shown in Fig. 1.

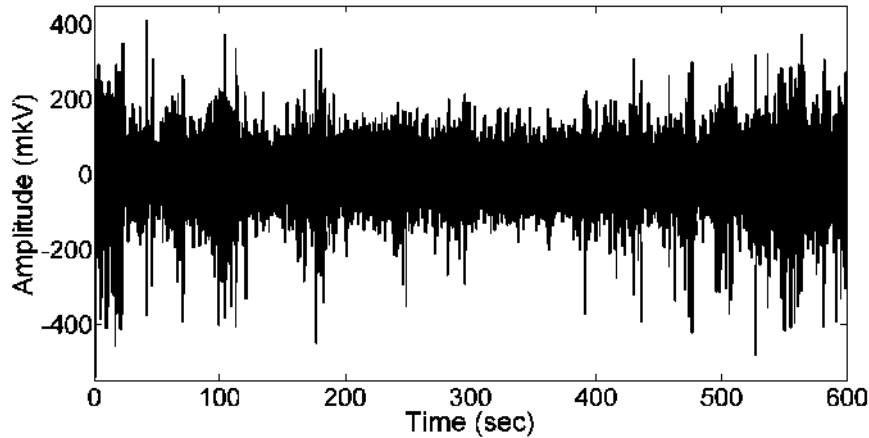


Fig. 1. 10 minute signal fragment after filtration. The sampling frequency is 250 Hz.

The automatic detection method of high-amplitude brain rhythmic electrical activity is based on the analysis of wavelet spectrogram ridges [17]. To calculate the wavelet spectrograms, a complex Morlet wavelet transform was used (2):

$$W(\tau, f) = \frac{1}{\sqrt{f}} \int x(t) * \psi\left(\frac{t-\tau}{f}\right) dt \quad (2)$$

In the formula (2)  $x(t)$  refers to the source signal, and  $\psi(\eta)$  refers to the Morlet mother function:

$$\psi(\eta) = \frac{1}{\sqrt{\pi F_b}} e^{2i\pi F_c \eta} e^{-\frac{\eta^2}{F_b}} \quad (3)$$

The coefficients  $F_b = F_c = 1$ . The power spectrum density (PSD) of a time-frequency signal is calculated according to function (4):

$$S_x = |W(\tau, f)|^2 \quad (4)$$

The ridge consists of the points  $y(i)$  with the maximum values of the power spectral density in each time count of the wavelet-spectrogram:

$$y(i) = \max_{f \in (2-20 \text{ Hz})} (S_x(\tau_i, f)) \quad (5)$$

Usually, the neurophysiologist examines long-term EEG recordings, in which he extracts fragments with high-amplitude activity and, in his experience, classifies them into sleep spindles or epileptiform discharges. Fig. 2 shows examples of wavelet spectrograms of rat EEG signals the day after TBI with ED and SS and their ridges  $y(i)$ .

The entire 10-minute time interval has both interesting events for us and background activity. Therefore, to extract the points of ridges corresponding to SS or ED, it is necessary to delete the ridge points corresponding to the background.

Fig. 3 shows wavelet-spectrogram ridge of a 10-minute rat EEG signal.



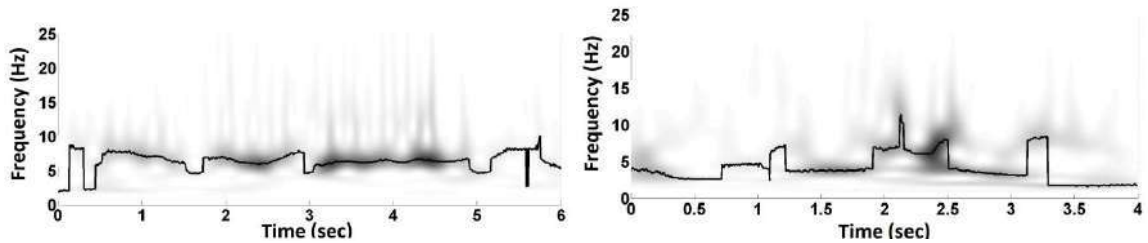


Fig. 2. Wavelet spectrograms and their ridges of rat EEG signal the day after the traumatic brain injury. To the left is ED and to the right is SS.

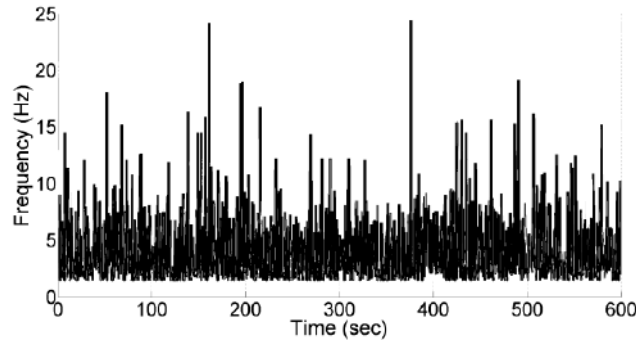


Fig. 3. Wavelet-spectrogram ridge of a 10-minute rat EEG signal.

SS and ED are characterized by an increased value of the spectral power density (PSD) as compared to the background. To select a positive ridge background clipping threshold  $Tr > 0$ , a histogram of the PSD at the points of the ridge is analyzed (Fig. 4). In the histogram, the PSD values are divided into 100 equal intervals.

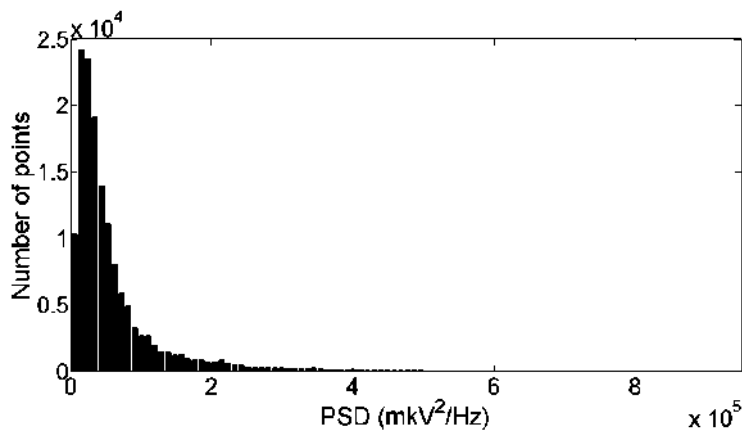


Fig. 4. One hundred interval PSD histogram at the ridge points of wavelet-spectrogram (10-minute record).

To calculate the histogram, the function "hist" was used. One of the output arguments of this function is an array of 100 PSD values. Each PSD value from this array was considered as a threshold  $Tr$ . The ridge points  $y(i) < Tr$  were assigned the value  $y(i) = 0$ . The remaining points of the ridge between the points  $y(i) = 0$  with the values  $y(i) \geq Tr$  are combined into a vector, which we will call an event. In Fig. 5 shows a histogram of the number of detected events, depending on the selected threshold value of PSD ( $Tr$ ).

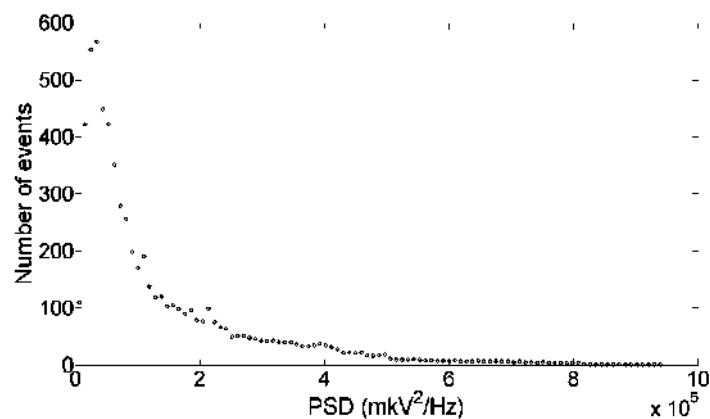


Fig. 5. Dependence of detected events number from PSD threshold value ( $Tr$ ).

We select a threshold value  $Tr$  to include all high-amplitude events present in the signal. Namely, the  $Tr$  value at which the maximum number of events was found (Fig. 5).

In the future, the beginning and the end of each event were calculated. The threshold value  $Tr$  is higher than the maximum value of the amplitude characteristic of the background activity. Consequently, the values of the beginning and end of found events do not correspond to the true ones. Therefore, the vector artificially expanded. Let the origin of the vector correspond to the point  $k$  of the ridge  $y(i)$ . We consider the points of the ridge  $k$  and  $k-1$ , if  $y(k-1) > y(k)$ , then the point  $k$  is considered the beginning of the event, otherwise the left shift along the ridge of the wavelet spectrogram continues until a local minimum is reached. A similar operation was done to calculate the end of the event, only the advance along the ridge was made to the right. Fig. 6 shows a 10-minute fragment of a filtered rat EEG signal with isolated high-amplitude events on it, and on wavelet-spectrogram ridge of this signal.

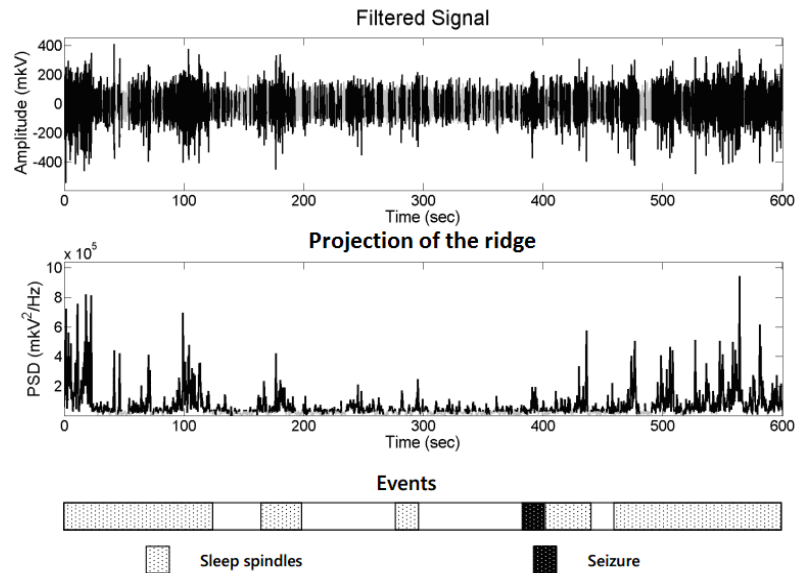


Fig. 6. Top image is 10 minute fragment of the filtered rat EEG signal with highlighted high-amplitude events (black lines). The middle figure is the ridge of the wavelet-spectrogram of this signal in the projection on the axis of the PSD-time with the selected events. The bottom figure is the events highlighted by the expert.

Fragments of a signal with epileptiform discharges are of immense importance in the study of PE. But they may not have such high amplitude, as, for example, in the time interval from 0 seconds to 200 seconds or from 440 seconds to 600 seconds (Fig.6.). Consider the minute section of the current 10-minute recording, at which the expert detected a discharge with smaller amplitude than the other events (Fig. 7). The presence of such events makes the detection process more difficult.

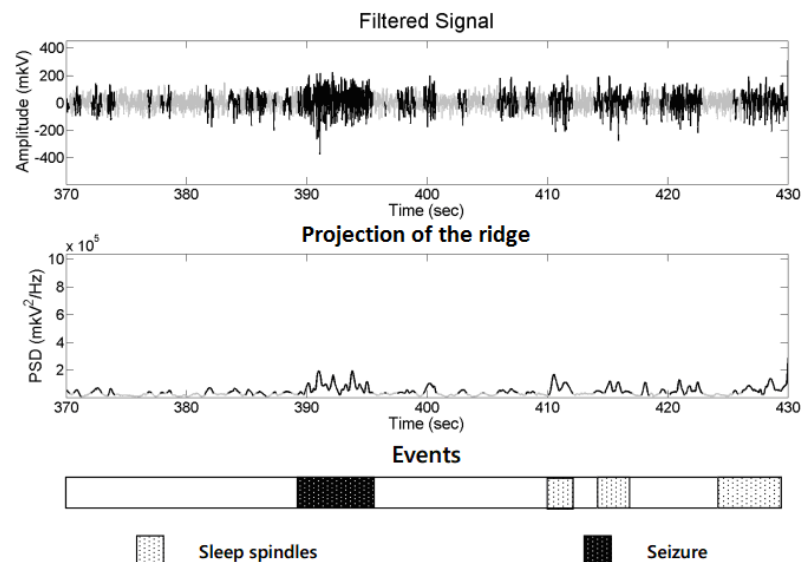


Fig. 7. Top image is minute fragment of the filtered rat EEG signal with highlighted high-amplitude events (black lines). The middle figure is the ridge of the wavelet-spectrogram of this signal in the projection on the axis of the PSD-time with the selected events. The bottom figure is the events highlighted by the expert.

The epileptiform discharge is in the time interval from 390 seconds to 395 seconds. As can be seen from Fig. 7, this method allows identifying regions with epileptiform discharges, but also other events are detected.

Additional conditions for the selection of high-amplitude events were given by expert-neurophysiologist. If there is a time delay between two events of not more than  $1/7$  second, then these two events are considered as one. Also, events longer than 0.5

seconds were considered uninformative and were removed from consideration. In Fig. 8 shows the result of reliable events, taking into account the conditions set by the expert.

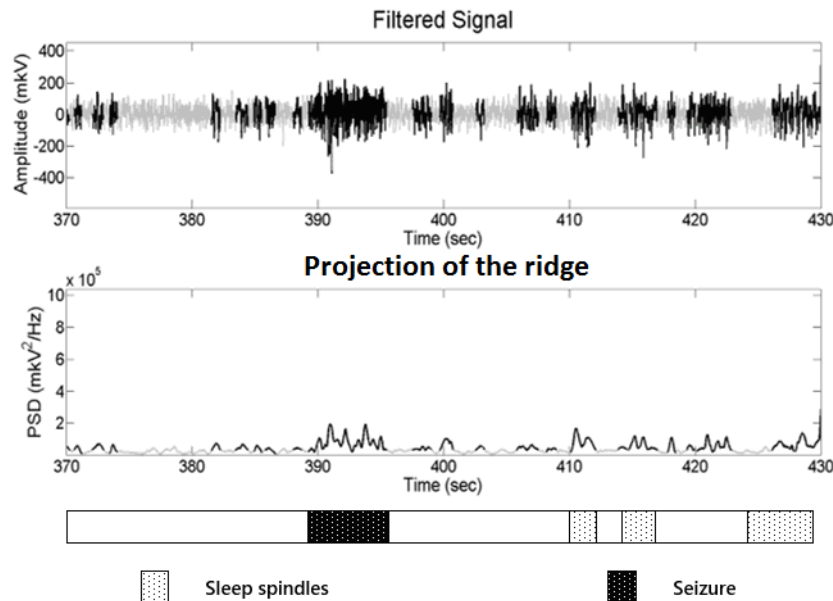


Fig. 8. Top image is minute fragment of the filtered rat EEG signal with highlighted high-amplitude events (black lines). The middle figure is the ridge of the wavelet-spectrogram of this signal in the projection on the axis of the PSD-time with the selected events. The bottom figure is the events highlighted by the expert.

The calculated areas are different from the background, but need further classification. They can contain both carotid spindles, epileptiform discharges, and other high-amplitude activity, which neurophysiologists have not detected. Beginning, end, duration, minimum, maximum and average value of frequencies, maximum PSD value were calculated for each event. After analyzing these parameters, events will be classified as epileptiform discharges, or as high-amplitude activity that are not ED.

### 3. Conclusion

The paper describes a method for automatic detection of high-amplitude rhythmic activity, based on the analysis of wavelet-spectrogram ridges. As a result of algorithm work, areas with high amplitude rhythmic activity on the electroencephalogram were allocated. Also, the parameters of the allocated parts of signal were calculated. With the help of this method, all epileptiform activity found by the expert, as well as sleep spindles and other high-amplitude activity. This method allows collecting a large group of events that will permit the classification of epileptiform discharges not only with background activity, but also with other events.

### Acknowledgements

This research was done at the expense of the grant of the Russian Science Foundation (project 16-11-10258).

### References

- [1] Annegers JF, Hauser WA, Coan SP, Rocca WA. A population-based study of seizures after traumatic brain injuries. *NEJM* 1998; 338: 20–24.
- [2] Pitkanen A, Immonen RJ, Grohn OHJ, Kharatishvili I. From traumatic brain injury to posttraumatic epilepsy: what animal models tell us about the process and treatment options. *Epilepsia* 2009; 50: 21–29.
- [3] Kabadi SV, Hilton GD, Stoica BA, Zapple DN, Faden AI. Fluid-percussion-induced traumatic brain injury model in rats. *Nature Protocols* 2010; 5(9): 1552–1563.
- [4] Hopfengartner R, Kasper BS, Graf W. Automatic seizure detection in long-term scalp EEG using an adaptive thresholding technique: a validation study for clinical routine. *Clinical Neurophysiology* 2014; 125(7): 1346–1352.
- [5] Divya S, Priyadharsini SS. Classification of EEG Signal for Epileptic Seizure Detection using EMD and ELM. *International journal for trends in engineering & technology* 2015; 3(2): 68–74.
- [6] Fergus P, Hignett D, Hussain A, Al-Jumeily D, Abdel-Aziz K. Automatic Epileptic Seizure Detection Using Scalp EEG and Advanced Artificial Intelligence Techniques. *BioMed research international*, 2015.
- [7] Sitnikova EY, Grubov VV, Khramov AE, Koronovskii AA. Developmental changes in the frequency-time structure of sleep spindles on the EEG in rats with a genetic predisposition to absence epilepsy (WAG/Rij). *Neuroscience and Behavioral Physiology* 2014; 44(3): 301–309.
- [8] Gloor P. Generalized cortico-reticular epilepsies: some considerations on the pathophysiology of generalized bilaterally synchronous spike and wave discharge. *Epilepsia* 1968; 9(3): 249–263.
- [9] Gloor P. Generalized epilepsy with bilateral synchronous spike and wave discharge. New findings concerning its physiological mechanisms. *Electroencephalography and Clinical Neurophysiology. Supplement* 1978; 34: 245–249.
- [10] Kostopoulos GK. Spike-and-wave discharges of absence seizures as a transformation of sleep spindles: the continuing development of a hypothesis. *Clinical Neurophysiology* 2000; 111: 27–38.
- [11] Jankel WR, Niedermeyer E. Sleep spindles. *Journal of clinical neurophysiology* 1985; 2(1): 1–36.
- [12] Jobert M. Topographical analysis of sleep spindle activity. *Neuropsychobiology* 1992; 26(4): 210–217.

- [13] Gabova AV, Bosnyakova DY, Bosnyakov MS, Shatskova AB, Kuznetsova GD. The Time–Frequency Structure of the Spike–Wave Discharges in Genetic Absence Epilepsy. *Doklady Biological Sciences*. Kluwer Academic Publishers-Plenum Publishers 2004; 396(1-6): 194–197.
- [14] Bosnyakova DY, Obukhov YuV. Extraction of dominant feature in biomedical signals. *Pattern Recognition and Image Analysis* 2005; 15(3): 513–515.
- [15] Bosnyakova D, Gabova A, Kuznetsova G. Time–frequency analysis of spike-wave discharges using a modified wavelet transform. *Journal of neuroscience methods* 2006; 154(1): 80–88.
- [16] Bosnyakova D, Gabova A, Zharikova A. Some peculiarities of time-frequency dynamics of spike-wave discharges in human and rat. *Clinical Neurophysiology* 2007; 118(8): 1736–1743.
- [17] Malla S. *Wavelets in Signal Processing*. Moscow: Mir, 2005; 671 p. (in Russian)

# The application of OpenCL to accelerate the lossless image compression algorithm based on cascading fragmentation and pixels sequence ordering

A. Khokhlachev<sup>1</sup>, V. Smirnov<sup>1</sup>, A. Korobeynikov<sup>1</sup>

<sup>1</sup>*Kalashnikov Izhevsk State Technical University, Studencheskaya 7, 426069, Izhevsk, Russia*

---

## Abstract

The previous papers of the authors offer approach to building the ordered sequence of image pixels at lossless compression, which comprises methods of cascading fragmentation and the use of bypasses code book. For fragment sized 6\*6 the code book contains 22144 various bypasses, the cost of coding to be estimated for every one of them. The search of optimal bypass is an exhaustive search type. The present paper describes ways of increasing the image lossless compression rate by using parallel computation based on OpenCL. Algorithm functions with great runtime were changed in order to transfer calculations to OpenCL using GPU/CPU. The acceleration degree for different algorithm functions gained in experiments amounted to 3..32.

*Keywords:* lossless image compression; cascading fragmentation; pixels sequence ordering; optimal bypass; code book; computational acceleration; parallel computing; open computing language (OpenCL); graphics processing unit (GPU); central processing unit (CPU); Haar integral-valued wavelet transformation; interchannel decorrelation

---

## 1. Introduction

At the moment there exist both a large number of compression algorithms of particular data classes and universal compression algorithms. This work will address the lossless image compression algorithm based on optimization of bypass image being developed by the authors and described in [1...3]. When processing test images [7], the algorithm gives the average compression ratio of 1.54, which matches the analogues [5]. Let us consider test results by groups of images: 1) in group «2.1.\*.tiff» by 1.426 2) in group «2.2.\*.tiff» by 1.547 3) in group «4.1.\*.tiff» by 1.622 4) in group «4.2.\*.tiff» by 1.522 [5]. In addition, the algorithm has some other advantages [5].

To achieve a high compression ratio it is necessary to use a number of demanding algorithm functions, which leads to longer image processing program runtime. Presently, parallel computing is there. The aim of this work is to apply OpenCL to speed up the lossless compression algorithm. To achieve this goal it is necessary to: analyze duration of program execution; find the algorithm functions with time-consuming calculations; consider transfer of these functions to GPU. Image processing performed in the algorithm is based on handling particular fragments, therefore, in general case, such tasks can be carried out simultaneously. Furthermore, it is possible to perform the preprocessing functions for image fragments in parallel as well.

## 2. Basic algorithm

The basic algorithm inherently consists in cascading fragmentation of image [1], the search of the fragment optimal bypass (path) [2], and dynamic programming of pixels delta-code at fragment bypass [6]. After encoding, the obtained data is further compressed by Deflate algorithm using standard libraries. The compression ratio depends on the class of the image being compressed, and on average equals 1.54 for the array of test images [5].

The runtime of image compression program depends on the processed image size. Due to a number of algorithmic solutions such as cascading fragmentation, and the use of bypasses codebook instead of calculating the possible bypasses for each image fragment, the runtime was reduced. However, the image compression duration is still high enough [5]: 1) in group «2.1.\*.tiff» - 101 seconds 2) in group «2.2.\*.tiff» - 404 seconds 3) in group «4.1.\*.tiff» - 24 seconds 4) in group «4.2.\*.tiff» - 141 seconds.

In computational terms the most complex of the basic algorithm functions is to estimate the encoding cost for all possible bypasses. Meanwhile, this algorithm function is suitable for parallelization, since the optimal bypass choice uses exhaustive search of obtained cost estimates. For a fragment sized 6\*6 the total bypasses number from the upper left corner is 22144.

To use all multi-core CPU resources it is necessary to effectively implement paralleling of functions between all cores. The basic program features parallel execution of optimal fragment bypass search cycle done with *.Net Framework* standard classes (SSE instructions). It is possible to use a more powerful CPU, but even in this case, the speed increase will not be significant.

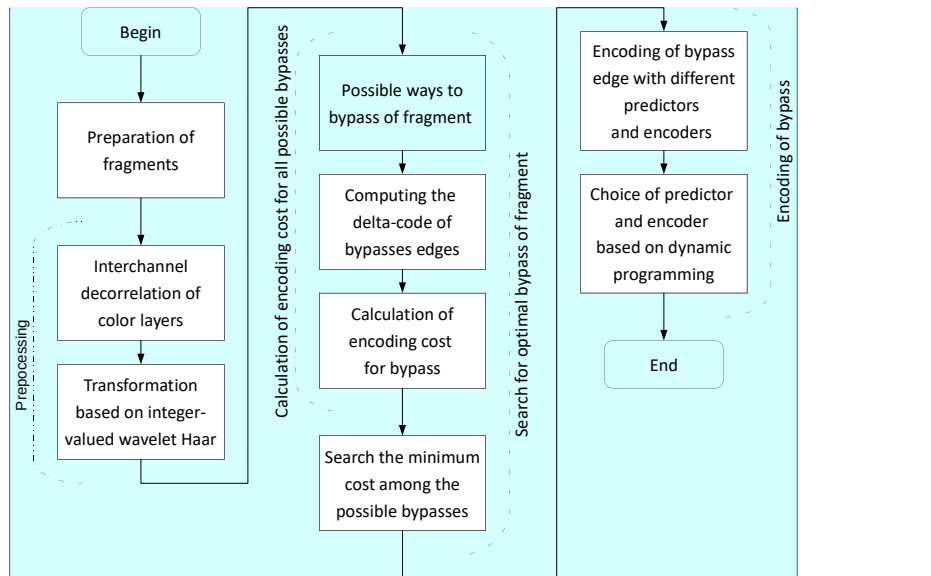
In recent years the increasing number of programs with parallel data processing use GPU computing [7]. This is dictated by a growing gap in performance between CPU and GPU.

Taking into account the above said, it was decided to move part of the compression algorithm functions to GPU. Obviously, this will require some significant changes in the functions, but it will allow for significant decrease in the program runtime without changing the basic algorithm.

Currently there are several approaches to programs execution on GPU. OpenCL is an open standard [8], which can execute programs on both CPU and GPU of different manufacturers. Therefore, in this research, to speed the algorithm, OpenCL was chosen.

At the moment there exist quite a big number of various compression algorithms in general and algorithms for images in particular. Images compressed both as lossy and lossless are widely and effectively used. For example, lossless compression is used in PNG files where the actual compression is implemented with Deflate algorithm [9, 10], which is a combination of LZ and Huffman algorithms in its turn. There are no free turn-key programs available for lossless image compression making use of OpenCL. WinZip is an example of the lossless compression program based on universal algorithm and using OpenCL, which provides for performance increase of about 45% [11].

In addition to the basic algorithm, image preprocessing was implemented which was described in the authors' previous works: interchannel decorrelation of image color layers [12] and the transformation of pixels matrix based on integer-valued Haar wavelets [13]. These functions can be easily threaded for the implementation on OpenCL.



**Примечание [K1]:** По рисунку: preprocessing (не хватает буквы), Haar integer-valued wavelet (порядок слов), to bypass fragments (не нужен предлог of), search of (предлог нужен)),

Fig. 1. Lossless image compression algorithm.

### 3. Methods of acceleration

The general problem solved in present research is changing the compression software in order to transfer part of calculations to OpenCL. Image compression algorithm is shown in Fig. 1.

#### 3.1. Preparation of fragments

The function receives separate color layers of an image. The function output is arrays of separate fragments of fixed size. Pixel values of the fragment nodes beyond the image borders are virtual pixels and the values of these pixels are set as constant (white pixels on Fig. 2). The top left pixels of each fragment on level 0 constitute the fragments on level 1 and so on, as long as the fragments number on a level is more than one. Data structure passed to the OpenCL kernel represents the matrix of image values, the output structure is the array of separate fragments [1].

#### 3.2. Preprocessing

##### 3.2.1. Interchannel decorrelation of color layers

This function is designed to calculate the interchannel decorrelation between the groups of color channels (layers) of the original image and to find the best variant to group them [12].

When function is started the arrays containing pixels values of all color channels of the fragment, and also the number of channels have to be conveyed (Fig. 3). In addition, data on the possible grouping of channels is needed.

Formula for calculating interchannel decorrelation for arbitrary channels number based on the mean and interchannel differences is applied [12]:

$$P^1 = \text{Round}\left(\frac{\sum_{i=2}^n X^i}{n}\right) \quad (1)$$

$$P^i = X^1 - X^i, i = \overline{2, n}$$

where  $Round$  is the rounding operation to the nearest integer;  $X^k$  – pixels value on each of the channels;  $k$  – channel index;  $n$  – the number of processed channels.

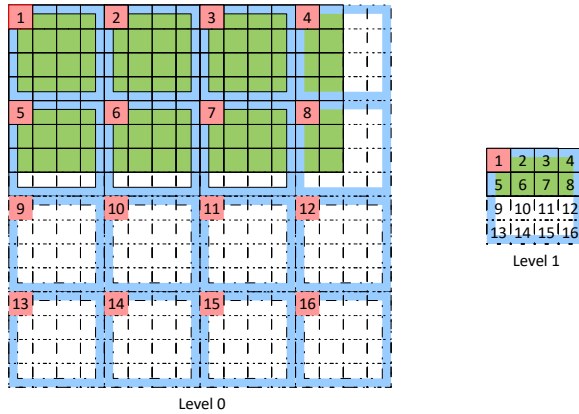


Fig. 2. Preparation of fragments.

The color channels can be independent from each other, therefore, the grouping variant with a minimum encoding costs estimate has to be selected. It is necessary to implement the decorrelation formulas for all dependent channels groups. The minimum channels number in the group is 2. If the image consists of 3 channels (24 bits per pixel), we get the following grouping variants:

$$(X^1 X^2 X^3) X^1 (X^2 X^3) X^2 (X^1 X^3) X^3 (X^1 X^2) X^1 X^2 X^3 \quad (2)$$

where decorrelation formulas are to be applied to groups of channels in parenthesis.

The calculation of decorrelation is performed for all possible groups ( $g=1..G$ ). The result is the index of  $g$  grouping:

$$g = \underset{g}{\operatorname{argmin}} \left( \sum_{g=1}^G \left( \sum_{i=1}^n \sum_{j=1}^k \operatorname{Cost}(P^{ij}) \right) \right) \quad (3)$$

where  $P_j^i$  – is the pixel value after the interchannel decorrelation for the grouping index  $g$ ;  $i$  – channel index;  $j$  – pixel index;  $n$  – number of channels;  $k$  – number of pixels number in the fragment.

$\operatorname{Cost}$  is the a some estimation-function of encoding costs estimate, for example, the length of the Fibonacci code which encoding the value  $P_j^i$  value, or the estimated length of binary coding:

$$\log_2 |P^{ij} + 1| + 1 \quad (4)$$

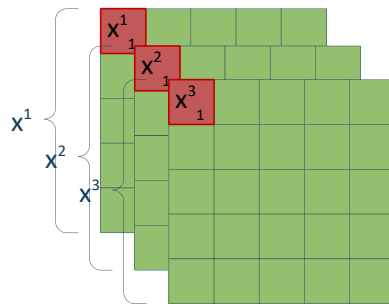


Fig. 3. Interchannel decorrelation of color layers.

### 3.2.2. Transformation based on Haar integer-valued wavelets-Haar

This function is designed-intended for cascading processing of each image fragment by applying an integer-valued version of the Haar wavelet transform [13].

The function is-passed-receives an array containing a single channel of image and the number of the processed fragments in by width and height-which-need-to-processed.

In the-course-of-the-algorithm-execution-it-uses-additional arrays for each executable OpenCL kernel with-athe-size-equal-to-of one fragment-each-are-used; for storing the-intermediate-interim results of the cascading conversion.

The function ~~result is outputs~~ an array of ~~the size equal to of the~~ original image.

Image matrix ~~should has to~~ be divided into blocks of sized  $2 \times 2$ . Then ~~calculated~~ the values for  $a, h, v, d$  ~~are to be found by the~~ formula [13]:

$$\begin{aligned} c_2 &= x_1 - x_2 \\ c_3 &= x_1 - x_3 \\ c_4 &= x_1 - x_4 \end{aligned} \tag{5}$$

$$c_1 = x_1 - \text{Round}\left(\frac{1}{4} \sum_{i=2}^4 c_i\right) = x_1 - z_1 \approx \frac{1}{4} \sum_{i=1}^4 x_i$$

$$\begin{aligned} a &= c_1 \\ h &= -\text{Round}(d/2) + c_3 \\ v &= -\text{Round}(d/2) + c_2 \\ d &= c_3 - c_4 + c_2 \end{aligned} \tag{6}$$

Where  $\text{Round}$  is the rounding operation to the nearest integer;  $x_i$  – original pixels of the block.

The obtained values of  $a, h, v, d$  ~~should have to~~ be recorded in positions of the matrix, as shown in Fig. 4. Calculation should be carried out as ~~multiresolution at multiple scale, by~~ repeating the transform on the blocks consisting of grouped values  $a^i$ , ~~each time and reducing their size in 2 times by half for in~~ each coordinate ~~every time, as long as it is possible to form  $2 \times 2$  block from  $a^i$  values on a subsequent scale.~~ Cascading transform will stop then the block  $a^i$  with size  $2 \times 2$  is absent.

It should be noted that ~~when with~~ the fragment sized of  $2^m \times 2^m$  it is possible to use preprocessing (interchannel decorrelation, and Haar transform) after the function of ~~dividing on the~~ fragmentations. In this case, the Haar transformation is possible only within the same fragment. ~~With In~~ this approach, the fragment encoding ~~is~~ completely independently of the other fragments and therefore the decoding ~~is~~ possible for a single fragment.

**Примечание [K2]:** Resolution не уверена насколько здесь обосновано применять его в значении масштаб???

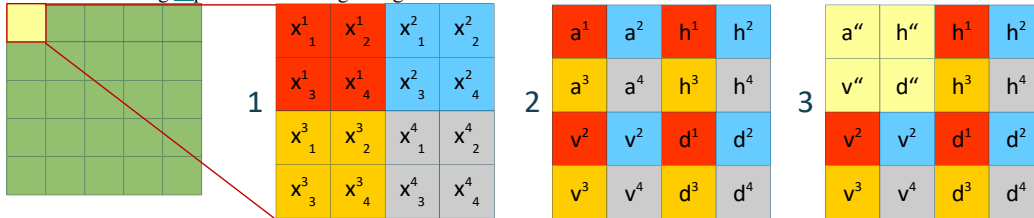


Fig. 4. Multiresolution-Multiple-scale transformation based on Haar wavelet-Haar.

### 3.3. Search for optimal bypass of fragment

The function receives the fragments obtained after preprocessing functions (integer-valued Haar transform, interchannel decorrelation of color layers).

#### 3.3.1. Calculation of encoding cost for all possible bypasses of fragment

##### 3.3.1.1. Possible ways to bypass of a fragment

Encoding and decoding algorithms have information about all the possible bypasses for a given fragment size, is known for encoding and decoding algorithms.

All bypasses (paths) have been calculated in advance and constitute stored in the bypasses codebook [2].

Therefore, in encoding and decoding only need to know the bypass index, but not the edges of bypass (path), is the only prerequisite for encoding and decoding, but not the edges of bypass (path).

##### 3.3.1.2. Computing the delta-code of bypasses edges

This function is designed to calculate the difference of values for all pairs of nodes that make up the edge on a given fragment bypass [2]. In the course of the function The algorithm uses the previously prepared fragments.

The result is a list of arrays containing the delta-code of all edges for each fragment.

It is necessary for each fragment you need to make compile an array of differences between the nodes values (delta-code) connecting the edge  $e$  is calculated according to the which is done with formula:

$$\Delta_e = x_{start(e)} - x_{stop(e)} \tag{7}$$

Where  $x_{start(e)}, x_{stop(e)}$  are the pixel values are connected by an edge  $e$ ;  $start(e)$  and  $stop(e)$  are nodes indexes of edge  $e$ .

##### 3.3.1.3. Calculation of encoding costs for bypass

For each fragment, estimates the encoding costs for each of the possible bypasses are calculated. The function receives the previously prepared fragments and details data one of all the fragment bypasses in from the codebook.



The result is the fragment-specific array containing estimated encoding costs for each bypass of the fragment (Fig. 5, Table 1).

Estimation of the encoding costs of a bypass through all edges (with its all delta-codes) of bypass edges it is possible to produce can be done in different ways. The cost of bypass for each fragment is needed to find the cost of the bypass:

$$\Sigma_s = \sum_{e=1}^E Cost(\Delta_e) \cdot z_s^e \tag{8}$$

where  $E$  is the length of bypass (the number of edges);  $e$  - edge index;  $S$  - the number of bypasses;  $s$  - bypass index;  $\Delta_e$  - delta-code of edge;  $z_s^e$  - presence of the edge  $e$  in bypass  $s$ ;  $Cost$  - is the some estimation function of encoding costs, it is similar to the  $Cost$  in interchannel decorrelation function.

It should be noted that the estimate of bypasses encoding costs based on the table 1 is effective from the point of view of parallel computing. In this function there is relies on parallel processing of all possible bypasses downloaded from the codebook, for all fragments making up, forming the processed image.

### 3.3.2. Search of the minimum among the possible bypasses

After the estimates of encoding costs of all paths is selected are estimated, and the save path of each fragment bypass with the smallest estimate for each fragment is picked and saved.

$$s = \underset{s}{\operatorname{argmin}}(\Sigma_s) \tag{9}$$

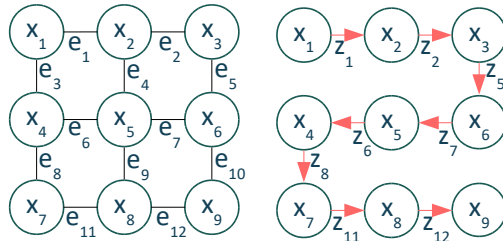


Fig. 5. Example of bypass of a 3\*3 fragment.

Table 1. Calculation of encoding cost for bypass in 3\*3 fragment.

$e$	$\Delta_e$	$z_1^e$	$z_2^e$	...	$z_8^e$
1	$\Delta_1$	1	1	...	0
2	$\Delta_2$	1	1	...	1
3	$\Delta_3$	0	0	...	1
4	$\Delta_4$	0	0	...	1
5	$\Delta_5$	1	1	...	1
6	$\Delta_6$	1	1	...	0
7	$\Delta_7$	1	0	...	0
8	$\Delta_8$	1	1	...	1
9	$\Delta_9$	0	1	...	1
10	$\Delta_{10}$	0	1	...	1
11	$\Delta_{11}$	1	0	...	1
12	$\Delta_{12}$	1	1	...	0

### 3.4. Encoding of bypass

This function is designed to encode the previously found optimal bypass. That is, the obtained array of bypass nodes values with a minimum encoding costs, must be handled by the processed with predictor and encoded by the encoder.

It should be noted that the encoding of bypass which previously was found as optimal bypass can be performed in various ways. In particular, there can be used certain known generic methods can be used: Huffman algorithm or arithmetic coding.

As for the considered suggested algorithm it is offered to perform the algorithm the bypass encoding of bypass is suggested which employs using a more sophisticated method [6]: 1) using a set of predictors and encoders for to encode the bypass edge; 2) using dynamic programming for choice of choose predictors and encoders for edge in purpose to optimize (to minimize) of the total bypass encoding costs of bypass.

1.1.1.3.4.1. Encoding of bypass edge with different predictors and encoders

In the simplest-most common case, the edge delta-code described above, and Fibonacci codes are used as encoders. In this case, the application of dynamic programming to select the predictor and the encoder is not required. In a more complex cases the number of choices of predictors and encoders variants may be more than one. For example, it is possible case with the predictors are possible not only on the basis of not only the finite difference of the first degree, but higher degrees, and Rice codes with different bases can be used as encoders with the encoders with use Rice codes with different bases. The use of a set of predictors and a set of encoders increases the resulting image compression ratio, but this raises the problem of choosing the best predictor and encoder for the current section of the bypass array.

**Отформатировано:**  
 многоуровневый + Уровень: 3 +  
 Стиль нумерации: 1, 2, 3, ... + Начать  
 с: 1 + Выравнивание: слева +  
 Выровнять по: 0 см + Отступ: 0 см

1.1.2.3.4.2. Choice of predictor and encoder based on dynamic programming

In this embodiment, this variation of compression algorithm to encode with which each bypass edge is encoded uses the most optimal encoding parameters (predictor/encoder) based on defined by dynamic programming [6]. In start of the When function is started, it is loaded the table of encoding cost for every encoder for values of every predictors for all pixels on edges of the optimal bypass is downloaded and executed.

**Отформатировано:**  
 многоуровневый + Уровень: 3 +  
 Стиль нумерации: 1, 2, 3, ... + Начать  
 с: 1 + Выравнивание: слева +  
 Выровнять по: 0 см + Отступ: 0 см

In the result the function creates produces a data file containing information with on the size of the encoded using encoders predictors values for edges and additional information – optimal switching of the predictors/encoders for edges of bypass encoding [6].

Due to the complexity of the dynamic programming algorithm it was found possible to transfer to run on OpenCL managed to transfer only a small part, responsible for the coding directly to OpenCL. This part contains branching, and is switching large sections of the algorithm takes place. These operations are an integral part of the algorithm, or changing the calculations flow in aim of parallel execution without the use of branches is not possible.

**Отформатировано:**  
 многоуровневый + Уровень: 1 +  
 Стиль нумерации: 1, 2, 3, ... + Начать  
 с: 1 + Выравнивание: слева +  
 Выровнять по: 0 см + Отступ: 0 см

2.4. Results and Discussion

Screen form the interface of the developed software is shown in Fig. 6. The program displays the following information: used hardware processor device processing unit and software platform being used; the number of files to be processed; the current processed file; the execution duration-time of the compression program particular functions; the execution duration of overall compression time; the compression ratio.

To use Hardware requirements for OpenCL acceleration requires the presence of a GPU or CPU with support of OpenCL 1.2 [8]. You must install the appropriate OpenCL support software which distributed with equipment is to be installed.

To The compilation of the developed image compression program requires the following software components [8]: 1) DotNetZip library, for the final compression of results using the with Deflate algorithm; 2) to provide OpenCL bindings for C#, use the Cloo library from OpenTK to link OpenCL to C#; 3) to compile and execute kernels on the GPU, Ionic.Zip.dll library is used (To compress the encoding results used library Ionic.Zip.dll. In addition is used Other requirements include a set of libraries to support work OpenCL running, bindings these libraries to .Net Framework and the source codes of the OpenCL kernels compiled in course of program execution.

The basic program required about 250 MB of RAM. When algorithm was adapted the algorithm for accelerating on OpenCL was added using large-volume arrays were added of large-volume and hence memory required demand increased to 900 MB.

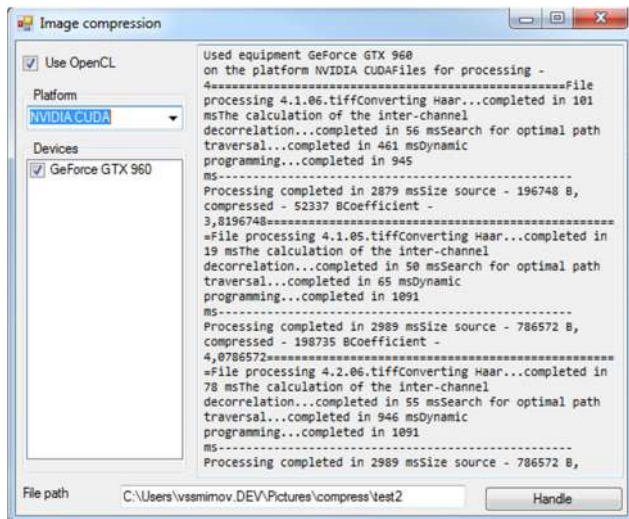


Fig. 6. ~~Screen form~~Interface of the developed software.

Test ~~batch-sample~~ of images ~~designed-is intended~~ to assess acceleration of all core functions of modified program ~~in comparing with relative to basic one~~. To estimate the dependence of the program ~~speed-runtime on the to-images size, the batch have the-images of different sizes are sampled. To check-used~~For test purposes 4 images from the standard set provided by the Institute of signal ~~processing~~ and images ~~processing were used~~: 4.1.06.tiff, 4.2.05.tiff, 4.2.06.tiff, 4.2.07.tiff. The color depth of the images ~~are-is~~ 24 bit. Image sizes ~~are~~: 256\*256, 512\*512, 1024\*1024, 2048\*2048.

At ~~t~~Testing was ~~performed-using~~ image compression was performed with ~~bothas-the-the~~ basic program on the CPU AMD Phenom II X4 955 ~~platform~~ and a program using OpenCL. For testing OpenCL parallel processing was ~~used-different 4 different devices were used~~: 1) GPU AMD Radeon HD6850; 2) GPU Nvidia GTX 960; 3) CPU AMD Phenom II X4 955; 4) CPU AMD FX-4300. The time spent on ~~the-particular functions of the algorithm, and the total processing time for each image are given in Table 2 and Fig. 7, where F1is-~~ integer-valued Haar transform, F2 ~~-~~ interchannel decorrelation of color layers, F3 - search of the optimal bypass, F4 - encoding bypass using dynamic programming.

When testing for ~~e~~Each fragment ~~hadwas~~ fixed size: 6\*6 pixels, and the number of bypasses: 22144 ~~at testing~~.

It should be noted, that ~~when-using-the-with-GPU Nvidia GTX 960 configuration-, according to Profiler, the load does not exceed 60% while-despite numerous s-the-high-number-of-~~ processing devices and high work frequency. ~~-according to Profiler the load does not exceed 6~~ Compression-The image of size 2048\*2048 pixels ~~size could not be compressed~~ on the GPU AMD Radeon HD6850 ~~failed-to-produce~~ due to ~~the-lack-of-insufficient~~ graphics memory. In ~~the-future, to avoid this-situationsuch failures, the necessary-modification-of-the-program needs to be modified: to-run-the-performed~~ calculations flow should be divided into several ~~groups-threads~~ and processed sequentially.

In the basic program ~~were-not implementedthe~~ functions of the integer-valued Haar transform and ~~the-interchannel decorrelation were not implemented~~, and therefore, testing of these functions was ~~not~~ carried out.

Testing ~~showed-yielded~~ approximately the same reduction in ~~the-compression-total-overall compression time when-usingwith~~ both CPU and GPU ~~application~~. The larger the size of the processed image, the greater the acceleration obtained as long as there is memory available ~~to-for~~ OpenCL.

GPU ~~showed-the-best results-performed better in the-for~~ searching of the optimal bypass ~~task~~. CPU ~~well-with-the-function-of handles~~ dynamic programming ~~well, due to because-of-presences of~~ a large number of branches in the function, despite the small number of ~~processorof processor~~ cores.

~~Time spent on c~~Calculating ~~ons-of~~ interchannel decorrelation and integer-valued Haar transform ~~is-performed-using~~ OpenCL ~~for-a-short-timeis insignificant~~ compared to total compression time.

Table 2. Results of processing of test images.

Used program / device	Image size, pixels	Execution time of compression particular functions, milliseconds					Acceleration, Times			
		F1	F2	F3	F4	Total	F3	F4	Total	
Program on OpenCL	AMD Radeon HD 6850	256*256	71	65	677	528	2581	2,2	9,0	2,8
		512*512	43	63	913	1129	4643	6,1	13,4	5,2
		1024*1024	66	187	2017	4492	15217	12,9	13,5	6,4
		2048*2048	151	438	—	—	—	—	—	—
	Nvidia GeForce GTX 960	256*256	23	20	160	360	2057	9,4	13,2	3,5
		512*512	11	36	411	1153	4402	13,6	13,1	5,5
		1024*1024	33	148	1474	4226	13416	17,7	14,4	7,3
		2048*2048	122	675	5354	14530	50031	31,8	15,9	8,6
	AMD FX-4300	256*256	34	30	350	499	2581	4,3	9,5	2,8
		512*512	18	48	913	1146	4207	6,1	13,2	5,7
		1024*1024	45	206	3324	4093	15514	7,9	14,8	6,3
		2048*2048	172	920	13164	13629	58567	12,9	17,0	7,4
AMD Phenom II X4 955	256*256	31	32	455	664	2407	3,3	7,2	3,0	
	512*512	21	56	1 378	1222	5981	4,0	12,4	4,0	
	1024*1024	67	239	4571	4477	16777	5,7	13,6	5,8	
	2048*2048	227	992	18432	12804	61350	9,2	18,1	7,0	
Basic program / AMD Phenom II X4 955	256*256	—	—	1504	4751	7273	1,0	1,0	1,0	
	512*512	—	—	5570	15155	24183	1,0	1,0	1,0	
	1024*1024	—	—	26107	60698	97297	1,0	1,0	1,0	
	2048*2048	—	—	170161	231398	431555	1,0	1,0	1,0	

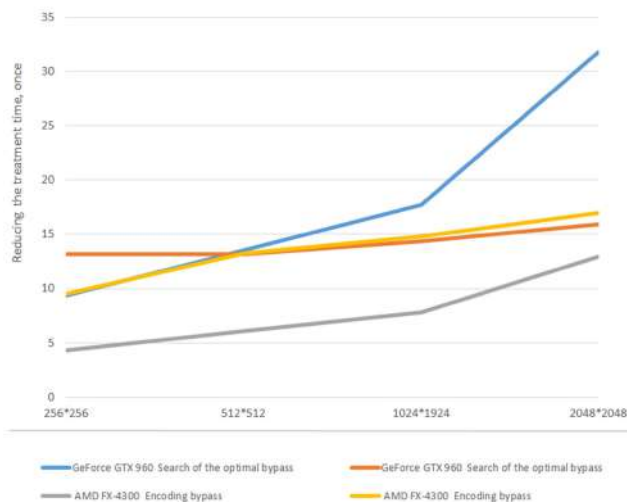


Fig. 7. Comparison of image compression acceleration.

### 3.5. Conclusion

In the course of this work was modified the basic program for lossless image compression without losses was modified, with the aim of increasing shortening its runtime speed. The parallel processing based on OpenCL was used for program acceleration. This solution significantly affected the processing speed, enabling making it possible to reduce computational time. This modification will allow provide for more efficient use of the program in the future, will facilitate future further research aimed at improving the compression ratio.

The changing of optimal bypass search function allowed for to obtain the acceleration up to 32-fold acceleration on the large images. This acceleration has been achieved because of executing OpenCL functions executed on OpenCL are almost linear, and branching, even where they are when it is the case, is limited to have only a few simple operations. Furthermore, for future program modification the acceleration of this function is important for future program modification because it is makes possible to use fragments of larger size that which was previously impossible unattainable earlier due to too much great execution time. Among other things Moreover, fragments with the sized of  $2^k * 2^k$  will effectively allow applying the integer-valued Haar transformation for the fragment to them, and will allow to compressing every each fragment separately.

Somewhat worse is the situation with As regards dynamic programming, the prospects are not as bright during encoding fragment bypass. Speed Performance managed to increase was gained mostly by due to ordinary conventional parallel execution of some operations, shutdown of cancel of operations which need only used solely for debugging purposes, and the use of the packet data read operations. The part that runs on OpenCL gives the increase in performance is of only about 30% compared with to ordinary parallel computing. On the other hand, even this result is relatively good enough, given the fact provided that OpenCL function has rather a wide large enough branching. It should be nNoted that the bypass encoding can be performed in various ways, for example, e.g. with Huffman algorithm or arithmetic coding.

### References

- [1] Smirnov VS, Korobeynikov AV. Cascade Image Splitting into Fragments at Lossless Compression on Basis of Image Bypass Optimization. Bulletin of Kalashnikov ISTU 2012; 2: 143–144.
- [2] Korobeynikov AV, Smirnov VS. Optimal Bypass Definition with Code Book Application at Images Lossless Compression. Bulletin of Kalashnikov ISTU 2012; 3: 114–115.
- [3] Smirnov VS, Korobeynikov AV. Ordering the numeric sequence of image pixels at lossless compression. I International Forum “Instrumentation Engineering, Electronics and Telecommunications (November, 25–27, 2015, Izhevsk, Russian Federation), 2015; 175–180.
- [4] Sample images from the site of University of Southern California. URL: <http://sipi.usc.edu/database/database.php?volume=misc> (2017-01-10).
- [5] Smirnov VS, Korobeynikov AV. The results of testing lossless compression algorithm based on cascade fragmentation method and ordering pixels sequence. II International Forum “Instrumentation Engineering, Electronics and Telecommunications (November, 23–25, 2016, Izhevsk, Russian Federation), 2016.
- [6] Korobeynikov AV. The Use of Dynamic Programming and Fibonacci Codes for Interchannel Decorrelation. The Three-Channel Signals Lossless Compression. Bulletin of KIGIT 2010; 1: 72–81. URL: <http://elibrary.ru/item.asp?id=18348092> (2017-01-10).
- [7] Denny Atkin. Computer Shopper: The Right GPU for You. URL: <http://www.computershopper.com/feature/the-right-gpu-for-you> (2017-01-10).
- [8] Official webpage of the standart OpenCL. URL: <https://www.khronos.org/opencl/> (2017-01-10).
- [9] Portable Network Graphics (PNG) Specification (Second Edition). URL: <https://www.w3.org/TR/PNG/> (2017-01-10).
- [10] PNG Home Site. URL: <http://www.libpng.org/pub/png/> (2017-01-10).
- [11] WinZip official webpage. URL: <http://www.winzip.com/win/ru/index.htm> (2017-01-10).
- [12] Franchenko RS, Korobeynikov AV. Interchannel Decorrelation for Any Number of Channels at Lossless Compression of Multichannel Signals. Bulletin of Kalashnikov ISTU 2010; 1: 87–88.

Отформатировано:  
 многоуровневый + Уровень: 1 +  
 Стиль нумерации: 1, 2, 3, ... + Начать  
 с: 1 + Выравнивание: слева +  
 Выровнять по: 0 см + Отступ: 0 см

- [13] Smirnov VS, Korobeynikov AV. Lossless Image Compression Based On Integral-Valued Haar Wavelets. *Intelligent Systems in Manufacturing. Bulletin of Kalashnikov ISTU* 2013; 2: 158–160.

# Methods of IPD normalization to counteract IP timing covert channels

K. Kogos<sup>1</sup>, A. Sokolov<sup>1</sup>

<sup>1</sup>National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), 31 Kashirskoe Sh., 115409, Moscow, Russia

## Abstract

Covert channels are used for information transmission in a manner that is not intended for communication and is difficult to detect. We propose a technique to prevent the information leakage via IP covert timing channels by inter-packet delays normalization in the process of packets sending. Recommendations for using the counteraction methods and choosing parameters were given. The advantage of our approach is that the covert channel is eliminated or limited preliminary, whereas state of the art methods focus on detecting active IP covert channels that may be insecure.

*Keywords:* Covert channel; IP timing channel; elimination; limitation; traffic normalization; inter-packet delays; capacity

## 1. Introduction

Covert channels were introduced by Lampson as channels not intended for information transfer at all [1]. TCSEC defines covert channel as any communication channel that can be exploited by a process to transfer information in a manner that violates the system's security policy [2].

Covert channels were classified into storage and timing channels. Storage channels involve the direct or indirect writing of a storage location by the sender and the direct or indirect reading of it by the receiver. Timing channels include the sender signaling information by modulating the use of resources over time such that the receiver can observe it and decode the information.

Information in covert timing channel can be encoded by varying packets transfer rates (or inter-packet times) [3, 4, 5, 6] and by packet sorting [7]. Storage channels in networks can be encoded in packet lengths [8, 9] or packet header fields (TTL, TOS, ID, Checksum, etc.) [10, 11, 12, 13]. Network covert channels are described on Fig. 1.

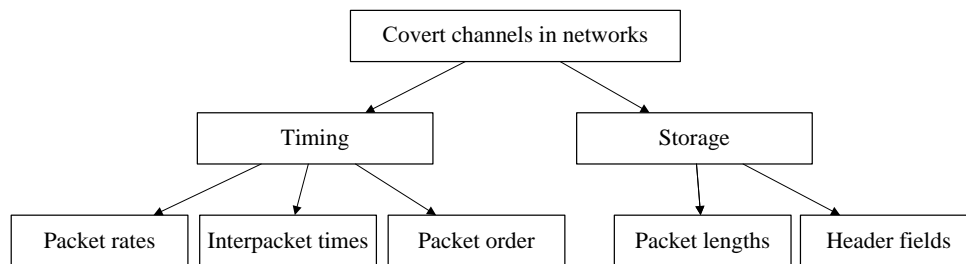


Fig. 1. Types of network covert channels.

Fig. 2 illustrates the main stages of covert channels counteraction.

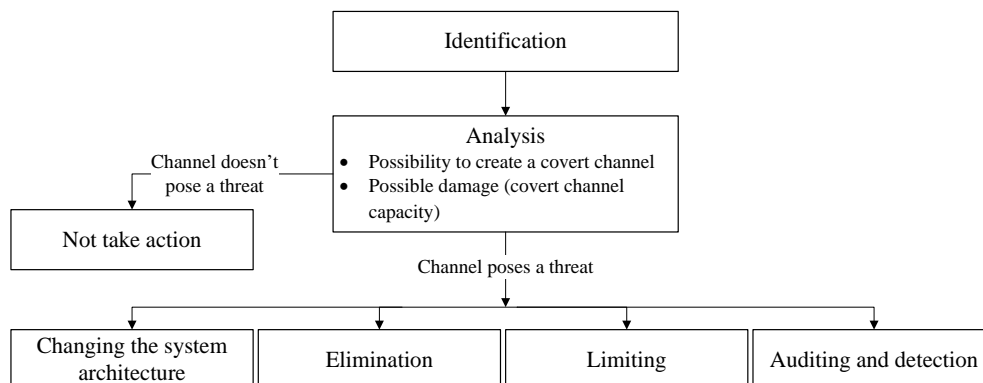


Fig. 2. Covert channels handling.

The identification problem is to find the potential covert channels that can be realized in the analyzed system. The second step is the analysis of identified channels to assess the threat level of each covert channel. If channel poses a threat to the protected system the following measures can be applied: elimination, limiting, detection. Ideally covert channels should be identified and removed during the design phase. Covert channels in networks can be eliminated by traffic encryption and normalization (protocol headers, packet lengths, inter-packet times). If a channel cannot be eliminated its capacity should be

reduced by using limiting techniques [14, 15, 16, 17, 18]. Auditing and detection methods can be used to detect the operating covert channels [4, 19, 20, 21, 22]. These methods are based on the detection of non-standard or abnormal behavior. Covert timing channels in networks can be eliminated only by normalizing inter-packet times. But this measure reduces the communication channel bandwidth. Method parameters must be correctly selected to minimize the negative impact on network performance.

The rest of the paper is organized as follows. First, we give an overview of existing methods of covert channels construction and counteraction in Chapter 2. In Chapter 3, we introduce recommendations on the choice of parameters of covert channel elimination method. In Chapter 4, we consider two ways to limit covert channels capacity. In Chapter 5, we provide experimental results to demonstrate counteraction methods effect on network performance. Our conclusions are presented in Chapter 6.

## 2. Related Work

### 2.1. Methods of covert timing channels construction

Covert information can be encoded by varying packet rates or inter-packet times. The covert sender varies packet rate between two (binary channel) or more packet rates each time interval. The receiver decodes the covert information by measuring the rate in each time interval. The sender and receiver need a mechanism for synchronization of the time intervals. Timing channel where the sender either transmits or stays silent in each time interval (on/off channel) is a special case of binary channel [3]. Authors of [5] implemented the on/off timing channel. In their scheme the covert data is divided into frames and synchronization between sender and receiver is achieved through a special start sequence at the beginning of each frame. There are variants of the timing channels that does not require synchronization between sender and receiver because the covert information is encoded directly in the inter-packet times of transmitted packets [23, 24].

Authors of [25] developed an indirect covert channel that exploits the fact that a host's CPU temperature is proportional to the number of packets per time unit it processes and a host's system clock skew depends on the temperature. The channel requires an intermediary that receives and sends packets to both covert sender and receiver. The covert sender either sends packets to the intermediary or stays silent. The covert receiver estimates the intermediary's clock skew by observing a series of timestamps in packets sent by the intermediary. There are other implementations of thermal covert channels [26].

Covert timing channel can be organized through packet sorting [7]. Sender can transmit a maximum of  $\log_2 n!$  bits because a set of  $n$  packets can be arranged in any  $n!$  ways. This approach requires per packet sequence numbers to determine the original packet order. The method only modifies the sequence numbers, thus keeping payload unchanged.

### 2.2. Methods of covert channels counteraction

Admissible covert channel capacity depends on the kind of protected information and on the amount of leaked information, which is critical. TCSEC assumes that covert channels with maximum bandwidths of less than 1 bit per second are acceptable in most application environments [2]. According to IBM guidelines, channels with bandwidths lower than 0.1 bits per second can exist. There is no special need to counteract them. Channels in range from 0.1 to 100 bits per second can exist when absolutely necessary [27].

Capacity of covert timing channels in networks can be limited by adding random delays to the packets. Fig. 3 shows the framework of using traffic control module [18]. Network covert timing channel exploitation takes place here. An innocent process request the OS kernel to send a network packet, then covert message sender can somehow interfere with this procedure (for example, delay response), after that the remote covert channel receiver eavesdrops related packets and decodes the message. No matter whether there are covert channels, the traffic controller will get in on the network packet send out procedure.

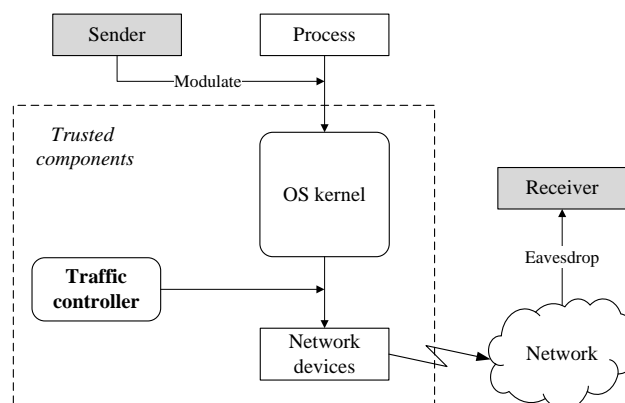


Fig. 3. Framework of using traffic controller.

For each network connection, traffic control module maintains some information (network address, port number, connection type, previous packet's outgoing time, etc.). When an application sends out a packet, traffic controller will intercept the packet, look up the network connection information and add a random delay to the packet (fixed delays could be easy filtered by covert channel users). Delay of  $n$ th packet will be calculated according to the formula:

$$T_n = f(\Delta t_n, k) = \text{Rand}(k) \cdot \Delta t_n, \quad (1)$$

where  $\Delta t_n$  denotes the time interval between current and previous packet-sending request,  $k$  is a configurable parameter ( $0 < k < 1$ ),  $\text{Rand}(k)$  function generates a random number ranged from 0 to  $k$ . Hence,  $T_n$  will be a random value from 0 and up to  $k \cdot \Delta t_n$ . Experimental results shows that the covert communication achieved nearly 100% encode/decode correctness when traffic control was disabled. With the traffic control enabled, the error rate rapidly raised to about 50%.

Gateway is often used to prevent the data transmission from higher security level to lower. Gateway is located between the sender with low security level and receiver with high security level (Fig. 4). When the gateway receives a packet from low it stores it into a buffer and sends an acknowledgment (ACK) to low. Then it transmits the packet to high and waits for an ACK. If the ACK is received the gateway removes the packet from the buffer.

However, when the buffer is full the gateway must wait for high to acknowledge a received packet until another packet from low can be stored in the buffer. The time of sending an ACK to low is directly related to the time of receiving an ACK from high. High can ensure the buffer is always filled and vary the rate of its ACKs. In this manner, he can exploit the covert channel. The PUMP model reduces this covert channel capacity [16, 17].

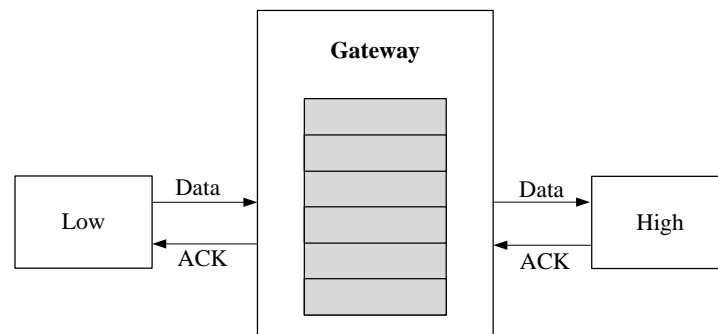


Fig. 4. Message passing from low to high using the gateway.

Network covert timing channels can be eliminated only by normalizing inter-packet times. But this measure reduces the communication channel bandwidth. Method parameters must be correctly selected to minimize the negative impact on network performance.

### 3. Covert timing channels elimination

Inter-packet times normalization makes it necessary to delay the transmission of packets and generate dummy packets. It reduces the network performance. So, method of covert channels elimination should be used only if the leakage of a very small amount information is unacceptable. Parameters of inter-packet times normalization method must be correctly selected to minimize the negative impact on communication channel capacity.

Input data for the calculation of the best inter-packet time value  $kt$  can include:

1. empirical distribution of inter-packet time over a long period of time,
2. maximum acceptable packet queue delay  $lt$ ,
3.  $\varepsilon$  equal to the allowable part of packets which may be delayed for a time greater than  $lt$ .

Following conditions must be met when inter-packet times normalization to  $kt$  is performed:

1. communication channel bandwidth is not less than the set value,
2. percentage of packets which are delayed for a time greater than  $lt$  is not more than  $\varepsilon$ ,
3. number of dummy packets is minimal.

One of the following values can be used instead of  $\varepsilon$  and  $lt$ :

1. maximum allowable average packet delay  $d_{avg,t}$ ,
2. maximum acceptable part of dummy packets.

Suppose we have a distribution of inter-packet time (Fig. 5). The minimum value of the inter-packet interval is equal to  $t$  and maximum equal to  $mt$ .



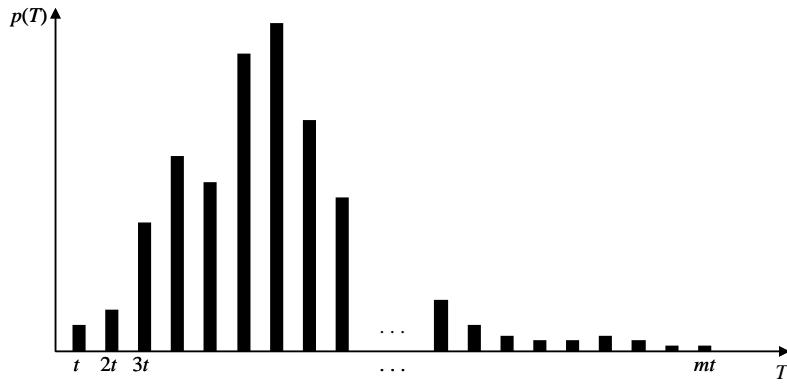


Fig. 5. Inter-packet time distribution.

Let the inter-packet times are normalized to  $kt$ . The device processes packets for an infinitely small time and its queue is empty at the moment. When two packets with  $at$  interval arrive to the device, there will be the following. The second packet will be delayed for  $kt - at$ , if  $at \leq kt$ . The second packet will be delayed for  $\left(\left\lceil \frac{at}{kt} \right\rceil - 1\right)kt + (kt - at) = \left\lceil \frac{at}{kt} \right\rceil kt - at$  and  $\left\lceil \frac{at}{kt} \right\rceil - 1$  dummy packets will be generated and sent by the device, if  $at > kt$ .

So, when  $n+1$  packets with  $a_1t, a_2t, \dots, a_nt$  intervals come to the device, delay of the  $(i+1)$ th packet is equal to

$$d_it = d_{i-1}t + (N_{d_i} + 1)kt - a_it, \tag{2}$$

where  $d_0t = 0$  and  $N_{d_i}$  is a number of dummy packets sent after receiving the  $i$ th packet (during the  $a_it$ ):

$$N_{d_i} = \left\lceil \frac{a_it - d_{i-1}t}{kt} \right\rceil - 1. \tag{3}$$

The smaller the inter-packet time  $kt$ , the less packet queue delay and the greater the number of dummy packets.

Let the inter-packet time is a discrete random variable  $\xi$  obeying the distribution law in Table 1.

Table 1. Distribution law of  $\xi$ .

$\xi$	$t$	$2t$	...	$(m-1)t$	$mt$
$P(\xi = it)$	$p_1$	$p_2$	...	$p_{m-1}$	$p_m$

Queue delay of  $(n+1)$ th packet is given by:

$$dt = (n + N_d)kt - \sum_{i=1}^n a_it, \tag{4}$$

where  $N_d$  is a number of dummy packets sent during the  $\sum_{i=1}^n a_it$  after receiving the first packet.

The inter-packet time after normalization should not exceed the average value in the initial distribution to avoid the constant increase in queue length. That is, the following inequality must be satisfied:

$$kt \leq E(\xi), \tag{5}$$

where  $E(\xi)$  is the expected value of a variable  $\xi$ .

One should choose a value of  $kt$  for which this probability is not greater than  $\varepsilon$ . Furthermore, the value of probability should be as close to  $\varepsilon$  as possible to minimize the amount of dummy packets. In choosing the value of  $kt$  based on the maximum acceptable part of dummy packets ( $\frac{N_d}{n + N_d}$ ) one should select the minimum suitable  $kt$  value to minimize packet delays.

We consider two use cases of communication channel:

1. file transfer only,
2. real-time data transfer (e.g. VoIP, Skype).

The maximum packet queue delay is not too important, if the channel is not being used for real-time data transmission. Allowable average packet delay or acceptable percentage of dummy packets should be used as input data in this case. If you use a channel for real-time data transfer, it is essential to ensure good communication quality. Therefore, the inter-packet time  $kt$  should be calculated based on the maximum acceptable packet delay. For example, packet jitter should not exceed 30 milliseconds to provide an acceptable quality of a Skype call [28, 29].

#### 4. Covert timing channels limitation

If a non-zero covert channel capacity is allowed one can use partial inter-packet times normalization. Such methods have less negative impact on the communication channel.

##### 4.1. Normalization by several inter-packet times

Let two values of inter-packet times after traffic normalization be allowed:  $k_1t$  and  $k_2t$ . Inter-packet time equal to  $k_1t$  will be observed at the output if the queue is not empty in  $k_1t$  after sending the last packet. If the queue is empty at this moment the inter-packet time equal to  $k_2t$  will appear at the output. It will be a dummy packet if the queue also is empty in  $k_2t$  after sending the last packet.

Violator can affect the inter-packet times by passing packets to the channel and use covert channel. Let the random variable  $X$  take the values 0 or 1 in accordance with the inter-packet times ( $k_1t$  or  $k_2t$ ) at the output.  $p$  is the probability of observing packets with an interval equal to  $k_1t$ . Then entropy of  $X$  is equal to:

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p). \quad (6)$$

The average time between two consecutive outgoing packets is  $pk_1t + (1-p)k_2t$ . So, capacity of covert timing channel that can be built is:

$$C = \frac{-p \log_2 p - (1-p) \log_2 (1-p)}{pk_1t + (1-p)k_2t}, \quad (7)$$

where  $(1-p)^{k_1t} = p^{k_2t}$  holds.

It is possible to use more than two values of inter-packet delays.

##### 4.2. Normalization by several packet transfer rates

Let there are two inter-packet delays:  $k_1t$  and  $k_2t$  ( $k_1t < k_2t$ ) which correspond to two fixed packet transfer rates. During each interval  $T$  inter-packet times are fixed and equal to  $k_1t$  or  $k_2t$  (packets are transmitted at a constant rate). Rate can change or remain the same at the time points  $T \cdot i$  ( $i = 1, 2, \dots$ ). Selected transfer rate depends on the number of packets received at the last part of the  $T$  and the number of packets in the queue. Lower rate (which correspond to  $k_2t$ ) will be set at the time  $T \cdot j$  if

$$\frac{T'}{N_{T'} + N_q} > k_2t, \quad (8)$$

where  $N_{T'}$  is the number of packets received during the time interval  $(T \cdot j - T', T \cdot j)$ ;  $N_q$  is the amount of packets in the queue at the moment. Otherwise, a high data transfer rate will be established.

When using this method, covert channel capacity is:

$$C = \frac{\log_2 r}{T}, \quad (9)$$

where  $r$  is the number of transfer rates. In this case  $r = 2$ .

Parameter  $T$  should be chosen depending on the predetermined allowable capacity of covert channel  $C_a$ .

$$T = \frac{\log_2 r}{C_a}. \quad (10)$$

## 5. Experimental results

This chapter provides experimental results to demonstrate the effect of inter-packet times normalization on network performance. Two use cases of network are reviewed: file transfer only and real-time data transfer. For each of these cases we have two empirical distribution of inter-packet time (under high and low network load). The best values of inter-packet time was calculated for several input data sets.

### 5.1. Covert channels elimination during file transfer

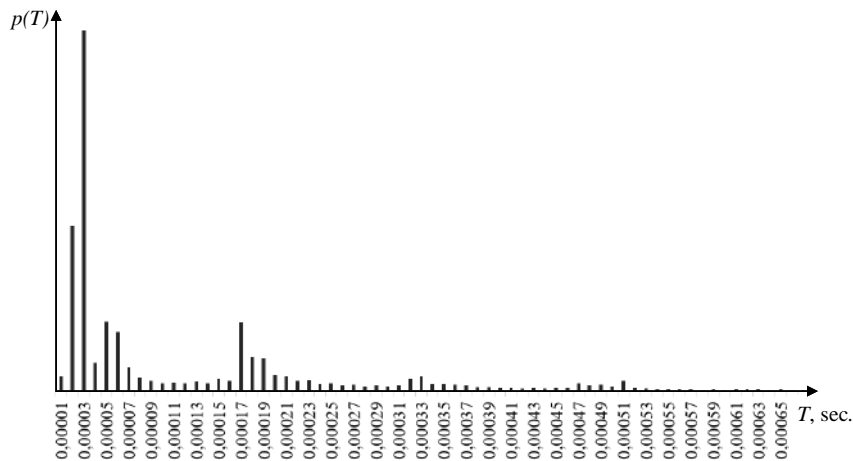


Fig. 6. Inter-packet time distribution under high network load ( $E(T) = 0.00017$  sec.).

Table 2. Results of calculation of  $kt$  based on the acceptable part of dummy packets (high network load).

$\frac{N_d}{n + N_d}$	$kt$ , sec.	$d_{avg}$ , sec.
<b>0.1</b>	0.00016	0.00611
<b>0.3</b>	0.00012	0.00053
<b>0.5</b>	0.00009	0.00019

Table 3. Results of calculation of  $kt$  based on the acceptable average packet delay (high network load).

$d_{avg}$ , sec.	$kt$ , sec.	$\frac{N_d}{n + N_d}$
<b>0.5</b>	0.00017	0.002
<b>1.0</b>	0.00017	0.002

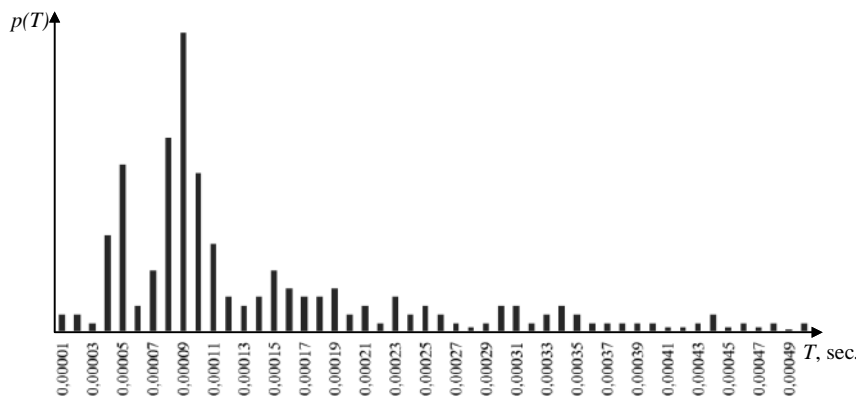


Fig. 7. Inter-packet time distribution under low network load ( $E(T) = 2.33749$  sec.).

Table 4. Results of calculation of  $kt$  based on the acceptable part of dummy packets (low network load).

Table 5. Results of calculation of  $kt$  based on the acceptable average packet delay (low network load).

$\frac{N_d}{n + N_d}$	$kt$ , sec.	$d_{avg}t$ , sec.
<b>0.1</b>	2.09752	47.1353
<b>0.3</b>	1.64243	12.9241
<b>0.5</b>	1.16614	5.45441

$d_{avg}t$ , sec.	$kt$ , sec.	$\frac{N_d}{n + N_d}$
<b>0.5</b>	0.25739	0.89
<b>1.0</b>	0.41155	0.82
<b>5.0</b>	1.11260	0.52

5.2. Covert channels elimination during real-time data transfer

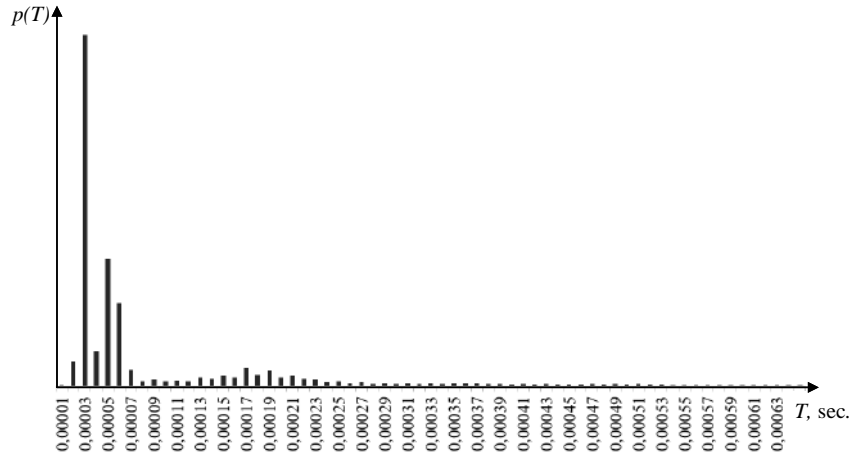


Fig. 8. Inter-packet time distribution under high network load ( $E(T) = 0.00025$  sec.).

Table 6. Results of calculation of  $kt$  based on the maximum acceptable packet delay (high network load;  $\epsilon = 0.001$ ).

$lt$ , sec.	$kt$ , sec.	$d_{avg}t$ , sec.	$\frac{N_d}{n + N_d}$
<b>0.005</b>	0.00012	0.00063	0.52
<b>0.010</b>	0.00014	0.00117	0.44
<b>0.020</b>	0.00016	0.00216	0.36

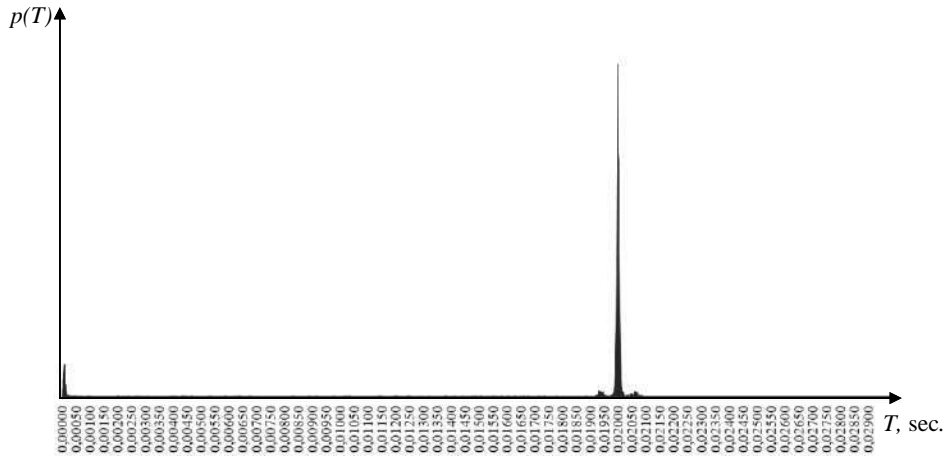


Fig. 9. Inter-packet time distribution under low network load ( $E(T) = 0.02472$  sec.).

Table 7. Results of calculation of  $kt$  based on the maximum acceptable packet delay (low network load;  $\epsilon = 0.001$ ).

$lt$ , sec.	$kt$ , sec.	$d_{avg}t$ , sec.	$\frac{N_d}{n + N_d}$
<b>0.005</b>	0.00192	0.00108	0.92
<b>0.010</b>	0.00367	0.00209	0.85
<b>0.020</b>	0.00720	0.00421	0.71

### 5.3. Covert channels limitation

The following dependencies were identified using a covert channel limit method that allows two packet transfer rates (Fig. 10, 11).

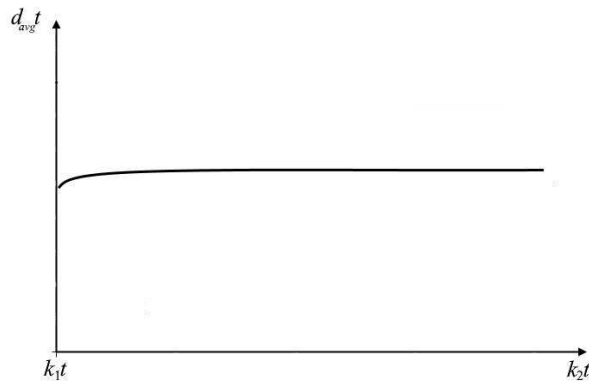


Fig. 10. The dependence of average packet delay on  $k_2t$  for a fixed  $k_1t$  (normalization by 2 packet transfer rates).

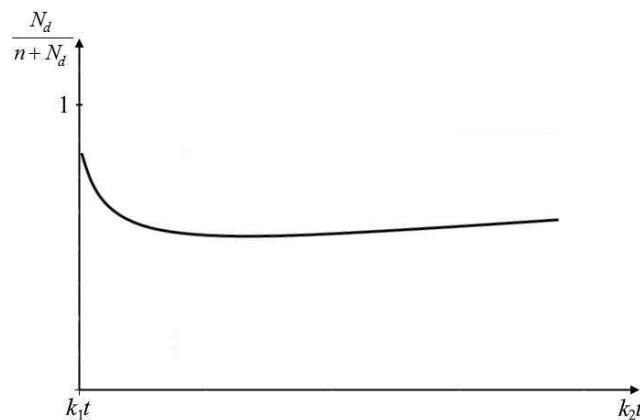


Fig. 11. The dependence of the part of dummy packets on  $k_2t$  for a fixed  $k_1t$  (normalization by 2 packet transfer rates).

### 5.4. Comparison of counteraction methods

Three techniques of inter-packet delays normalization were applied under the same conditions. The requirement for the average packet delay was specified. The parameters of each method have been calculated to ensure a minimum effect on the communication channel performance. The values of the covert channel capacity and part of dummy packets are shown in the Table 8.

Table 8. Comparison of covert channels counteraction methods.

Average packet delay, sec.	Normalization method	Part of dummy packets	Covert channel capacity, bit/sec.
0.1	One inter-packet time	0.911	0
	Two inter-packet times	0.340	664
	Two packet transfer rates	0.602	1
0.5	One inter-packet time	0.898	0
	Two inter-packet times	0.100	13
	Two packet transfer rates	0.546	1

## 6. Conclusions

Inter-packet times normalization makes it necessary to delay the transmission of packets and generate dummy packets. Parameters of inter-packet times normalization method must be correctly selected to minimize the negative impact on communication channel capacity. Channel performance requirements may be different. They depend on how you use the channel. The results also show that the packet delays and the number of dummy packets are strongly depend on the communication channel load.

Complete normalization of inter-packet delays is only way to eliminate covert timing channels. However, this measure greatly reduces the communication channel capacity and should be used only if the leakage of a very small amount information is unacceptable. In other case, one can use methods of partial inter-packet times normalization which limit covert channel capacity.

## Acknowledgements

This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Professional Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

## References

- [1] Lampson BW. A note on the confinement problem. *Communications of the ACM*, 1973: 613–615.
- [2] Department of defense standard. Department of defense trusted computer system evaluation criteria, 1985; 116 p.
- [3] Padlipsky MA, Snow DW, Karger PA. Limitations of end-to-end encryption in secure computer networks: technical report. Bedford: MITRE Corporation, 1978; 11 p.
- [4] Brodley C, Cabuk S, Shields C. IP covert timing channels: design and detection. *Proc. CCS 2004*: 178–187.
- [5] Cabuk S. Network covert channels: design, analysis, detection, and elimination : for the degree of doctor of philosophy. Indiana: Perdue University, 2006; 111 p.
- [6] Hovhannisyian H, Lu K, Wang J. A novel high-speed IP-timing covert channel: Design and evaluation. *Proc. 2015 IEEE International Conference 2015*: 7198-7203.
- [7] Ahsan K, Kundur D. Practical data hiding in TCP/IP. *Proc. ACM Wksp. Multimedia Security*, 2002; 8 p.
- [8] Yao Q, Zhang P. Coverting channel based on packet length. *Computer Engineering* 2008; 34.
- [9] Zhang L, Liu G, Dai Y. Network Packet Length Covert Channel Based on Empirical Distribution Function. *Journal of networks* 2014; 9(6): 1440–1446.
- [10] Kundur D, Ahsan K. Practical Internet Steganography: Data Hiding in IP. *Proc. Texas Wksp. Security of Information Systems*, 2003.
- [11] Lucena NB, Lewandowski G, Chapin SJ. Covert Channels in IPv6. *Proc. Privacy Enhancing Technologies* 2005: 147–166.
- [12] Alsaffar H, Johnson D. Covert channel using the IP timestamp option of an IPv4 packet. *Proc. The International Conference on Electrical and Bio-medical Engineering* 2015: 48–51.
- [13] Mavani M, Ragha L. Covert channel in IPv6 destination option extension header. *Proc. 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications*, 2014.
- [14] Hu WM. Reducing timing channels with fuzzy time. *Journal of Computer Security* 1992; 1(3-4): 362–372.
- [15] Trostle J. T. Modelling a fuzzy time system. *Journal of Computer Security* 1993; 2(4): 291–310.
- [16] Kang MH, Moskowitz IS. A pump for rapid, reliable, secure communication. *Proc. First ACM Conference on computer and communications security* 1993: 119–129.
- [17] Kang MH, Lee DC, Moskowitz IS. A network version of the pump. *Proc. 1995 IEEE Computer society symposium on research in security and privacy* 1995; 144–154.
- [18] Wang Y, Chen P, Ge Y., Mao B, Xie L. Traffic controller: a practical approach to block network covert timing channel. *Proc. International Conference on Availability, Reliability and Security* 2009: 349–354.
- [19] Mahajan AN, Shaikh IR. Detecting Covert Channels in TCP/IP Header with the Use of Naive Bayes Classifier. *International Journal of Computer Science and Mobile Computing* 2015; 4(6): 1008–1012.
- [20] Rezaei F, Hempel M, Shrestha PL, Rakshit SM, Sharif H. Detecting covert timing channels using non-parametric statistical approaches. *Proc. IEEE International Wireless Communications and Mobile Computing Conference* 2015; 102–107.
- [21] Venkataramani G, Chen J, Doroslovacki M. Detecting Hardware Covert Timing Channels. *Journal IEEE Micro* 2016; 36(5): 17–27.
- [22] Rezaei F. A Novel Approach towards Real-Time Covert Timing Channel Detection : for the degree of doctor of philosophy. Linclon: The University of Nebraska, 2015; 136 p.
- [23] Berk V, Giani A, Cybenko G. Detection of covert channel encoding in network packet delays : technical report. New Hampshire: Thayer school of engineering of Dartmouth college, 2005; 11 p.
- [24] Sellke SH, Wang C-C, Bagchi S. TCP/IP Timing Channels: Theory to Implementation. Indiana: Purdue University, 2009; 9 p.
- [25] Murdoch SJ. Hot or not: revealing hidden services by their clock skew. *Proc. 13th ACM conference on computer and communications security* 2006; 27–36.
- [26] Masti RJ, Rai D, Ranganathan A, Muller C, Thiele L, Capkun S. Thermal Covert Channels on Multi-core Platforms. *Proc. 24th USENIX Security Symposium* 2015; 865–880.
- [27] IBM Knowledge Center. URL: <http://www-01.ibm.com/support/knowledgecenter> (05.01.2017).
- [28] Inside Skype for Business. URL: <http://blog.insidelync.com/2012/06/a-primer-on-lync-audio-quality-metrics/> (05.01.2017).
- [29] Alreja A. Understanding quality of experience alerting. Redmond: Microsoft Corporation, 2011; 15 pp.

# Automatic adjustment of image processing pipeline

D.A. Kolchaev<sup>1</sup>, E.R. Muratov<sup>1</sup>, M.B. Nikiforov<sup>1</sup>

<sup>1</sup>Ryazan State Radio Engineering University, Gagarina, 59/1, 390005, Ryazan, Russia

---

## Abstract

Mathematical processing of images in real-time vision systems can be conventionally divided into two stages: preprocessing (filtering, contrasting, protection from natural distortions, etc.) and final one (imposition, visualization, solution of the navigation task, etc.). Mentioned tasks can be solved by a lot of known and specially developed methods with various degrees of efficiency. The present paper suggests a mathematical criterion model and algorithm of automatic selection of the most effective method at each stage of the image processing pipeline in relation to the current situation at its input.

*Keywords:* image processing pipeline; real time; automatic algorithm selection

---

## 1. Introduction

Important stage in vision systems is preprocessing which includes contrasting of images and compensation of interferences. Since contemporary algorithms of enhancement provide different results which depend on both processed image and control parameters transmitted to these algorithms then there is a necessity of dynamical selection of the algorithm depending on the plot. Combination of all variants of algorithms and also all variants for control parameters provides a set of solutions with high dimensionality. Solution of this issue is usage of the automatic system for selection of enhancement algorithms and also selection of optimal control parameters for the series of  $t$  frames obtained from the video sequence [1].

## 2. Task definition

Let a set of contrasting algorithms be  $A$  and a set  $B$  be variations of the method for interference compensation then  $A_i$  is  $i$ -algorithm of contrasting and  $B_j$  is  $j$ -variation of the interference compensation method. Hence, direct multiplication of a set  $A$  by a set  $B$  provides a set  $M$  which describes all possible variants of the mutual application of these algorithms.  $M = A \times B$ .  $(A_i, B_j)$  is one of selection variants,  $(A_i, B_j) \in M$ . Initial task is the following: to find such element  $(A_i, B_j)$  from the set  $M$  when quality index of the resulting image is the best and noise index is located in the range corresponding to algorithm  $B_j$ . For current task we have  $i, j = 0..3$  because three algorithms of contrasting and three variants of noise time filtering usage have been selected experimentally.

## 3. Algorithm for automatic selection of combination of methods for image enhancement and interference compensation

Algorithm allows finding the best variant of processing by selecting a certain combination providing the best result from the point of view of some objective index from a set of interference compensation methods with various control parameters and from several algorithms of image contrasting. Processing results are estimated after every  $t$  frames after accumulation of the sequence of four reference frames where the best processing algorithms correspond to each frame. Frequency of these algorithm repetitions is estimated and algorithms which repetition frequency is the highest are selected. The algorithm allows selecting the best element from the set  $M$  for each frame and recording this element to the stack. When the stack is filled and the most effective processing algorithm is selected, it is applied to the current image. After following  $n$  frames the element which was first recorded for the stack is removed and the best element from the set  $M$  occupies the stack top. Procedure of selection of application of the algorithm is repeated. Stack is organized according to the principle FILO (First-In-Last-Out).

Main blocks of the algorithm are a block of quality assessment and block of noise evaluation.

Integral performance index (IPI) is calculated as an amount of average brightness, brightness root-mean-square deviation, normalized contrast index, number of informative levels and entropy. Number of tonal gradations characterizes a number of various informative levels being present on the image and it is determined by the image histogram [2]:

$$G(Z_i) = \begin{cases} 0, & \text{if } Z_i = 0, \\ 1, & \text{if } Z_i > 0, \end{cases}$$

where  $Z_i$  – a number of point which brightness is equal to  $i$ .

One of the most significant characteristics of the image is average brightness  $\bar{L}$  [3] calculated according to the following formula:

$$\bar{L} = \frac{\sum_{y=1}^N \sum_{x=1}^W L_{xy}}{HW},$$

where  $H$ ,  $W$  – a height and width of the image, and  $L_{xy}$  – brightness of the element of the current image with coordinates  $x$  and  $y$ .

Such objective characteristics as root-mean-square deviation ( $\sigma$ ) and entropy ( $\varepsilon$ ) are used for quantitate quality assessment. Root-mean-square deviation is equal to notions - local contrast and accuracy to some extent. Entropy is a measure of quantity of information in the image.

Task to estimate image quality has a multicriterion nature, so an additive generalized criterion  $F$  is introduced as follows:

$$F = \sum_{i=1}^p \beta_i f_i,$$

where  $\beta_i$  – weight coefficients,  $\sum_{i=1}^p \beta_i = 1$  – a condition of normalization  $F$ ,  $f_i$  – partial normalized criteria,  $p$  – a number of partial criteria.

Normalized partial indices of the contrast and numbers of informative levels are determined as follows:

$$K_n = \frac{(L_{\max} - L_{\min})}{255},$$

$$N_n = \frac{N}{N_{\max}},$$

where  $L_{\max} = \max(L_{xy})$ ,  $L_{\min} = \min(L_{xy})$  – maximum and minimum values of brightness of image elements,  $N$  – a number of informative levels being different from null,  $N_{\max} = 256$  – a maximum number of informative levels in digital images for visualization.

Shannon entropy estimation can be calculated for any half-tone (including television and thermal) image. In this case estimation of distribution of possibilities of gray shades in the half-tone image is calculated [3]. Calculation of the entropy is performed based on the image histogram which distribution of frequencies is described by a simple expression:

$$p_i = \frac{N_i}{HW}, \quad (1)$$

where  $N_i$  – a number of elements having  $i$  level.

Calculation of the image entropy is performed according to formula:

$$\varepsilon = -\sum_i p_i \log_2(p_i). \quad (2)$$

For normalization entropy values can be divided into a coefficient being an entropy maximum for such number of levels. For the half-tone image with 256 brightness gradations it is equal to 8. So, the image entropy value can vary from 0 to 1.

Image dispersion is calculated as following:

$$\sigma^2 = \sum_i p_i (i - \bar{i})^2, \quad (3)$$

where  $\bar{i} = \sum_i i p_i$ ,  $i$  – a level of quantization.

For the half-tone image estimation of dispersion of distribution of possibilities of gray shades is calculated. Experimentally it was determined on a series of different images that root-mean-square deviation varies within the range from  $\approx 0$  to  $\approx 100$ , then the mean value is 50, consequently  $\sigma_n$ :

$$\sigma_n = \begin{cases} \frac{\sigma}{50}, & \sigma \leq 50, \\ \frac{(100 - \sigma)}{50}, & 50 < \sigma \leq 100, \\ 0, & \sigma > 100. \end{cases}$$

For average brightness, values belonging to middle of the range are preferable, and on boundaries of the brightness range its

$$\text{value is minimum: } \bar{L}_n = \begin{cases} \frac{\bar{L}}{128}, & \bar{L} \leq 107, \\ \frac{(255 - \bar{L})}{128}, & \bar{L} > 147, \\ 1, & \bar{L} \in \{108 \div 147\}. \end{cases}$$

Entropy achieves its maximum under the uniform distribution law. Image entropy having the range from 0 to 255 brightness gradations cannot exceed 8. Normalized value of the entropy has the form:  $\varepsilon_n = \frac{\varepsilon}{8}$ .



Main complication of particular index application is selection of weight coefficients taking into consideration influence of corresponding particular indices on the generalized criterion as a whole. Fishburne criterion is used for selection of initial values of these coefficients [3, 4]:

$$\beta_i = \frac{2(p-i+1)}{p(p+1)}.$$

For this purpose partial criteria are divided into groups by priorities:  $(L_n, \sigma_n)$ ;  $(K_n, N_n)$  and  $(\varepsilon_n)$ . Root-mean-square deviation contribution weightiness to the value of the quality function integral index is described by this index meaning: it determines accuracy and to some extent perceives intense noise. Then indices are arranged by descending of influence inside separated priority groups.

Taking into account above-mentioned facts, integral performance index (IPI) of the image brightness component has the form:

$$\text{IPI} = 0,33\bar{L}_n + 0,27\sigma_n + 0,20K_n + 0,13N_n + 0,07\varepsilon_n.$$

Correction of coefficients under partial indices is performed by the method of expert evaluations, besides their amount should be equal to one.

Noise power for the whole image is calculated in the block of noise evaluation. For this purpose, image is divided into windows of size 3x3 and value of the neighbor pixel brightness which is located diagonally to the central one is subtracted from the central pixel brightness. Result of subtraction is raised to the square and summed up by all pixels of the image.

Described algorithm allows automatically selecting a method of contrasting and also a mode of interference compensation. However, usage of this algorithm on a video sequence leads to appearance of areas where a sharp jump in brightness occurs (other algorithm of contrasting is chosen), such event negatively influences on perception of video information by an operator. The method described below is suggested to be used for solution of this issue.

#### 4. Interpolation method of proportional application of two boundary algorithms

Let's suppose that  $k$  is a number of the video frame where replacement of the algorithm occurs, then  $k + t = k'$  is a number of the following frame where analysis and selection of the enhancement algorithm are performed, then  $k_T$  is a current video frame, besides  $k \leq k_T \leq k'$ . Let's designate  $A_k$  as an enhancement algorithm on the  $k$ -frame and  $A_{k'}$  as an enhancement algorithm after  $t$  frames after  $k$ -frame. Interpolation method is the following: for each  $k_T$  frame, proportion of two algorithms  $A_k$  and  $A_{k'}$  is calculated, so, the closer  $k_T$  is to  $k'$ , the higher coefficient the result of algorithm  $A_{k'}$  is used with, and consequently, the lower coefficient the result of algorithm  $A_k$  is used with. Reversed situation can occur similarly when  $k_T$  is closer to boundary  $k$ . Hence, formula for calculation of the resulting image pixel brightness depending on value  $k_T$  has the following form:

$$I_{x,y} = I_{x,y}^{A_k(k_T)} * \left(1 - \frac{k_T - k}{t}\right) + I_{x,y}^{A_{k'}(k_T)} * \left(\frac{k_T - k}{t}\right), \quad (4)$$

where  $x,y$  – pixel coordinates,  $I_{x,y}^{A_k(k_m)}$  – a buffer with a result of the first algorithm for the current frame,  $I_{x,y}^{A_{k'}(k_m)}$  – a buffer with a result of the second algorithm for the same frame,  $A_k(k_T)$ ,  $A_{k'}(k_T)$  – results of operation of algorithms selected on  $k$  and  $k'$  frames correspondingly,  $t$  – a number of frames between moments when automatic choice of the algorithm happens.

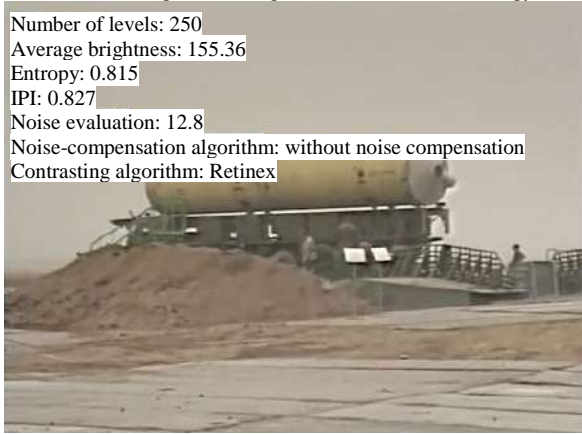
Such formula allows gradually changing applied algorithms without sharp bursts on the resulting image but requires processing of the image by two algorithms that decreases resulting performance. The algorithm based on this method begins operation with obtaining of results of the automatic selection for  $k$  and  $k'$  frames. Then each frame is processed by two algorithms  $A_k$  and  $A_{k'}$ , processing results are stored in two buffers of images. Each pixel of the resulting image is formed according to formula 4. After achievement of the interval boundary ( $k_T = k'$ ),  $A_k = A_{k'}$ , and value  $A_{k'}$  is obtained from automatic selection of the following enhancement algorithm.

#### 5. Results of the algorithm for automatic selection of the enhancement method

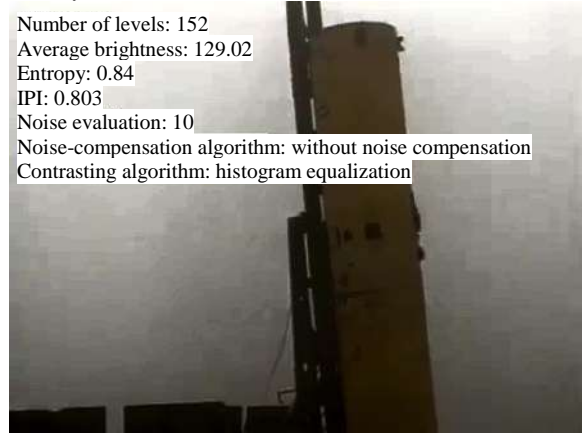
Fig.1 shows four frames from a video fragment and current variant of enhancement provided by the algorithm of automatic selection with interval  $t = 50$  frames. In this case result of the automatic selection for the  $k+t$  frame is different from other frames of this sequence. Such choice leads to the fact that if brightness sharply changes on one frame of the video fragment (e.g. light flash has occurred) then enhancement of following  $n$  frames is performed with the ineffectively selected algorithm. For solution of this issue stack is used. The stack contains four best elements of the set  $M$  which were obtained as a result of preliminary operation of the algorithm. When stack is filled completely, selection of the enhancement algorithm is performed by calculating frequencies of repetitions  $A_i$  and  $B_j$  in this stack. For sequence in Fig.1 the following repetition frequencies are obtained:

$$F(A_2) = 1; F(A_3) = 3; F(B_0) = 4.$$

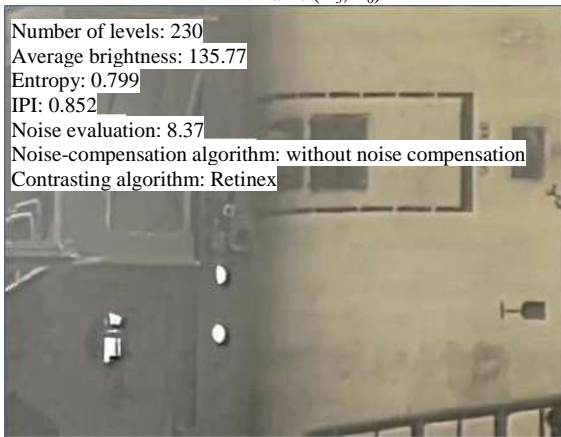
So, the best element is  $(A_3, B_0) \in M$ .



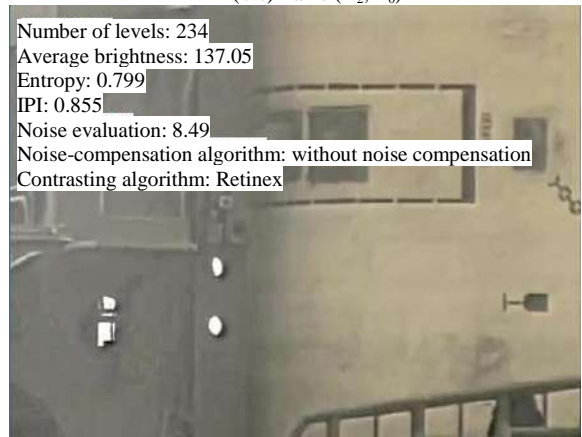
$k$  frame ( $A_3, B_0$ )



$(k+t)$  frame ( $A_2, B_0$ )



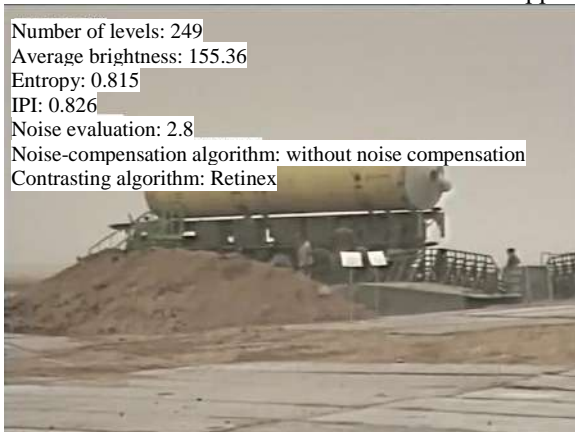
$(k+2*n)$  frame ( $A_3, B_0$ )



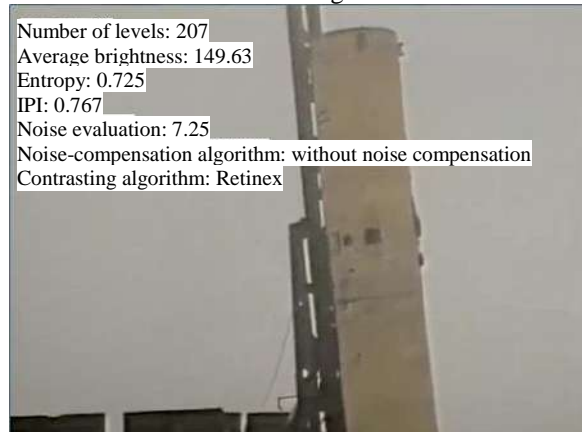
$(k+3*n)$  frame ( $A_3, B_0$ )

Fig. 1. Results of four frame evaluation.

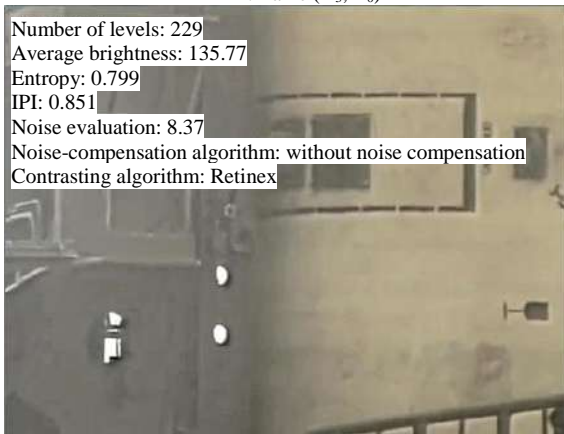
Let's consider the same four frames but now with application of the stack. Result is shown in Fig.2.



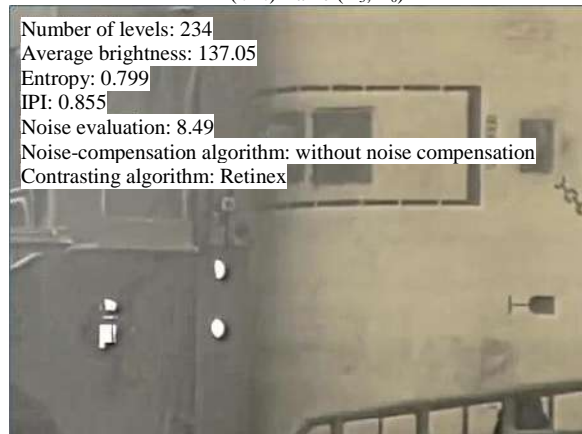
$k$  frame ( $A_3, B_0$ )



$(k+n)$  frame ( $A_3, B_0$ )



$(k+2*n)$  frame ( $A_3, B_0$ )



$(k+3*n)$  frame ( $A_3, B_0$ )

Fig.2. Result of the algorithm operation for automatic selecting using a stack.

We can see that the frame with number  $(k+t)$  is now processed by the same sequence of algorithms as other frames. It allows avoiding undesirable darkening.

Fig.3 shows a diagram of dependence of IPI on a number of the frame. Numbers of frames are marked in horizontal direction and values of quality indices are marked in vertical direction, A0-A3 – contrasting algorithms (A0 –without contrasting, A1 – linear stretch of the histogram [5], A2 – histogram equalization [6], A3 – Multi Scale Retinex with Color Restoration [7]), Auto – IPI values under automatic selection of contrasting algorithms.

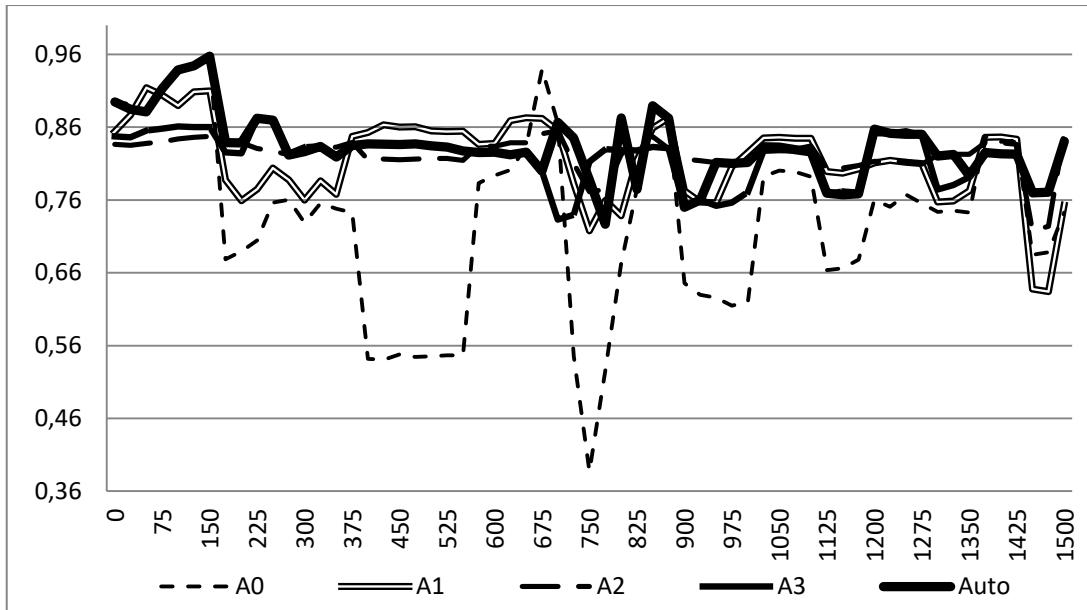


Fig.3. Diagram of dependence of IPI on a number of the frame.

On the diagram we can see areas which value of IPI has not the best value for, it is described by delay occurred because the stack is used and evaluation is performed after every  $t=50$  frames.

## 6. Conclusion

Application of the suggested algorithm for automatic selection of the enhancement method allows automatically finding the best combination of methods for image enhancement. The algorithm provides a possibility to change both a number of used methods of enhancement and also methods for image evaluation not causing significant changes in the algorithm structure. It allows implementing new enhancement algorithms and also new methods of evaluation. Application of the stack for accumulation of processing results and interpolation method of proportional application of two boundary algorithms allows avoiding errors in choice of the best algorithms.

## References

- [1] Elesina SI. Imposition of images in correlation-extreme navigation systems. Edited by Kostyashkina LN, Nikiforov MB. Moscow: Radio Engineering, 2015; 208 p.
- [2] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 2004; 13(4): 600– 612.
- [3] Gurov VS. Image processing in aviation vision systems: monograph. Edited by Kostyashkina LN, Nikiforov MB. Moscow: FIZMATLIT, 2016; 240p.
- [4] Fishburne PS. Theory of utility for decision-making. Moscow: Nauka, 1978; 352 p.
- [5] Gonzalez RC, Woods RE. Digital Image Processing. Prentice Hall, 2 edition, 2002.
- [6] Singh RP, Dixit M. Histogram Equalization: A Strong Technique for Image Enhancement. International Journal of Signal Processing, Image Processing and Pattern Recognition 2015; 8(8): 345– 352.
- [7] Petro AB, Sbert CS, Morel JM. Multiscale Retinex. Image Processing On Line 2014; 4: 71–88.

# Heuristic Malware Detection Mechanism Based on Executable Files Static Analysis

A.V. Kozachok<sup>1</sup>, M.V. Bochkov<sup>2</sup>, E.V. Kochetkov<sup>1</sup>

<sup>1</sup>Academy of the Federal Guard Service, 35, Priborostroitel'naya Street, Oryol, 302034, Russia  
<sup>2</sup>Business risk educational center, 27, Professor Popov Street, Saint-Petersburg, 197022, Russia

---

## Abstract

To ensure the protection of information processed by computer systems is currently the most important task in the construction and operation of the automated systems. The paper presents the application justification of a new set of features distinguished at the stage of the static analysis of the executable files to address the problem of malicious code detection. In the course of study, following problems were solved: development of the executable files classifier in the absence of a priori data concerning their functionality; designing class models of uninfected files and malware during the learning process; development of malicious code detection procedure using the neural networks mathematical apparatus and decision tree composition relating to the set of features specified on the basis of the executable files static analysis. The paper also describes the functional model of malware detection system using the executable files static analysis. The conclusion contains the results of experimental evaluation of the developed detection mechanism efficiency on the basis of neural networks and decision tree composition. The obtained data confirmed the hypothesis about the possibility of constructing the heuristic malware analyzer on the basis of features distinguished during the static analysis of the executable files. However, the approach based on the decision tree composition enables to obtain a significantly lower false negative rate probability with the specified initial data and classifier parameter values relating to neural networks.

*Keywords:* Anti-virus protection, Malware, Neural networks, Decision trees, Heuristic analysis, Machine learning

---

## 1. Introduction

Security of information processed by computer systems poses the most important task for building and operating of automated systems today. Along with that, one of the most dangerous threat is computer malware that can modify (delete) user data, steal confidential information, slowdown or disable operating system. This research substantiates the possibility of using feature space based on the static analysis of executable files for solving the task of heuristic malware detection using the neural networks and decision trees composition mathematical apparatus.

Today there exist different malware detection techniques. The most widely known techniques are the following ones [1]:

- signature-based search (malicious code detection based on byte sequence which definitely characterizes it);
- heuristic search (code detection based on indirect attributes which characterize it as being malicious);
- behavioral mechanisms that affect executing forbidden operations by different processes (e.g., access to critical memory areas or executable code injection into other processes).

All above-listed techniques have essential weak points, i.e. they possess limited capabilities for detection of modified and new viruses, or require the user to be involved in the decision-making process with respect to file belonging to a certain class.

Today the antivirus software cannot guarantee 100% malware protection. The results of tests performed by AV-Comparatives in March 2017 show that heuristic detection rate of new malware strains amounts to approximately 95-98% for most of the modern antivirus software [2].

To detect new malware, heuristic methods or more generally statistical approaches are the most promising research trends nowadays. Some of them based on structural analysis and executable file features [3, 4, 5, 6]. One of the solutions for increasing the effectiveness of heuristic malware detection process is the development of new tools and techniques for malware detection. The purpose of this study is to substantiate the possibility to build a heuristic technique for malware detection based on static analysis of executable files. The distinctive feature of the approach suggested consists in the use of new feature space for building a heuristic detector based on the well-known machine learning techniques, i.e. neural networks and decision trees composition. In this context, a decision on the malware presence will be taken according to a certain law based on availability or absence of totality of features from criteria array defined at the stage of executable file static analysis.

## 2. Forming feature space based on static executable files analysis

In order to substantiate the possibility of using suggested feature space for solving the task of heuristic malware detection, the neural networks and decision trees composition technique has been applied in this study. Let us consider the totality of the features being studied.

The whole feature space may be divided into eight conditional feature groups. Group 1 comprises the features based on the results of characteristics evaluation for the following executable files parts: file header size, optional header, MS-DOS header, digital certificate. Since the structure of the headers has been defined, in case of their size change relevant attribute will be detected. Group 2 comprises the features associated with the use of packing, archiving and encrypting utilities for executable files such as UPX, MPRESS, PeCompact etc. Group 3 comprises information about dynamic libraries, as well as functions exported and imported by the executable file. The rate of certain API-functions and dynamic libraries usage by malware has been precomputed. As a result, two classes have been identified. The first class comprises API-functions by means of which malicious actions can be performed. The second class comprises the rest of the functions. In this context, belonging to a certain class of API-functions is to be regarded as a feature. Group 4 comprises data on digital certificates, namely, whether they are available in the file, whether data are out-of-date or have been recalled. Group 5 comprises the features based on the information about PE-file structure, namely, availability of anonymous section, whether overlay technique is applied, whether the first section is available for writing, whether the control function is transferred by the file to other files, entry point address is out of the file section boundaries, in the first or other sections, whether the last and other sections are of executable type. Feature group 6 is formed based on the manifest, its availability, correspondence of the manifest structure to standard format, whether the administration privilege is requested by the manifest etc. Group 7 comprises information about executable file interface with the operating system, namely, whether the use of Structured Exception Handling (SEH), Data Execution Prevention (DEP) and Address Space Layout Randomization (ASLR) is ignored, whether the application is executed in Visual Basic virtual machine, whether Thread Local Storage (TLS) is used. Group 8 comprises information not included in previous seven groups; for example, whether the file contains rigidly fixed IP-addresses, whether direct cookie links are present, whether databases are in use etc.

Each analyzed executable file is described in the form of a Boolean vector, where one means the feature is available, zero means no feature is available.

## 3. Static heuristic malware detection mechanisms

### 3.1. Machine learning detection mechanism based on static executable files analysis

To increase the effectiveness of heuristic analysis of executable files we suggest using the malware detection technique based on neural networks. Utilization of the neural network mathematical apparatus together with the created feature space will enable us to solve the following tasks:

1. generating class models in the course of learning (uninfected files and malware);
2. developing malware detection procedure through the use of feature vector based on static analysis of executable files;
3. classifying executable files without priori data on their infection with malicious code.

Solution of the task for developing the neural network malware detection technique based on static analysis of executable files comprises two main stages as follows:

1. learning the neural network which defines malware and uninfected file classes (learning subsystem);
2. calculating output values of neural network based on the sequence of features singled out of the files analyzed, and decision-making on files belonging to a certain class (classification subsystem).

Let us consider the learning process in detail. It consists of two main stages:

1. supervised learning the neural networks;
2. adjusting hyper parameter values for decision-making on the file infection status.

At the first stage, the neural network is learnt in a supervised mode. The network is provided with values of both input and priori known output signals, whereas weight coefficients are subject to corrective adjustment in order to increase the accuracy of the neural network learning.

The result of the first learning stage is a weighting coefficients matrix (model) of the learnt neural network.

At the second learning stage for the given learning set it is necessary to evaluate mistake probability values of the first and second grade depending on the hyper parameter values, and adjust them in accordance with the requirements to the first and second grade mistake criticality for further effective functioning of the neural network as a malware detection tool.

As a result of the second learning stage hyper parameter values to be used for executable file classification must be chosen. The learning procedure result is a weighting coefficients matrix and hyper parameter values.

The detection procedure consists of two main stages [7]. During the first stage the detection system input receives the feature sequence singled out of the file analyzed. Then the neural network output values are calculated using weighting coefficients matrices entered into the database.

Depending on output value classifier makes a decision on the analyzed file belonging to a certain software class.

### 3.2. Heuristic malware detection mechanism based on decision trees composition

As an alternative approach, in order to substantiate the possibility to build heuristic malware detection tool based on static analysis of executable files this study provides the results of classifier effectiveness evaluation based on the decision trees composition.

As a rule, composition of algorithms is regarded as a combination of  $N$  algorithms  $d_1(x), \dots, d_N(x)$  in a single one. The idea consists in learning the algorithms and averaging of the obtained responses:

$$a(x) = \frac{1}{N} \sum_{n=1}^N d_n(x). \quad (1)$$

This formula directly answers the regression problem. For the case of binary classification  $d_1(x), \dots, d_N(x)$  it is necessary to take a sign from the resulting formula:

$$a(x) = \text{sign} \frac{1}{N} \sum_{n=1}^N d_n(x). \quad (2)$$

To build a decision trees composition [8] first it is necessary to learn the basic  $N$  algorithms on different subsets singled out from the learning set. To single out random sets, an approach based on the random sets generation from the learning set through removal followed by return procedure (bootstrap) has been applied in this study. At the same time, the size of each subset amounts to  $0,632L$ , where  $L$  is the learning set size.

Additionally, random subspace technique [9] has been applied. The technique consists in choosing the random subset of features for learning each basic algorithm. The number of the features chosen is a hyperparameter of the given technique.

#### 4. Results and Discussion

At the learning stage for both approaches described above it is necessary to create two representative learning sets: uninfected files and malware. A test set is to be created using the files that are not included in the learning file set. During the pre-processing stage, a totality of features is singled out in the form of a Boolean vector from the executable files sent to the system input.

For learning block it is required to single out a totality of feature sequences from the whole totality of files of the representative learning set. For detection block it is required to single out a feature sequence from one file which was received at the classification system input.

Initial data used in the study:

- 1862 uninfected files (system and program files collected from different Windows operating systems);
- 1910 malware files (authors private collection);
- 353 features singled out based on static analysis of executable files.

As a result of performed static analysis of provided file sets, a feature vector has been generated for each executable file of both classes. Groups of performed experiments have confirmed the hypothesis that building a malicious code heuristic analyzer based on the static analysis of executable files is possible.

To evaluate the effectiveness of the neural network technique the following initial data have been used:

- learning rate factor (constant) 0.001;
- accuracy 0.0001;
- number of iterations for reaching the required learning accuracy has been limited by 200.

Hyper parameters to adjust were [10]:

- activation function selection:
  - *relu*, the rectified linear unit function, returns  $f(x) = \max(0, x)$ ;
  - *tanh*, the hyperbolic tan function, returns  $f(x) = \tanh(x)$ ;
  - *logistic*, the logistic sigmoid function, returns  $f(x) = 1/(1 + \exp(-x))$ ;
  - *identity*, no-op activation, useful to implement linear bottleneck, returns  $f(x) = x$ ;
- solver function selection:
  - *adam* refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba [11];
  - *sgd* refers to stochastic gradient descent;
  - *lbfgs* is an optimizer in the family of quasi-Newton methods;
- number of hidden layers and neurons.

The first group of experiments allowed us to select appropriate activation function and solver values. Figure 1 shows the results of experimental evaluation with crossvalidation based on our dataset divided in two sets train (0.7) and test (0.3) and combination of various parameter values. We used one hidden layer with one hundred neurons. As a score value we selected F-measure.

A box plot (box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the F-measure values while the whiskers extend to show the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the inter-quartile range. The mean value is shown by line inside box. Figure 1 shows us approximately equal three possible combination variants: *logistic\_lbfgs*, *tanh\_lbfgs*, *relu\_adam*.

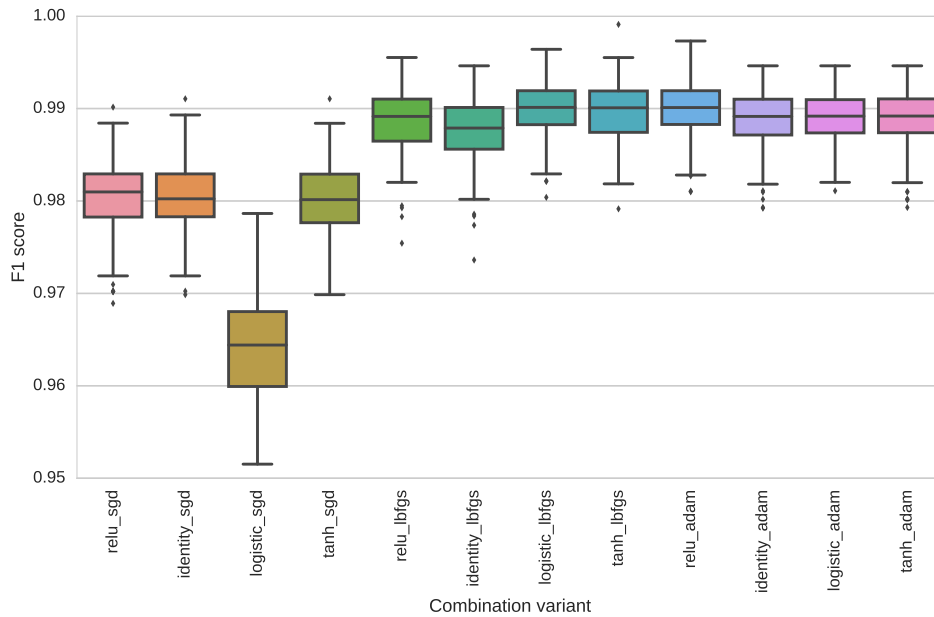


Figure 1: F-measure depending on activation function and solver combination variant.

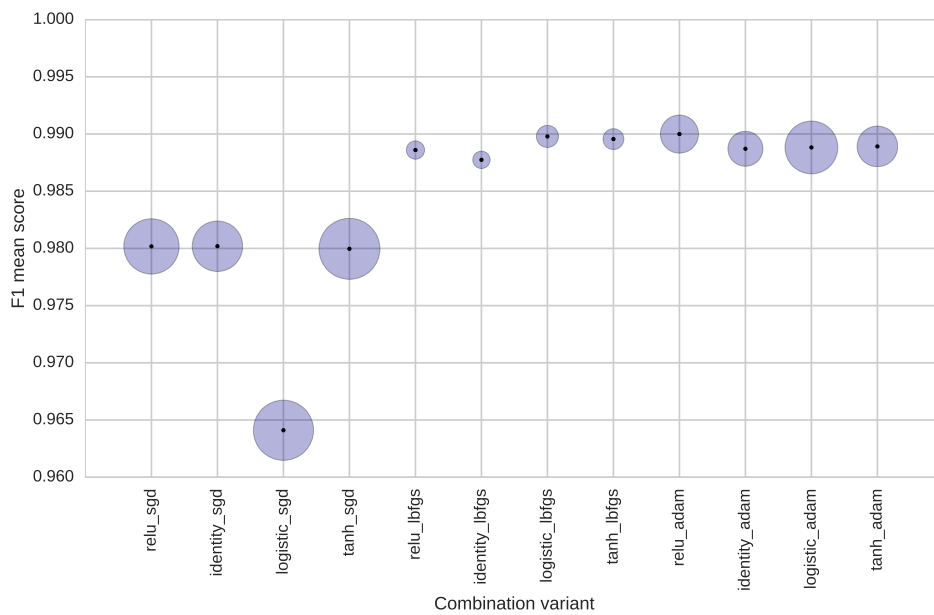


Figure 2: F-measure and crossvalidation time depending on activation function and solver combination variant.

Another group of experiments allowed us to select exactly one combination: tanh\_lbfgs, because it is the fastest one. Figure 2 shows mean F-measure value for all variants (black dots) and blue rounds show relative crossvalidation time.

The third group of experiments allowed us to select a number of hidden layers and neurons. The initial state was with one hidden layer with 10 neurons. We used maximum three hidden layers with 100 neurons. The step was equal to 10 neurons.

Table 1 shows 15 best F-measure values got during experimental evaluation sorted from max to min.



Table 1: F-measure mean crossvalidation score depending on number of hidden layers and neurons

Hidden layer neurons			F-measure mean score
1st	2nd	3rd	
20	70	80	0.99110
20	20	0	0.99109
50	10	100	0.99092
20	10	70	0.99083
20	40	40	0.99082
10	20	60	0.99074
20	30	70	0.99074
70	20	30	0.99074
20	40	0	0.99074
20	70	60	0.99073
20	50	100	0.99047
10	100	90	0.99047
20	50	50	0.99047
20	60	0	0.99047
20	10	50	0.99047

First two rows have closed values, but the first neural network configuration consists of three hidden layers, in spite of the second one with two layers. The first one consumes for about 60% much more time to classify a test set than the second.

The performed groups of experiments have enabled us to substantiate this hyper parameter values:

- activation function – *tanh*;
- solver function – *lbfgs*;
- 2 hidden layers with 20 neurons.

As a result of the developed malware detection technique based on neural network application the following values have been obtained:

- Accuracy = 0.99125;
- F-measure = 0.99110;
- Sensitivity = 0.99425;
- Specificity = 0.99409.

Then the classifier has been learnt based on the decision trees composition, and the obtained result has been cross validated. Figure 3 shows the relation between detection accuracy and number of decision trees. Along with that, the sets of both classes have been divided with the proportion of 0.7 (learning), 0.3 (test).

The performed experimental evaluation of the developed solution allows us to substantiate the following hyper parameters of the classifier:

- number of decision trees is 85;
- number of valuable features is 55 (figure 4).

The following evaluation values have been obtained when choosing the given parameters of the decision trees composition classifier:

- Accuracy = 0.99240;

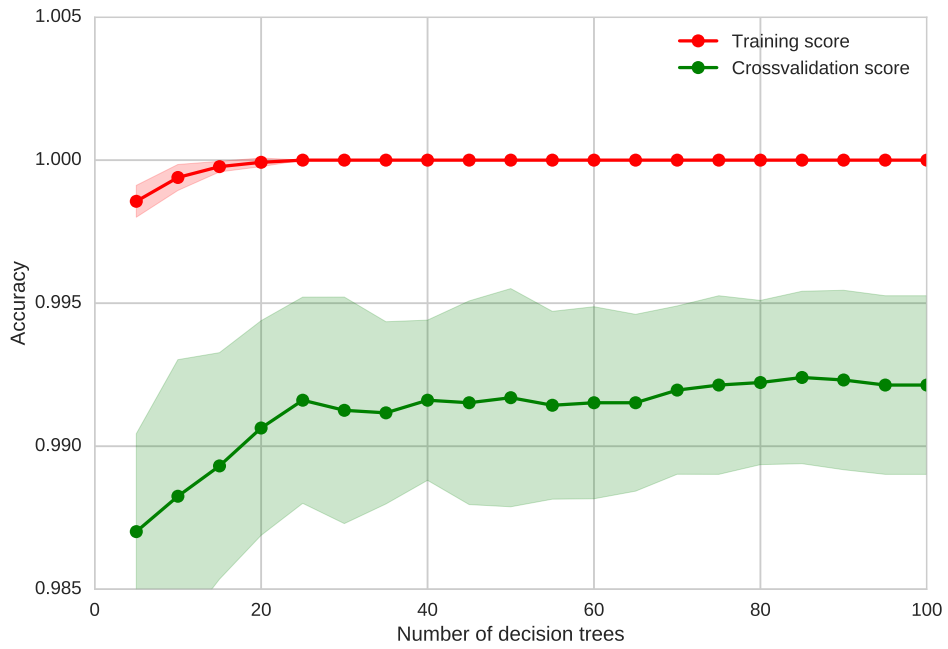


Figure 3: Accuracy value depending on number of trees.

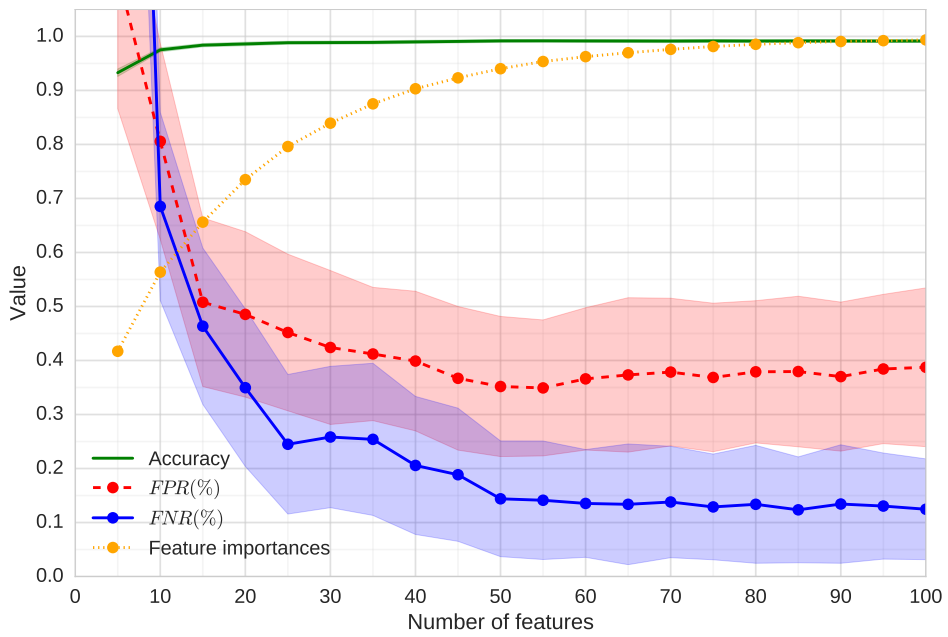


Figure 4: Accuracy value depending on number of features.

- F-measure = 0.99226;
- Sensitivity = 0.99616;
- Specificity = 0.99605.

## 5. Conclusion

It should be noted that both approaches have confirmed the hypothesis that building a malicious code heuristic analyzer based on features singled out during static analysis of executable files is possible. However, the approach based on the decision trees composition allows obtaining better accuracy value relative to neural network tool with the above-mentioned initial data and classifier hyper parameter values.

The obtained values of mistake probabilities for developed prototypes comply with the requirements of the Federal Service for Technical and Export Control [12] imposed to antivirus software.

In conclusion it should be noted that the suggested approach consisting in the application of the space of features singled out from the executable files at the stage of static analysis and well known machine learning techniques enables us to implement a new mechanism for heuristic detection of malicious code which provides the possibility to reveal new and modified malware samples.

## References

- [1] Kozachok, A. V. Mathematical model of destructive software recognition tools based on hidden markov models [Text] / A. V. Kozachok // "Vestnik SibGUT". — 2012. — Vol. 3. — P. 29–39. — (in Russian).
- [2] Anti-Virus Comparative — malware protection test [Electronic resource]. — 2017. — URL: [https://www.av-comparatives.org/wp-content/uploads/2017/04/avc\\_mpt\\_201703\\_en.pdf](https://www.av-comparatives.org/wp-content/uploads/2017/04/avc_mpt_201703_en.pdf).
- [3] Siddiqui, M. A survey of data mining techniques for malware detection using file features [Text] / Muazzam Siddiqui, Morgan C. Wang, Joochan Lee // Proceedings of the 46th Annual Southeast Regional Conference on XX. — ACM-SE 46. — New York, NY, USA : ACM, 2008. — P. 509–510. — URL: <http://doi.acm.org/10.1145/1593105.1593239>.
- [4] Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey [Text] / Asaf Shabtai, Robert Moskovitch, Yuval Elovici, Chanan Glezer // Information Security Technical Report. — 2009. — Vol. 14, no. 1. — P. 16–29. — Malware. URL: <http://www.sciencedirect.com/science/article/pii/S1363412709000041>.
- [5] Opem: A static-dynamic approach for machine-learning-based malware detection [Text] / Igor Santos, Jaime Devesa, Félix Brezo [et al.] // International Joint Conference CISIS12-ICEUTE' 12-SOCO' 12 Special Sessions / Springer. — Ostrava, Czech Republic : Springer-Verlag Berlin Heidelberg, 2013. — P. 271–280.
- [6] David, B. Structural analysis of binary executable headers for malware detection optimization [Text] / Baptiste David, Eric Filiol, Kévin Gallienne // Journal of Computer Virology and Hacking Techniques. — 2017. — Vol. 13, no. 2. — P. 87–93. — URL: <http://dx.doi.org/10.1007/s11416-016-0274-2>.
- [7] Dropout: A simple way to prevent neural networks from overfitting [Text] / Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky [et al.] // J. Mach. Learn. Res. — 2014. — jan. — Vol. 15, no. 1. — P. 1929–1958. — URL: <http://dl.acm.org/citation.cfm?id=2627435.2670313>.
- [8] Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees [Text] / Helmut Schmid. — Manchester, UK : UMIST, 1994.
- [9] Shi, T. Unsupervised learning with random forest predictors [Text] / Tao Shi, Steve Horvath // Journal of Computational and Graphical Statistics. — 2006. — Vol. 15, no. 1. — P. 118–138.
- [10] API design for machine learning software: experiences from the scikit-learn project [Text] / Lars Buitinck, Gilles Louppe, Mathieu Blondel [et al.] // ECML PKDD Workshop: Languages for Data Mining and Machine Learning. — [S. l. : s. n.], 2013. — P. 108–122.
- [11] Kingma, D. Adam: A method for stochastic optimization [Text] / Diederik Kingma, Jimmy Ba // arXiv preprint arXiv:1412.6980. — 2014.
- [12] Federal Service for Technology and Export Control. Informational report on antivirus software requirements approval [Text]. — 2012. — (in Russian).

# Retinamorphic bichromatic Schrödinger metamedia

V. Labunets<sup>1</sup>, I. Artemov<sup>1</sup>, V. Chasovskikh<sup>1</sup>, E. Ostheimer<sup>2</sup>

<sup>1</sup>Ural State Forest Engineering University, 620100, Ekaterinburg, Russia

<sup>2</sup>Capricat LLC, Pompano Beach, Florida, USA

---

## Abstract

In this work, we apply quantum cellular automata (QCA) to study pattern formation and image processing in quantum-diffusion Schrödinger metamedia with generalized complex diffusion coefficients. Generalized complex numbers have the real part and imaginary part with the imaginary unit  $i^2 = -1$  (classical case),  $i^2 = +1$  (double numbers) and  $i^2 = 0$  (dual numbers). They form three 2-D complex algebras. Discretization of the Schrödinger equation gives the quantum Schrödinger cellular automaton with various complex-valued physical parameters. The process of excitation in these media is described by the Schrödinger equations with the wave functions that have values in algebras of the generalized complex numbers. This medium can be used for creation of the eye-prosthesis (so called the "silicon eye"). The medium suggested can serve as the prosthesis prototype for perception of the bichromatic images.

*Keywords:* Schrödinger equation; Schrödinger transform of image; quantum metamedia; quantum cellular automata; silicon eye; quantum image processing

---

## 1. Introduction

The metamedia (metamaterials), in which the electro dynamical, thermal and other physical parameters have "exotic" values (negative, imaginary, complex or quaternion ones), shows us the wonderful diversity of dynamic behavior and self-organization types. It is becoming more and clearer that such systems are not exclusive: when researchers try to investigate the nature of complex systems - chemical, biological or physical, - they find many of certain examples. In particular, this fact mainly refers to biological systems, because these systems are always quite far from stable state and their parameters frequently have exotic values. A theoretical quantum brain model was proposed in [1] using a linear and nonlinear Schrödinger wave equation. The model proposes that there exists a quantum process (quantum part of the brain) that mediates the collective response of a neural lattice (classical part of the brain). Perception, emotion etc. are supposed to be emergent properties of such compound a (classical-quantum) neural circuits.

Linear and nonlinear Schrodinger equations [2,3] are important members of the family of methods for image processing, computer vision, and computer graphics. Schrödinger transform of image as a new tool for image analysis was first given in [3]. In the paper, exterior and interior of objects are obtained from Schrödinger transforms of original image and its inverse image. Neural networks and cellular automata (in form of a media) which are compatible with the theory of quantum mechanics and demonstrate the particle-wave nature of information have been analyzed in [4-6]. The studying of processes in such metamedia is very important for many branches of the system theory. There is no general theory of the metamedia yet, and every particular example of similar media, usually provides us with the examples of new dynamic or self-organization types.

In this work, we apply quantum cellular automata to study pattern formation and image processing in quantum-diffusion Schrodinger metamedia with generalized complex diffusion coefficients. Generalized complex numbers have the real part and imaginary part with the imaginary unit  $i^2 = -1$  (classical case),  $i^2 = +1$  (double numbers) and  $i^2 = 0$  (dual numbers). They form three 2-D complex algebras. Discretization of the Schrödinger equation gives the quantum Schrödinger cellular automaton with various complex-valued physical parameters. The process of excitation in these media is described by the Schrodinger equations with the wave functions that have values in algebras of the generalized complex numbers. This medium can be used for creation of the eye-prosthesis (so called the "silicon eye"). The medium suggested can serve as the prosthesis prototype for perception of the bichromatic images.

The rest of the paper is organized as follows: in Section 2, the object of the study (the Schrödinger equation) is described. In Section 3, a brief introduction to mathematical background (algebra  $A_2(\mathbf{R} | i)$ ,  $i^2 = 1, 0$  of generalized complex numbers  $\mathbf{z} = a + ib$ ) is given (subsection 3.1) in order to understand the concept behind the proposed method. In subsection 3.2, the proposed method based on Schrödinger equations is explained. Next, we defined Schrödinger transform of image, discussed its properties and the properties of the Schrödinger transforms are analyzed. In Section 4, the basic metamedia (the Schrödinger-Euclidean, Schrödinger-Minkowskian, Schrödinger-Galilean, Schrödinger-Yaglom) are devised and analyzed in detail. The simulation result and algorithm complexity are demonstrated too. Finally, we gave our conclusion in Section 5.

## 2. The object of the study

In this work the new metamedia with a complex diffusion coefficients are studied. We call such media the Schrödinger metamedia. Classical 2-D heat equation is:

$$\frac{\partial \varphi(x, y, t)}{\partial t} = D \left( \frac{\partial^2 \varphi(x, y, t)}{\partial x^2} + \frac{\partial^2 \varphi(x, y, t)}{\partial y^2} \right) + f(x, y, t), \quad (1)$$

where  $\varphi(x, y, t)$  is a function describing the media's excitement,  $f(x, y, t)$  is an exciting source (input signal) and  $D$  is a diffusion coefficient (real number).

The main purpose of this work is the investigation of derivative laws for Schrödinger metamedia with generalized complex diffusion coefficient in the form of quantum cellular automata. The generalized complex numbers [7] consist of a real part, an imaginary part and a generalized imaginary unit that have one of the following properties:  $i_-^2 = -1$  (a classical case),  $i_+^2 = +1$  (double numbers) and  $i_0^2 = 0$  (dual numbers). They form three 2-D complex algebras  $A_2(\mathbf{R} | i) := \{z = a + ib \mid a, b \in \mathbf{R}\}$ , where  $i = i_-, i_0, i_+$ . There is a specific type of excitable metamedium for each kind of complex numbers: for  $A_2(\mathbf{R} | i_-)$  - the Schrödinger-Euclidian metamedium (when  $\mathbf{D} = D_{cl} + i_- D_{qu}$ ), for  $A_2(\mathbf{R} | i_0)$  - the Schrödinger-Galilean metamedium (when  $\mathbf{D} = D_{cl} + i_0 D_{qu}$ ) and for  $A_2(\mathbf{R} | i_+)$  - the Schrödinger-Minkowskian metamedium (when  $\mathbf{D} = D_{cl} + i_+ D_{qu}$ ), where  $D_{cl}, iD_{qu}$  are classical and quantum diffusion coefficients, respectively.

Excitation of waves in metamedia are described by three Schrödinger equations with a  $A_2(\mathbf{R} | i)$ -valued wave functions  $\varphi(x, y, t)$ . The discretization of the Schrödinger equations gives us a metamedia models in the form of three excitable cellular automata. Their microelectronic realizations appear to be a programmable Schrodinger metamedia [8].

In this work, we study properties of the Schrödinger excitable metamedium in the form of a cellular automaton. The more detailed information about cellular automata can be found in [9]. The automaton's cells are located inside a 2D array. They can perform basic operations with complex numbers (in different complex algebras  $A_2(\mathbf{R} | i)$ ). These cells are able to inform the neighboring cells about their states. Such media possess large opportunities in processing of bichromatic images in comparison with the ordinary diffusion media with the real-valued diffusion coefficients. The latter media are used for creation of the eye-prosthesis (so called the "silicon eye"). The medium suggested can serve as the prosthesis prototype for perception of the bichromatic images [10-16].

### 3. Methods

#### 3.1. Mathematical background

We consider the algebraic and geometric properties of three 2-D complex algebras  $A_2(\mathbf{R} | i) := \{z = x + iy \mid x, y \in \mathbf{R}\}$ , where  $i = i_-, i_0, i_+$ . Additions for all three algebra are identical:  $\mathbf{z}_1 + \mathbf{z}_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$ , but multiplications are different [7]:

$$\mathbf{z}_1 \mathbf{z}_2 = (x_1 + iy_1)(x_2 + iy_2) = \begin{cases} (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1), & i^2 = -1, \\ (x_1 x_2 + y_1 y_2) + i(x_1 y_2 - x_2 y_1), & i^2 = +1, \\ x_1 x_2 + i(x_1 y_2 + x_2 y_1), & i^2 = 0. \end{cases}$$

The conjugation operation can be defined for  $A_2(\mathbf{R} | i)$ . It maps each number  $\mathbf{z} = x + iy$  to the number  $\bar{\mathbf{z}} = \overline{x + iy} := x - iy$ . It is possible to define a *pseudo norm* using conjugation.

**Definition 1.** The quadratic norm

$$\|\mathbf{z}\| = \mathbf{z} \bar{\mathbf{z}} = \begin{cases} x^2 + y^2, & \mathbf{z} \in A_2(\mathbf{R} | i_-), \\ x^2 - y^2, & \mathbf{z} \in A_2(\mathbf{R} | i_+), \\ x^2, & \mathbf{z} \in A_2(\mathbf{R} | i_0). \end{cases}$$

The conjugation operation can be defined is called the pseudonorm of the number  $\mathbf{z} = x + iy$ . It is easy to check that  $N(\mathbf{z}_1 \mathbf{z}_2) = N(\mathbf{z}_1) N(\mathbf{z}_2)$ .

**Definition 2.** The value of an arithmetical square root of the product of numbers  $\mathbf{z} \bar{\mathbf{z}} = N(\mathbf{z})$  is called an *absolute value of a generalized complex number z* and can be denoted as norm

$$|\mathbf{z}| = \sqrt{\mathbf{z} \bar{\mathbf{z}}} = \sqrt{x^2 - i^2 y^2} = \begin{cases} \sqrt{x^2 + y^2}, & \mathbf{z} \in A_2(\mathbf{R} | i_-), \\ \sqrt{x^2 - y^2}, & \mathbf{z} \in A_2(\mathbf{R} | i_+), \\ |x|, & \mathbf{z} \in A_2(\mathbf{R} | i_0). \end{cases} \quad (2)$$

This absolute value can be interpreted as a distance (elliptic, hyperbolic or parabolic) from origin to the point  $\mathbf{z}$ . In the first case, the absolute value is called *elliptic*, in the second case we are dealing with a *hyperbolic* value (it can also take imaginary values because of the result of subtraction operation  $x^2 - y^2$ ) and in the third case, it is called the *parabolic* absolute value. The generalized complex planes are turned into a 2-D pseudo metrical space if they are equipped the following pseudo metrics:

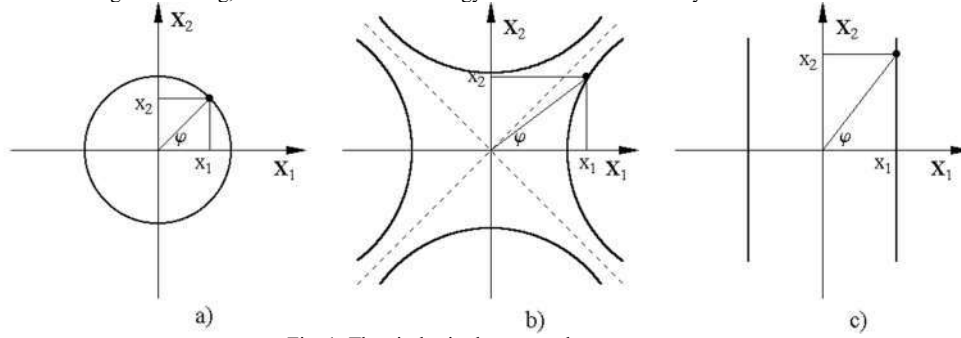


Fig. 1. The circles in three complex spaces.

$$\rho(\mathbf{z}_1, \mathbf{z}_2) := \sqrt{(\mathbf{z}_2 - \mathbf{z}_1)(\mathbf{z}_2 - \mathbf{z}_1)} = \begin{cases} \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, & \mathbf{z} \in A_2(\mathbf{R} | i_-), \\ \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}, & \mathbf{z} \in A_2(\mathbf{R} | i_+), \\ |x_2 - x_1|, & \mathbf{z} \in A_2(\mathbf{R} | i_0), \end{cases} \quad (3)$$

where  $\mathbf{z}_1 = x_1 + iy_1$ ,  $\mathbf{z}_2 = x_2 + iy_2$

The algebra  $A_2(\mathbf{R} | i)$  equipped with pseudo metrics, form three metrical spaces with corresponding geometries:  $A_2(\mathbf{R} | i_-)$  is transformed into the Euclidean geometry,  $A_2(\mathbf{R} | i_+)$  - into the Minkowskian geometry and  $A_2(\mathbf{R} | i_0)$  - into the Galilean geometry.

**Definition 3.** The set of all points in the generalized complex plane  $A_2(\mathbf{R} | i)$  satisfying the equation  $|\mathbf{z}|^2 = x^2 - i^2 y^2 = r^2$  is called  $A_2(\mathbf{R} | i)$ -circle of the radius  $r$  centered at the origin.

There are three types of circles:  $A_2(\mathbf{R} | i_0)$ -circle is the *classical Euclidean (elliptic) circle* (Fig.1a),  $A_2(\mathbf{R} | i_+)$ -circle is the *Minkowskian (hyperbolic) circle* (Fig.1b) and  $A_2(\mathbf{R} | i_0)$ -circle is the *Galilean (parabolic) circle* in the form of two parallel lines (Fig.1c). If  $\mathbf{z} = x + iy$  then the generalized complex number  $\mathbf{z}_0 = \mathbf{z} / |\mathbf{z}|$  has the unit modulus if  $|\mathbf{z}| \neq 0$ . It is easily see, that

$$\mathbf{z} = |\mathbf{z}| \left( \frac{x}{|\mathbf{z}|} + i \frac{y}{|\mathbf{z}|} \right) = |\mathbf{z}| (\mathbf{cos} \beta + i \cdot \mathbf{sin} \beta), \quad (4)$$

where

$$\mathbf{cos} \beta = \frac{x}{\sqrt{x^2 - i^2 y^2}} = \begin{cases} \frac{x}{\sqrt{x^2 + y^2}} = \cos \beta, & \text{if } i = i_-, \\ \frac{x}{\sqrt{x^2 - y^2}} = \text{ch} \beta, & \text{if } i = i_+, \\ \frac{x}{|x|} = \pm 1 = \text{cg} \beta, & \text{if } i = i_0, \end{cases} \quad \mathbf{sin} \beta = \frac{y}{\sqrt{x^2 - i^2 y^2}} = \begin{cases} \frac{y}{\sqrt{x^2 + y^2}} = \sin \beta, & \text{if } i = i_-, \\ \frac{y}{\sqrt{x^2 - y^2}} = \text{sh} \beta, & \text{if } i = i_+, \\ \frac{y}{|x|} = \text{sg} \beta, & \text{if } i = i_0. \end{cases} \quad (5)$$

Here  $\mathbf{cos} \beta$ ,  $\mathbf{sin} \beta$  are generalized trigonometric functions. In the first case ( $i = i_-$ ) generalized trigonometric functions coincide with classical (elliptic) functions:  $\mathbf{cos} \beta = \cos \beta$ ,  $\mathbf{sin} \beta = \sin \beta$ . In the second case ( $i = i_+$ ) they are equal to hyperbolic functions  $\mathbf{cos} \beta = \text{ch} \beta$ ,  $\mathbf{sin} \beta = \text{sh} \beta$ . The third case ( $i = i_0$ ) gives us new kinds of trigonometric functions:  $\mathbf{cos} \beta = \text{cg} \beta \equiv \pm 1$ ,  $\mathbf{sin} \beta = \text{sg} \beta \equiv \beta$  which will be called parabolic (or Galilean) functions.

According to (4)-(6), an arbitrary generalized complex number with the unit modulus has the following form

$$\mathbf{z} = e^{i\beta} = (\mathbf{cos} \beta + i \cdot \mathbf{sin} \beta) = \begin{cases} \cos \beta + i_- \cdot \sin \beta, & \text{if } i = i_-, \\ \text{ch} \beta + i_+ \cdot \text{sh} \beta, & \text{if } i = i_+, \\ \pm 1 + i_0 \cdot \beta, & \text{if } i = i_0. \end{cases} \quad (6)$$

In this work, we study the diffusion equation (or the heat equation) with a diffusion coefficient in the form of a generalized complex number and with  $A_2(\mathbf{R} | i)$ -valued wave function. We will call such equation the *generalized Schrödinger equation*.

### 3.2. The generalized Schrödinger equation and cellular automata

Consider the following 2-D Schrödinger equation

$$\frac{d}{dt} \varphi(x, y, t) = \mathbf{D} \cdot \left( \frac{d^2}{dx^2} + \frac{d^2}{dy^2} \right) \varphi(x, y, t) + f(x, y, t), \quad (7)$$

where  $\varphi(x, y, t)$  is a wave  $A_2(\mathbf{R} | i)$ -valued function. It describes the state  $\varphi(x, y, t)$  (in terms of generalized complex numbers) of a metamedium point with coordinates  $(x, y)$  at the moment  $t$ . In (7)  $\mathbf{D} = D_{cl} + iD_{qu}$  is an  $A_2(\mathbf{R} | i)$ -valued diffusion coefficient. If  $\mathbf{D} \equiv D_{cl} \in \mathbf{R}$  is a real number then (7) is an ordinary diffusion (or heat) equation in the real ordinary medium (we will call one as the *Fourier-Gauss medium*). If  $\mathbf{D} \equiv iD_{qu} \in \mathbf{C}$  is an imaginary number then (7) becomes an ordinary Schrödinger equation with the Plank's constant  $iD_{qu} \in i / 2m$ . If  $\mathbf{D} = D_{cl} + iD_{qu} = |\mathbf{D}| (\cos \beta + i \sin \beta) = |\mathbf{D}| e^{i\beta} \in A_2(\mathbf{R} | i)$ , then (7) is our

generalization of both diffusion and Schrodinger equations. In case of zero initial conditions, we can write the solution (7) in the form of the Cauchy integral:

$$\varphi(x, y, t) = \int_0^t \frac{1}{\left(2\sqrt{\pi\mathbf{D}(t-\tau)}\right)^2} \left( \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(x-\xi)^2 + (y-\eta)^2}{4\mathbf{D}(t-\tau)}} f(\xi, \eta, \tau) d\xi d\eta \right) d\tau. \quad (8)$$

This integral we will call the *generalized Schrödinger transform* (GST) of the initial image  $f(x, y, t)$ . If  $iD_{qu} \in i/2m \in \mathbf{C} = A_2(\mathbf{R}|i_-)$ , then GST is ordinary Schrödinger transform [2-6].

Let us introduce a 2-D *regular lattice* with nodes  $(x_n, y_m, t_k)$ , where  $x_{n+1} = x_n + h$ ,  $y_{m+1} = y_m + h$  and  $t_{k+1} = t_k + \tau$ . Here  $h$  and  $\tau$  are spaces between nodes on the space  $\mathbf{Z}_{Sp}^2 \subset \mathbf{R}^2$  and time  $\mathbf{Z}_t \subset \mathbf{R}_t$  lattices, respectively. For discrete Laplacian we use the following approximation:

$$\begin{aligned} d^2\varphi/dx^2 &= \varphi(x_n+1, y_m, t_k) + \varphi(x_n-1, y_m, t_k) - 2\varphi(x_n, y_m, t_k), \\ d^2\varphi/dy^2 &= \varphi(x_n, y_m+1, t_k) + \varphi(x_n, y_m-1, t_k) - 2\varphi(x_n, y_m, t_k), \\ d^2\varphi/dt &= \varphi(x_n, y_m, t_k+1) - \varphi(x_n, y_m, t_k). \end{aligned} \quad (9)$$

As a result, we get the 2-D discrete Schrödinger equation

$$\begin{aligned} \varphi(x_n, y_m, t_k+1) &= \varphi(x_n, y_m, t_k) + \\ &+ \mathbf{D} \cdot [\varphi(x_n+1, y_m, t_k) + \varphi(x_n-1, y_m, t_k) + \varphi(x_n, y_m+1, t_k) + \varphi(x_n, y_m-1, t_k) - 4\varphi(x_n, y_m, t_k)]. \end{aligned} \quad (10)$$

Now, we give the definition of a 2-D “cellular space” (2-D *regular lattice*) in which the cellular automaton is defined. A regular lattice  $\mathbf{Z}_{Sp}^2 \subset \mathbf{R}_{Sp}^2$  consists of a set of cells (elementary automata, or electrical circuits  $\mathbf{Aut}$ ), which homogeneously cover a 2-D Euclidean space. Each cell is labeled by its position  $\mathbf{Aut}(x_n, y_m) = \mathbf{Aut}(n, m)$ ,  $(n, m) \in \mathbf{Z}_{Sp}^2$ .

Regular, discrete, infinite network consisting of a large number of simple identical elements in the form of elementary automata  $\mathbf{Aut}(n, m)$  a copy of which will take place at each node  $(n, m)$  of the net is called the *cellular automaton* (see Fig.2 and Fig.3a). Each so decorated node will be called a *cell*  $\mathbf{Aut}(n, m)$  and will communicate with a finite number of other cells  $\mathbf{Aut}(i, k)$ , which determine its *neighborhood*  $(i, k) \in \mathbf{M}(m, n)$ , geometrically uniform  $\mathbf{M}(m, n) \equiv \mathbf{M}$ ,  $\forall \mathbf{M}(m, n) \in \mathbf{Z}_{Sp}^2$ . The neighborhood of the cell  $\mathbf{Aut}(n, m)$  (including the cell itself or not, in accordance with convention) is the set of all the cells  $\mathbf{Aut}(i, k)$ ,  $(i, k) \in \mathbf{M}(m, n)$  of the network which will locally determine the evolution of  $\mathbf{Aut}(n, m)$ . This local communication, which is *deterministic, uniform* and *synchronous* determines a *global evolution* of the cellular automaton, along *discrete time steps*  $t_{k+1} = t_k + \tau$ .

In the case of  $\mathbf{Z}_{Sp}^2$ , the classical neighborhoods are the von Neumann’s and Moore’s ones. They are known as the nearest neighbors neighborhoods, and defined according to the usual norms and the associated distances. More precisely, for  $(i, j) \in \mathbf{Z}_{Sp}^2$ ,  $\|(i, j)\|_1 = |i| + |j|$  and  $\|(i, j)\|_\infty = \max(|i|, |j|)$  will denote  $_1$ - and  $_\infty$ -norm respectively. Let  $\rho_1$  and  $\rho_\infty$  be the associated distances. Then Von Neumann and Moore neighborhoods (Fig.2) are  $\mathbf{M}_+(m, n) := \{(i, k) | \rho_1((m, n), (i, k)) \leq 1\}$  and  $\mathbf{M}(m, n) := \{(i, k) | \rho_\infty((m, n), (i, k)) \leq 1\}$ , respectively. To each cell  $\mathbf{Aut}(n, m)$  we assign an  $A_2(\mathbf{R}|i)$ -valued state  $\varphi(n, m, k) = \varphi(x_n, y_m, t_k)$  (i.e., the media’s excitement). The dynamics of the cellular automaton are determined by a local transition rule, which specifies the new state  $\varphi(n, m, k+1) = \varphi(x_n, y_m, t_{k+1})$  of a cell as a function of its interaction Von Neumann neighborhood configuration, according to (10), i.e.,

$$\begin{aligned} \varphi(n, m, k+1) &= \varphi(n, m, k) + \\ &+ \mathbf{D} \cdot [\varphi(n+1, m, k) + \varphi(n-1, m, k) + \varphi(n, m+1, k) + \varphi(n, m-1, k) - 4\varphi(n, m, k)]. \end{aligned} \quad (11)$$

This rule shows us the relation between a state  $\varphi(n, m, k+1)$  of the cell  $\mathbf{Aut}(n, m)$  at the current moment time  $k+1$  and the state  $\varphi(n, m, k)$  the same cell  $\mathbf{Aut}(n, m)$  and the states of the four neighboring cells  $\varphi(n+1, m, k)$ ,  $\varphi(n-1, m, k)$ ,  $\varphi(n, m+1, k)$ ,  $\varphi(n, m-1, k)$  at the previous moment time  $k$ .

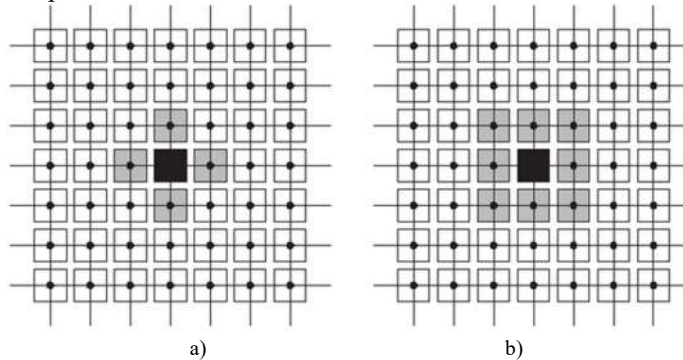


Fig. 2. Examples of interaction neighborhoods (gray and black cells) for the black cell in a 2-D square lattice  $\mathbf{Z}_{Sp}^2$ . Von Neumann neighborhoods  $\mathbf{M}_+(m, n)$  and b) Moore neighborhoods  $\mathbf{M}(m, n)$ .

The *global time evolution* of the cellular automaton depends on an algebraic nature of the number  $\mathbf{D}$ . If it is a real number  $\mathbf{D} \equiv D_{cl} \in \mathbf{R}$  then the automaton simulates the heat propagation on a 2-D plane. According to the results of analysis, in this case an elementary medium’s cell is an ordinary RC-circuit (see Fig. 3b). It is interesting to investigate the global time evolution of the Schrödinger cellular automaton with diffusion coefficient in the form of a generalized complex number

$\mathbf{D} = D_{cl} + iD_{qu} \in A_2(\mathbf{R} | i)$ , where  $i^2 = \pm 1, 0$ . The analysis shows that in this case the elementary cells of a 2-D Schrödinger cellular automata are not RC-circuits, but a 2-channel filters (see Fig. 3c).

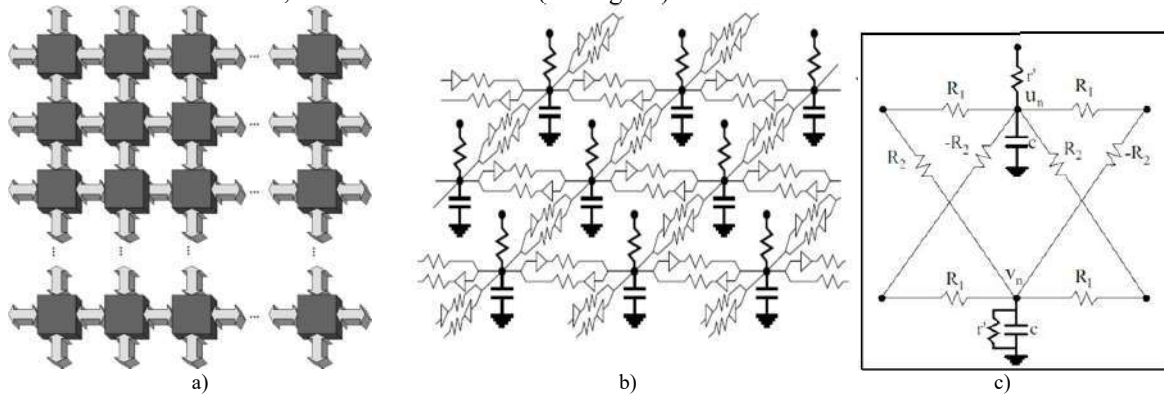


Fig. 3. a) The 2-D cellular automaton (the Schrödinger metamedium) and b) its equivalent electrical circuit in the form of spatially distributed RC-circuit that simulates a simple diffusion media, c) a single cell  $\mathbf{A}ut(n, m)$  of the 2-D cellular automaton in form of a 2-channel (complex) filter.

## 4. Results and Discussion

### 4.1. The Schrodinger-Euclidean metamedium

For studying the global time evolution of the Schrödinger cellular automaton we will use the fixed absolute value of  $\mathbf{D}$ , namely  $|\mathbf{D}| = 0.11$ , which provides quite fast process of diffusion propagation in the classical case with a real-valued diffusion coefficient  $\mathbf{D} \equiv D_0 = 0.11$ , but will not lead to the memory overflow because of extremely high values

For a classical complex case ( $i^2 = -1$ ), the diffusion coefficient can be represented in the polar form:  $\mathbf{D} = D_{cl} + iD_{qu} = \sqrt{D_{cl}^2 + D_{qu}^2} \cdot e^{i\beta} = D_0 \cdot e^{i\beta}$ , where  $D_0 = \sqrt{D_{cl}^2 + D_{qu}^2}$ ,  $\beta = \arctg(D_{qu} / D_{cl})$ . An ordinary diffusion occurs when  $\beta = 0$  (real-valued diffusion coefficient). The quantum diffusion (for free quantum particle) occurs when  $\beta = \pi / 2$  (a purely imaginary diffusion coefficient like the one in the Schrodinger equation). It is interesting to study how the global time evolution is changing when the angle  $\beta$  runs along interval  $0 \leq \beta \leq \pi / 2$ . For  $\beta = 0$  we have the *Fourier-Gaussian medium* (classical Newton world), and for  $\beta = \pi / 2$  we have the *Schrödinger-Euclidean medium* (quantum world). On the  $\beta$ 's increase a classical diffusion Fourier-Gaussian medium turns into the quantum Schrödinger-Euclidean medium.

On Fig. 4 and Fig. 5 the results of modeling for a complex diffusion coefficient  $D$  with different values of phase  $\beta$  are presented. Each picture is divided onto four parts: the bottom row represents a real  $\Re\{\varphi(x, y, t)\}$  and an imaginary  $\Im\{\varphi(x, y, t)\}$  components of a wave excitement in the form of  $A_2(\mathbf{R} | i)$ -valued function  $\varphi(x, y, t)$ , the absolute value  $|\varphi(x, y, t)|$  is presented in the top left quarter, the phase  $\text{Arg}\{\varphi(x, y, t)\}$  is shown in the top right quad.

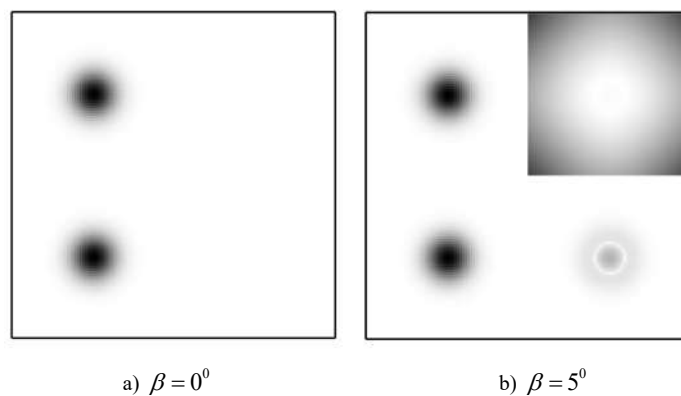


Fig. 4. The excitement of the Schrödinger-Euclidean metamedium at the time  $t_k = 128$  for two values of diffusion coefficient  $\mathbf{D} = D_0 \cdot e^{i\beta}$ , where  $\beta = 0^0$  (the Fourier-Gaussian medium) and  $\beta = 5^0$  (the Schrödinger-Euclidean metamedium).

At the initial moment of time a single cell  $\mathbf{A}ut(x_0, y_0)$  was being excited by the bichromatic Dirac delta-function  $\varphi(x_0, y_0, t = 0) = [\delta(x_0, y_0, 0) + i\delta(x_0, y_0, 0)] = \delta(x_0, y_0, 0)[1 + i]$ . In process of time, the excitement covers more and more cells of automaton. Fig. 4 shows the excitement of a metamedium with two diffusion coefficients:  $\beta = 0^0$  (a real-valued diffusion coefficient) and  $\beta = 5^0$  at the moment of time  $t_k = 128$ . It can be seen that when  $\beta = 0^0$  the excitement takes the form of the 2D Gaussian surface (see Fig. 4a and Fig. 5a) and describes an ordinary diffusion process. Dark intensities correspond to higher





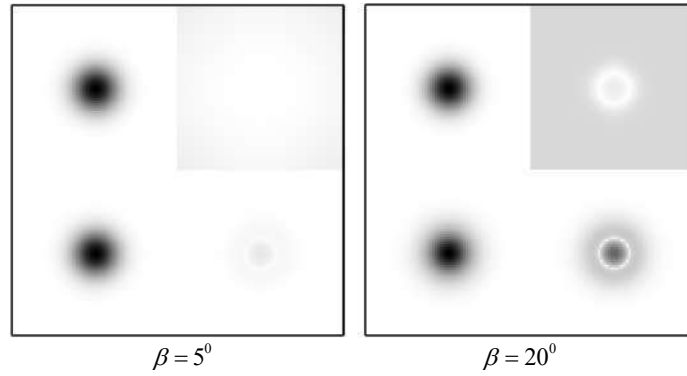


Fig. 7. The excitement of a Schrödinger-Minkowskian metamedium at the moment  $t_k = 128$  for two values of a diffusion coefficient's phase  $\mathbf{D} = D_0 \cdot e^{i \cdot \beta}$  ( $\beta = 5^0$  and  $\beta = 20^0$ ).

Unlike the previous case (when both real and imaginary values had the wave nature) in this case a real component has a smooth Gaussian form and real part has the form of a wave packet. It turned out that the frequency of fluctuations of real values' waves does not increase when the phase of a diffusion coefficient  $D$  grows. In the center of a phase image (top right quarter of pictures) the increase of an angle  $\beta$  leads to the sharper look of a zero phases ring.

#### 4.3. The Schrödinger-Galilean metamedium

In this case the diffusion coefficient  $\mathbf{D}$  is a dual number and wave function  $\varphi(x, y, t)$  takes its values in the algebra of dual numbers  $A_2(\mathbf{R} | i_0) := \{z = a + i_0 b \mid a, b \in \mathbf{R}\}$ , where  $i_0^2 = 0$ . Every dual number can be represented in the following polar form  $\mathbf{D} = D_{cl} + i_0 D_{qu} = |\mathbf{D}|(\pm 1 + i_0 \beta) = |D_{cl}| e^{i_0 \cdot \beta} \in A_2(\mathbf{R} | i_0)$ , where  $|\mathbf{D}| = |D_{cl}|$ ,  $\beta = D_{qu} / |D_{cl}|$ . On Fig. 8 we can see the form of excitement process after 128 iterations from the impact of the Dirac delta-function at the initial moment of time.

As in the previous case, a real component does not demonstrate a wave nature when an imaginary component does. What is more, when we increase the value of  $\beta$  to  $\beta \leq 45^0$  then the average value of a real part becomes lower than the average value of an imaginary part, and when  $\beta > 45^0$  an imaginary component begins to prevail over the real one. In addition, in this case the wave nature of the absolute value of a wave function is absent because of the fact that it does not include a non-zero imaginary part. The reason of this is that for dual numbers we have an equation  $|\mathbf{z}| = |x + i_0 y| = |x|$ . In this way the imaginary part of a wave function is "living on its own", it does not have an impact on an absolute value. So, it is like an invisible "ghost" that accompanies it.

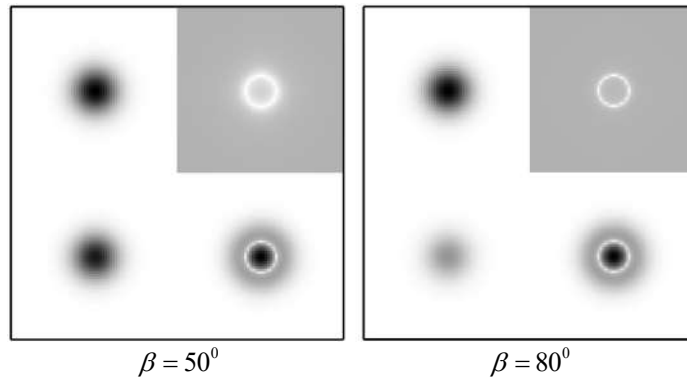


Fig. 8. The excitement of a Schrödinger-Minkowskian metamedium at the moment  $t_k = 128$  for two values of a diffusion coefficient's phase  $\mathbf{D} = D_0 \cdot e^{i \cdot \beta}$  ( $\beta = 5^0$  and  $\beta = 20^0$ ).

#### 4.4. The Schrödinger-Yaglom metamedium

The generalization of three algebra  $A_2(\mathbf{R} | i) := \{\mathbf{z} = a + ib \mid a, b \in \mathbf{R}\}$ , where  $i = i_-, i_0, i_+$ , is the *Yaglom algebra* [7]  $A_2(\mathbf{R} | i_k) := \{\mathbf{z} = a + i_k b \mid a, b \in \mathbf{R}\}$  in which we have  $i_k^2 = k \in \mathbf{R}$ , where  $k$  is an arbitrary real number (see Fig. 9). Particularly when  $k = 1, 0$  an algebra  $A_2(\mathbf{R} | i_k)$  turns into  $A_2(\mathbf{R} | i)$ . In this algebra, the addition and multiplication rules have the following form:

$$\begin{aligned} \mathbf{z}_1 + \mathbf{z}_2 &= (x_1 + i_k y_1) + (x_2 + i_k y_2) = (x_1 + x_2) + i_k (y_1 + y_2), \\ \mathbf{z}_1 \mathbf{z}_2 &= (x_1 + i_k y_1)(x_2 + i_k y_2) = (x_1 x_2 + k \cdot y_1 y_2) + i_k (x_1 y_2 + x_2 y_1). \end{aligned}$$

The conjugation operation can be defined in the algebra  $A_2(\mathbf{R} | i_k)$ . Such operation maps each number  $\mathbf{z} = x + i_k y$  in a new number  $\overline{\mathbf{z}} = \overline{x + i_k y} := x - i_k y$ . It is obvious that  $\|\mathbf{z}\| = \mathbf{z}\overline{\mathbf{z}} = x^2 - ky^2$ . It can be easily seen that

$$\mathbf{z} = |\mathbf{z}| \left( \frac{x}{|\mathbf{z}|} + i \frac{y}{|\mathbf{z}|} \right) = |\mathbf{z}| \cdot \left( \frac{x}{x^2 - k \cdot y^2} + i_k \cdot \frac{y}{x^2 - k \cdot y^2} \right) = |\mathbf{z}| \cdot (\cos_k \beta + i_k \cdot \sin_k \beta) = |\mathbf{z}| \cdot e^{i_k \beta}, \quad (12)$$

where

$$\cos_k \beta := \frac{x}{|\mathbf{z}|} = \frac{x}{\sqrt{x^2 - ky^2}}, \quad \sin_k \beta := \frac{y}{|\mathbf{z}|} = \frac{y}{\sqrt{x^2 - ky^2}}, \quad \text{tg}_k \beta = \frac{\sin_k \beta}{\cos_k \beta} = \frac{b}{a}. \quad (13)$$

In the considered case the diffusion coefficient  $D$  is the  $A_2(\mathbf{R} | i_k)$ -valued complex number and the wave function  $\varphi(x, y, t)$  takes its values in the algebra  $A_2(\mathbf{R} | i_k)$ , where  $i_k^2 = k$ . We will call the corresponding medium *the Schrödinger-Yaglom metamedium*. According to (12) every  $A_2(\mathbf{R} | i_k)$ -valued diffusion coefficient can be represented in a polar form:

$$\mathbf{D} = |\mathbf{D}| \left( \frac{D_{cl}}{|\mathbf{D}|} + i_k \frac{D_{qu}}{|\mathbf{D}|} \right) = |\mathbf{D}| \cdot \left( \frac{D_{cl}}{D_{cl}^2 - k \cdot D_{qu}^2} + i_k \cdot \frac{D_{qu}}{D_{cl}^2 - k \cdot D_{qu}^2} \right) = |\mathbf{D}| \cdot (\cos_k \beta + i_k \cdot \sin_k \beta) = |\mathbf{D}| \cdot e^{i_k \beta}.$$

Now  $\mathbf{D}$  depends on two parameters  $\beta$  and  $k$ . The results of modeling the Schrödinger-Yaglom metamedia for different values of  $k$  are shown on Fig. 10. It should be noted that the ring of zero phases (the bright one), which was inherent for the case with dual numbers ( $k = 0$ ) also is the first inner ring of phase fluctuations for the negative values of a parameter  $k$  (on a Fig. 10  $k = -0,25$  and  $k = -0,05$ ). We can see the second bright ring that is located after the first one and also after the first black ring. The second bright ring moves away from the point of origin when the absolute value of  $k$  is being decreased. When  $|k| \rightarrow 0$  (see Fig. 10c) the mentioned ring along with the first dark one tends to infinity.

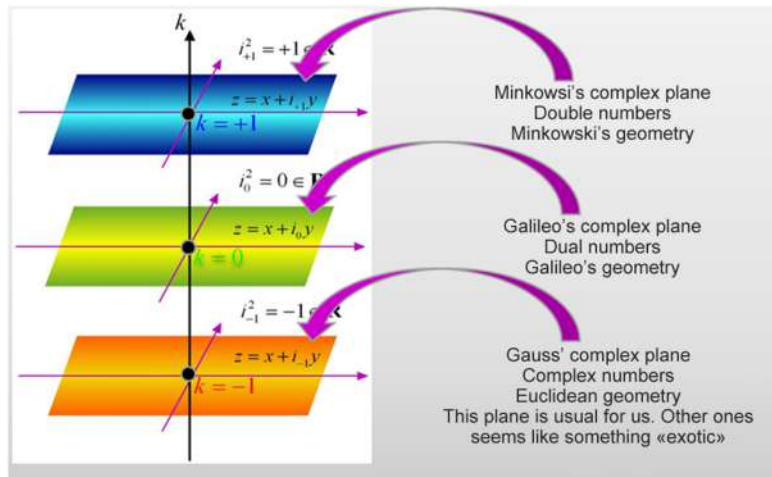


Fig. 9. In every plane, that crosses the vertical of the  $k$ -parameter axis, there is an algebra  $A_2(\mathbf{R} | i_k) := \{z = a + i_k b \mid a, b \in \mathbf{R}\}$ . Three planes that cross this axis at three points  $k = -1, 0$  represent three algebras of complex numbers that were considered before.

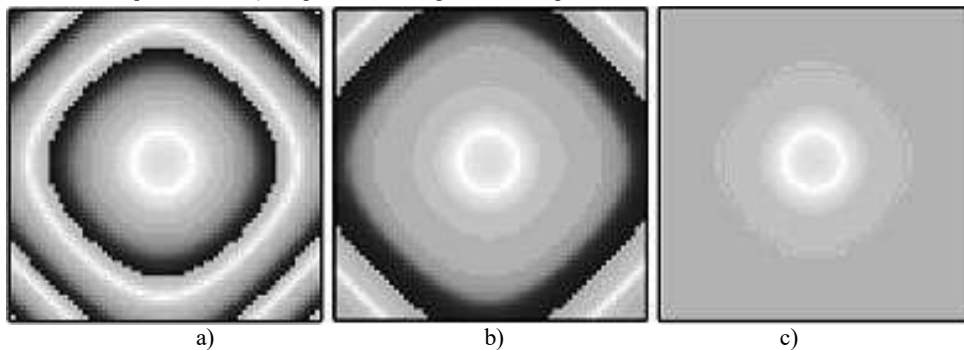


Fig. 10. The excitement of three Schrödinger-Yaglom metamedia at the moment  $t_k = 128$  for three values of the parameter  $k$ : a)  $k = -0,25$ , b)  $k = -0,05$ , c)  $|k| \rightarrow 0$  for identical values  $\arg\{\mathbf{D}\} = 40.5^\circ$  and  $|\mathbf{D}| = 0.07$ .

#### 4.5. The interference of two excitements

Because the excitement function  $\varphi(x, y, t)$  frequently has a wave nature, it is very interesting to study the interference picture of two excitements that appears simultaneously in the different points of a metamedium.

Fig.11a shows us a superposition of two excitations when a diffusion coefficient is a real number. In this case both excitation processes appear to be 2-D Gaussian surfaces that add up with each other in process of time.

More interesting results can be seen on Fig. 11b-c for the Schrödinger-Euclidean metamedium with  $\beta = \arg\{D\} = \pi/2$ ,  $i^2 = -1$ . In that case the interference of excitations occurs, like it happens in a classical quantum mechanics. The results of an interference for the Schrodinger-Galilean metamedium with a dual diffusion coefficient ( $i^2 = 0$ ) are presented on Fig. 11d. Let us note that white rings of the zero phases don't add up with each other like it happens in the case of a classical interference. They are smoothly connecting instead.

## 5. Conclusion

The metamedia with a generalized complex diffusion coefficients were first studied. Their time evolutions are described with generalized Schrodinger equations. The implementation of such metamedia with a cellular automaton was considered. In addition, this work contains the results of modeling, which shown the complex character of such media's behavior. Our future work will be focused on using commutative and Clifford algebras for hyperspectral image processing and pattern recognition.

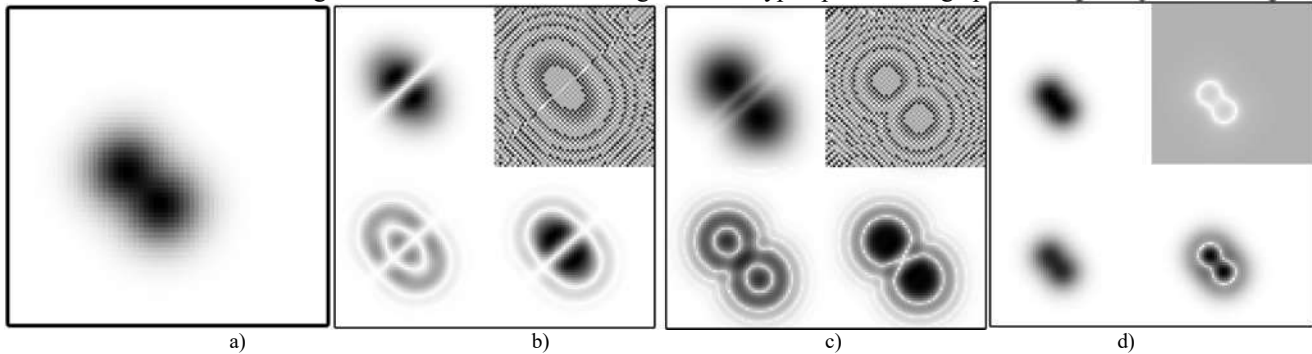


Fig. 11. The interference picture of two excitations in a) the Fourier-Gaussian medium with  $\beta = \arg\{D\} = 0^\circ$ ,  $i^2 = -1$  (real diffusion coefficient); b)-c) the Schrödinger-Euclidean diffusion media (complex diffusion coefficient): b) two closely located points were excited by the Dirac delta-functions at the initial moment of time, c) one points were located relatively far from each other, d) the interference picture of two excitations in the Schrodinger-Galilean metamedium (it has a dual diffusion coefficient).

## Acknowledgements

This work was supported by grants the RFBR № 17-07-00886, № 17-29-03369 and by Ural State Forest University Engineering's Center of Excellence in "Quantum and Classical Information Technologies for Remote Sensing Systems".

## References

- [1] Behera L, Kar I, Elitzur A. Quantum Brain: A Recurrent Quantum Neural Network Model to Describe Eye Tracking of Moving Targets, 2000. URL: <http://arxiv.org:q-bio/quant-ph/0407001v1>.
- [2] Nagasawa M. Schrodinger equations and diffusion theory. Monographs in mathematics. Birkhauser Verlag, Basel, Switzerland, 1993; 86: 238 p.
- [3] Lou L, Zhan X, Fu Z, Ding M. Method of Boundary Extraction Based on Schrödinger Equation. Proceedings of the 21th Congress of the International Society for Photogrammetry and Remote Sensing. Beijing, China 2008; B5(2): 813–816.
- [4] Hagan S, Hameroff SR, Tuzyinski JA. Quantum Computation in Brain Microtubules. Decoherence and Biological Feasibility, Physical Review E, American Physical Society 2002; 65: 1–11.
- [5] Perus M, Bischof H, Caulfield J, Loo CK. Quantum Implementable Selective Reconstruction of High Resolution Images. Applied Optics 2004; 43: 6134–6138.
- [6] Rigatos GG. Quantum Wave-Packets in Fuzzy Automata and Neural Associative Memories. International Journal of Modern Physics C, World Scientific 2007; 18(9): 209–221.
- [7] Yaglom I. Complex numbers in geometry. New York.: Academic press 1968; 242: 203–205.
- [8] Labunets V. Excitable Schrodinger metamedia. 23rd Internation Crimean Conference. Microwave and Telecommunication Technology. Conference proceedings 2013; I: 12–16.
- [9] Wolfram S. Cellular automata as models of complexity. Reprinted from Nature. Macmillan Journals Ltd 1985; 311(5985): 419–424.
- [10] Obeid I, Morizio J, Moxon K, Nicoletis MA, Wolf PD. Two Multichannel Integrated Circuits for Neural Recording and Signal Processing. IEEE Trans Biomed. Eng. 2003; 50: 255–258.
- [11] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Normann R, Solzbacher F. A Low-Power Integrated Circuit for a Wireless 100-Electrode Neural Recording System. International Solid State Circuits Conference 2006; 30.
- [12] Ruedi PF, Heim P, Kaess F, Grenet E, Heitger F, Burgi P-Y, Gyger S, Nussbaum P. A 128 128, pixel 120-dB dynamic-range vision-sensor chip for image contrast and orientation extraction. IEEE J. Solid-State Circuits 2003; 38: 2325–2333.
- [13] Lichtsteiner P, Posch C, Delbruck T. A 128 128 120 dB 30mW asynchronous vision sensor that responds to relative intensity change. IEEE J. Solid-State Circuits 2008; 43: 566–576.

# Retinamorphic color Schrödinger metamedia

V. Labunets<sup>1</sup>, I. Artemov<sup>1</sup>, V. Chasovskikh<sup>1</sup>, E. Ostheimer<sup>2</sup>

<sup>1</sup>Ural State Forest Engineering University, 620100, Ekaterinburg, Russia

<sup>2</sup>Capricat LLC, Pompano Beach, Florida, USA

---

## Abstract

In this work, we use quantum color cellular automata to study pattern formation and image processing in quantum-diffusion Schrödinger systems with triplet-valued (color-valued) diffusion coefficients. Triplet numbers have the real part and two imaginary parts (with two imaginary units  $\varepsilon^1$  and  $\varepsilon^2$ , where  $\varepsilon^3 = 1$ ). They form 3-D triplet algebra. Discretization of the Schrödinger equation gives quantum color cellular automata with various triplet-valued physical parameters. The process of excitation in these media is described by the color Schrödinger equations with the wave functions that have values in triplet algebras. The color Schrödinger metamedia can be used for creation of the eye-prosthesis. The color metamedium suggested can serve as the prosthesis prototype for perception of the color images.

*Keywords:* Color Schrödinger equation; color Schrödinger transform; color metamedia; cellular automata; silicon eye; color image processing

---

## 1. Introduction

Diffusion and Schrödinger equations (linear and nonlinear) [1,2] and real-valued Gauss, complex-valued Fresnel and Schrödinger transforms associated with them are important members of the family of methods for image processing, computer vision, and computer graphics. Schrödinger transform of image as a new tool for image analysis was first given in [2]. Neural networks and cellular automata (in form of a media) which are compatible with the theory of quantum mechanics and demonstrate the particle-wave nature of information have been analyzed in [3-5]. The studying of processes in such metamedia is very important for many branches of the system theory. There is no general theory of the metamedia yet, and every particular example of similar media, usually provides us with the examples of new dynamic or self-organization types.

In this work, we apply quantum cellular automata to study pattern formation and image processing in quantum-diffusion Schrödinger metamedia with triplet-valued diffusion coefficients. Triplet (color) numbers [6] contain one real and two imaginary components with two hyper-imaginary units  $\varepsilon^1$  and  $\varepsilon^2$  and the following property  $\varepsilon^3 = 1$ :  $C = r + g\varepsilon^1 + b\varepsilon^2$ , where  $r, g, b$  are real numbers. The numbers  $C = r + g\varepsilon^1 + b\varepsilon^2$  are called *triplet* or *color numbers*. They form a 3-D triplet (color) algebra  $A_3(\varepsilon) = A_3(\mathbf{R} | 1, \varepsilon^1, \varepsilon^2) := \{C = r + g\varepsilon^1 + b\varepsilon^2 \mid r, g, b \in \mathbf{R}\}$ . If the diffusion coefficient in the Fourier diffusion or Planck's constant in the Schrödinger equations are a triplet number  $D = r + g\varepsilon^1 + b\varepsilon^2$  then both equations are turned into *the color Schrodinger equation*. It describes the process of excitement in the so-called color Schrodinger metamedium with  $A_3(\varepsilon)$ -valued (color) wave function  $\varphi(x, y, t) = \varphi_r(x, y, t) + \varphi_g(x, y, t)\varepsilon^1 + \varphi_b(x, y, t)\varepsilon^2$ .

In this work, we study properties of the color Schrödinger excitable metamedium in the form of a cellular automaton. The more detailed information about cellular automata can be found in [7]. The automaton's cells are located inside a 2D array. They can perform basic operations with triple (color) numbers (in color algebra  $A_3(\mathbf{R} | 1, \varepsilon^1, \varepsilon^2)$ ). These cells are able to inform the neighboring cells about their states. Such media possess large opportunities in processing of color images in comparison with the ordinary diffusion media with the real-valued diffusion coefficient.

The rest of the paper is organized as follows: in Section 2, the object of the study (the color Schrödinger equation) is described. In Section 3, a brief introduction to mathematical background (color algebra  $A_3 = A_3(\mathbf{R} | 1, \varepsilon^1, \varepsilon^2)$  of triplet numbers  $C = r + g\varepsilon^1 + b\varepsilon^2$ ) is given (subsection 3.1) in order to understand the concept behind the proposed method. In subsection 3.2, the proposed method based on color Schrödinger equations is explained. Next, we defined Schrödinger transform of color image, discussed its properties. In Section 4, the basic color metamedia (the color Schrödinger-Euclidean, color Schrödinger-Minkowskian, color Schrödinger-Galilean and color Schrödinger-Yaglom) are devised and analyzed in detail. The simulation result and algorithm complexity are demonstrated too. Finally, we gave our conclusion in Section 5.

## 2. The object of the study

In this work, we apply quantum cellular automata to study pattern formation and image processing in color quantum-diffusion Schrödinger metamedia with triplet-valued diffusion coefficients:

$$\frac{\partial \varphi(x, y, t)}{\partial t} = D \left( \frac{\partial^2 \varphi(x, y, t)}{\partial x^2} + \frac{\partial^2 \varphi(x, y, t)}{\partial y^2} \right) + f(x, y, t), \quad (1)$$

where  $\varphi(x, y, t) = \varphi_r(x, y, t) + \varphi_g(x, y, t)\varepsilon^1 + \varphi_b(x, y, t)\varepsilon^2$  is a color wave function that describes excitement of medium,  $f(x, y, t) = f_r(x, y, t) + f_g(x, y, t)\varepsilon^1 + f_b(x, y, t)\varepsilon^2$  is an exciting color source (input color signal) and  $D = r + g\varepsilon^1 + b\varepsilon^2$  is color-valued diffusion coefficient. It describes the process of excitement in the so-called color Schrodinger metamedium with  $A_3(\varepsilon)$ -valued (color) wave function  $\varphi(x, y, t)$ . Discretization of the color Schrödinger equation gives a color quantum Schrödinger

cellular automaton with various triple-valued physical parameters. Their microelectronic realizations appear to be a programmable Schrodinger metamedia [8]. The main purpose of this work is the investigation of time evolution for color Schrödinger metamedia in the form of quantum cellular automata with triplet diffusion coefficients. The automaton's cells are located inside a 2D array. They can perform basic operations with triple (color) numbers (in color algebra  $A_3(\mathbf{R} | 1, \varepsilon^1, \varepsilon^2)$ ). These cells are able to inform the neighboring cells about their states. Such media possess large opportunities in processing of color images in comparison with the ordinary diffusion media with the real-valued diffusion coefficients. The latter media are used for creation of the eye-prosthesis (so called the "silicon eye"). The medium suggested can serve as the prosthesis prototype for perception of the color images [9-15].

### 3. Methods

#### 3.1. Mathematical background. Triplet algebra

Let us consider the algebraic and geometric properties of the triplet algebra  $A_3 = A_3(\mathbf{R} | 1, \varepsilon^1, \varepsilon^2) := \{C = r + g\varepsilon^1 + b\varepsilon^2 \mid r, g, b \in \mathbf{R}\}$ . The addition and product of two triplet numbers  $C_1 = (r_1 + g_1\varepsilon + b_1\varepsilon^2)$  and  $C_2 = (r_2 + g_2\varepsilon + b_2\varepsilon^2)$  are given by [6]:

$$C_1 + C_2 = (r_1 + g_1\varepsilon^1 + b_1\varepsilon^2) + (r_2 + g_2\varepsilon^1 + b_2\varepsilon^2) = (r_1 + r_2) + (g_1 + g_2)\varepsilon^1 + (b_1 + b_2)\varepsilon^2,$$

$$C_1 \cdot C_2 = (r_1 + g_1 \cdot \varepsilon + b_1 \cdot \varepsilon^2) \cdot (r_2 + g_2 \cdot \varepsilon + b_2 \cdot \varepsilon^2) = (r_1 r_2 + b_1 g_2 + g_1 b_2) + (g_1 r_2 + r_1 g_2 + b_1 b_2)\varepsilon + (b_1 r_2 + g_1 g_2 + r_1 b_2)\varepsilon^2.$$

It is useful to introduce the following triplet numbers  $\mathbf{e}_{lum} := (1 + \varepsilon + \varepsilon^2)/3$ ,  $\mathbf{E}_{chr} := (1 + \omega_3\varepsilon^2 + \omega_3^2\varepsilon)/3$ , where  $\omega_3 = \exp(i \cdot 2\pi/3)$ . It is easy to check  $\mathbf{e}_{lum}^2 = \mathbf{e}_{lum}$ ,  $\mathbf{E}_{chr}^2 = \mathbf{E}_{chr}$ ,  $\mathbf{e}_{lum}\mathbf{E}_{chr} = \mathbf{E}_{chr}\mathbf{e}_{lum} = 0$ . Hence,  $\mathbf{e}_{lum}, \mathbf{E}_{chr}$  are orthogonal idempotents (projectors) and every triplet (color) number  $C = r + g\varepsilon + b\varepsilon^2$  can be represented in the form of the linear combination of a "scalar"  $a_{lum} \cdot \mathbf{e}_{lum}$  and "complex"  $z_{chr} \cdot \mathbf{E}_{chr}$  components  $C = a_{lum} \cdot \mathbf{e}_{lum} + z_{chr} \cdot \mathbf{E}_{chr} = (a_{lum}, z_{chr})$  in the idempotent basis  $\{\mathbf{e}_{lum}, \mathbf{E}_{chr}\}$ , where  $a_{lum} \cdot \mathbf{e}_{lum} \equiv C \cdot \mathbf{e}_{lum}$ ,  $z_{chr} \cdot \mathbf{E}_{chr} = C \cdot \mathbf{E}_{chr}$ , because

$$C \cdot \mathbf{e}_{lu} = (a_{lu} \cdot \mathbf{e}_{lu} + z_{ch} \cdot \mathbf{E}_{ch}) \cdot \mathbf{e}_{lu} = a_{lu} \cdot \mathbf{e}_{lu}^2 + z_{ch} \cdot \mathbf{E}_{ch} \mathbf{e}_{lu} = a_{lu} \cdot \mathbf{e}_{lu} =$$

$$C \cdot \mathbf{E}_{ch} = (a_{lu} \cdot \mathbf{e}_{lu} + z_{ch} \cdot \mathbf{E}_{ch}) \cdot \mathbf{E}_{ch} = a_{lu} \cdot \mathbf{e}_{lu} \mathbf{E}_{ch} + z_{ch} \cdot \mathbf{E}_{ch}^2 = z_{ch} \cdot \mathbf{E}_{ch}.$$

We will call real numbers  $a_{lum} \in \mathbf{R}$  the *luminance numbers* and complex numbers  $z_{chr} \in \mathbf{C}$  - the *chromatic numbers*. Obviously,

$$a_{lum} \cdot \mathbf{e}_{lum} = C \cdot \mathbf{e}_{lum} = (r + g\varepsilon^1 + b\varepsilon^2) \frac{1 + \varepsilon^1 + \varepsilon^2}{3} = (r + g + b) \frac{1 + \varepsilon^1 + \varepsilon^2}{3},$$

$$z_{chr} \cdot \mathbf{E}_{chr} = C \cdot \mathbf{E}_{chr} = (r + g\varepsilon^1 + b\varepsilon^2) \frac{1 + \omega^1\varepsilon^1 + \omega^2\varepsilon^2}{3} = (r + g\omega^1 + b\omega^2) \frac{1 + \omega^1\varepsilon^1 + \omega^2\varepsilon^2}{3}.$$

Hence,  $a_{lum} = r + g + b$ ,  $z_{chr} = r + g\omega^1 + b\omega^2 = \left(r - \frac{g+b}{2}\right) + i \frac{\sqrt{3}}{2}(g-b)$ . In the new duplex representation two main arithmetic operations have the simplest form:

$$C + B = (a_{lum} \cdot \mathbf{e}_{lum} + z_{chr} \cdot \mathbf{E}_{chr}) + (b_{lum} \cdot \mathbf{e}_{lum} + w_{chr} \cdot \mathbf{E}_{chr}) = (a_{lum} + b_{lum}) \cdot \mathbf{e}_{lum} + (z_{chr} + w_{chr}) \cdot \mathbf{E}_{chr},$$

$$C \cdot B = (a_{lum} \cdot \mathbf{e}_{lum} + z_{chr} \cdot \mathbf{E}_{chr}) \cdot (b_{lum} \cdot \mathbf{e}_{lum} + w_{chr} \cdot \mathbf{E}_{chr}) = (a_{lum} b_{lum}) \cdot \mathbf{e}_{lum} + (z_{chr} w_{chr}) \cdot \mathbf{E}_{chr}.$$

Consequently, a color algebra  $A_3(\mathcal{E})$  is the direct sum of real  $\mathbf{R}$  and complex  $\mathbf{C}$  fields:  $A_3(\mathcal{E}) = \mathbf{R} \cdot \mathbf{e}_{lu} + \mathbf{C} \cdot \mathbf{E}_{ch} = \mathbf{R} \oplus \mathbf{C}$ . It is known that every 2-D complex number  $z = x + iy$  can be represented geometrically by the modulus  $\rho = |z| = \sqrt{x^2 + y^2}$  and by the polar angle  $\theta = \arctg(x/y)$ . The modulus  $\rho$  is multiplicative and the polar angle  $\theta$  is additive upon the multiplication of ordinary complex numbers. The triplet numbers introduced in this section have the form  $C = r + g\varepsilon^1 + b\varepsilon^2$ , the variables  $r, g$  and  $b$  being real numbers. In a geometric representation, the triplet number  $C = r + g\varepsilon^1 + b\varepsilon^2$ , is represented by the point  $C(r, g, b)$ , or as a 3-D vector with coordinates  $(r, g, b)$  in the 3-D color space  $\mathbf{R}_{col}^3$  (see Fig. 1).

Let the point  $C$  be the point of the origin of the  $R, G, B$  axes, and  $T_{Ach}$  will be the line, which contains the points with equal coordinates  $r = g = b$  (it is called an *achromatic diagonal*). The luminance numbers  $a_{lum}$  lie on this achromatic diagonal. Also let  $\Delta_M(a_{lum})$  be the plane  $r + g + b = a_{lum}$  that is perpendicular to an achromatic axis  $T_{Ach}$ ; this plane intersects it on a range  $a_{lum}$  from the point of origin  $C$ . It is called a *chromatic plane*. It contains chromatic numbers  $Z_{chr}$ .

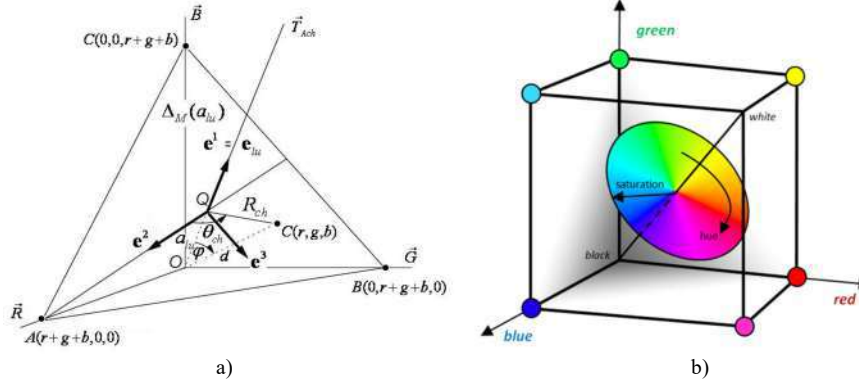


Fig. 1. a) The geometrical representation of a triplet number  $C \equiv r + g\varepsilon + b\varepsilon^2$  in the form of a 3D vector  $C \equiv (r, g, b) \in \mathbf{R}_{col}^3$  or as a point  $C \equiv C(r, g, b) \in \mathbf{R}_{col}^3$  in 3-D color space  $\mathbf{R}_{col}^3$ . The geometrical characteristics have the following values:  $a_{lum} = (r + g + b) / \sqrt{3}$ ,

$$d = \sqrt{r^2 + g^2 + b^2}, R_{chr} = \sqrt{d^2 - a_{lum}^2}, \theta_{chr} = \arg(z_{chr})$$

b) the color cube, it's achromatic diagonal and chromatic plane.

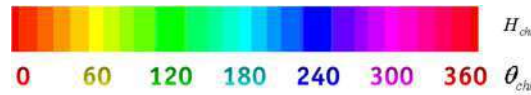


Fig. 2. The relation between an angle  $\theta_{chr}(x, y, t)$  and hue  $H_{chr}(x, y, t)$ .

Obviously, a vector  $C \equiv r + g\varepsilon + b\varepsilon^2 = (r, g, b)$  can be described: 1) by the projection  $a_{lum}$  of a line segment  $OC$  on a line  $T_{Ach}$ , i.e. by the luminance component and 2) by a complex number  $z_{chr}$  in the chromatic plane. Besides, the absolute value of this number appears to be the range  $|z_{chr}|$  from  $C(r, g, b)$  to this line, i.e. it describes the saturation (which is marked by the symbol  $S_{chr} = |z_{chr}|$ ) of a triplet number  $C = r + g\varepsilon + b\varepsilon^2$  and the azimuth angle  $\theta_{chr} = \arg(z_{chr})$  represents its color hue (which we can mark by the symbol  $H_{chr} = \theta_{chr} = \arg(z_{chr})$ ) according to Fig. 2.

### 3.2. The generalized Schrödinger equation and cellular automata

Let a diffusion coefficient  $D = r + g\varepsilon + b\varepsilon^2$  in the Schrödinger equation

$$\frac{\partial \varphi(x, y, t)}{\partial t} = (r + g\varepsilon + b\varepsilon^2) \cdot \left( \frac{\partial^2 \varphi(x, y, t)}{\partial x^2} + \frac{\partial^2 \varphi(x, y, t)}{\partial y^2} \right) + f(x, y, t), \quad (2)$$

be a triplet number, where  $\varphi(x, y, t) = \varphi_r(x, y, t) + \varphi_g(x, y, t)\varepsilon + \varphi_b(x, y, t)\varepsilon^2$  is a color wave function that describes excitement of medium,  $f(x, y, t) = f_r(x, y, t) + f_g(x, y, t)\varepsilon + f_b(x, y, t)\varepsilon^2$  is an exciting color source (input color signal). Color wave function  $\varphi(x, y, t)$  describes time evolution of state  $\varphi(x, y, t)$  (in terms of triplet numbers) of a metamedium point with coordinate  $(x, y)$ . If  $D \equiv D_{cl} = r \in \mathbf{R}$  is a real number and  $\varphi(x, y, t) = \varphi_r(x, y, t)$  then (2) is an ordinary diffusion (or heat) equation for the real ordinary medium (we will call one as the *Fourier-Gauss medium*). If  $D \equiv iD_{qu} \in \mathbf{C}$  is an imaginary number and  $\varphi(x, y, t) = \varphi_{cl}(x, y, t) + i\varphi_{qu}(x, y, t)$  then (2) becomes an ordinary Schrödinger equation with the Plank's constant  $iD_{qu} \in i / 2m$  for ordinary quantum Schrödinger medium. If  $D = D_{cl} + iD_{qu} \in A_2(\mathbf{R} | i)$  and  $\varphi(x, y, t) = \varphi_{cl}(x, y, t) + i\varphi_{qu}(x, y, t)$  then (2) is bichromatic Schrödinger equation for bichromatic quantum Schrödinger metamedium [16]. It is a generalization of both diffusion and Schrödinger metamedia.

In case of zero initial conditions, we can write the solution of (2) in the form of the Cauchy integral:

$$\varphi(x, y, t) = \int_0^t \frac{1}{\left(2\sqrt{\pi D(t-\tau)}\right)^2} \left( \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{(x-\xi)^2 + (y-\eta)^2}{4D(t-\tau)}} f(\xi, \eta, \tau) d\xi d\eta \right) d\tau. \quad (3)$$

This integral we will call the *color Schrödinger transform* (GST) of the initial image  $f(x, y, t)$ . If  $iD_{qu} \in i / 2m \in \mathbf{C} = A_2(\mathbf{R} | i_-)$ , then GST is ordinary Schrödinger transform [1-5].

Let us introduce a 2-D *regular lattice* with nodes  $(x_n, y_m, t_k)$ , where  $x_{n+1} = x_n + h$ ,  $y_{m+1} = y_m + h$  and  $t_{k+1} = t_k + \tau$ . Here  $h$  and  $\tau$  are spaces between nodes on the space  $\mathbf{Z}_{Sp}^2 \subset \mathbf{R}^2$  and time  $\mathbf{Z}_t \subset \mathbf{R}_t$  lattices, respectively. For discrete Laplacian we use the following approximation:

$$\begin{aligned} d^2 \varphi / dx^2 &= \varphi(x_n + 1, y_m, t_k) + \varphi(x_n - 1, y_m, t_k) - 2\varphi(x_n, y_m, t_k), \\ d^2 \varphi / dy^2 &= \varphi(x_n, y_m + 1, t_k) + \varphi(x_n, y_m - 1, t_k) - 2\varphi(x_n, y_m, t_k), \\ d^2 \varphi / dt &= \varphi(x_n, y_m, t_k + 1) - \varphi(x_n, y_m, t_k). \end{aligned} \quad (4)$$

As a result, we get the 2-D discrete color Schrödinger equation

$$\begin{aligned} \varphi(x_n, y_m, t_k + 1) = & \varphi(x_n, y_m, t_k) + \\ & + D \cdot [\varphi(x_n + 1, y_m, t_k) + \varphi(x_n - 1, y_m, t_k) + \varphi(x_n, y_m + 1, t_k) + \varphi(x_n, y_m - 1, t_k) - 4\varphi(x_n, y_m, t_k)]. \end{aligned} \quad (5)$$

Now, we give the definition of a 2-D “cellular space” (2-D *regular lattice*) in which the cellular automaton is defined. A regular lattice  $\mathbf{Z}_{sp}^2 \subset \mathbf{R}_{sp}^2$  consists of a set of cells (elementary automata, or electrical circuits **Aut**), which homogeneously cover a 2-D Euclidean space. Each cell is labeled by its position  $\mathbf{Aut}(x_n, y_m) = \mathbf{Aut}(n, m)$ ,  $(n, m) \in \mathbf{Z}_{sp}^2$

Regular, discrete, infinite network consisting of a large number of simple identical elements in the form of elementary automata  $\mathbf{Aut}(n, m)$  a copy of which will take place at each node  $(n, m)$  of the net is called the *cellular automaton* (see Fig.2 and Fig.3a in [16]). Each so decorated node will be called a *cell*  $\mathbf{Aut}(n, m)$  and will communicate with a finite number of other cells  $\mathbf{Aut}(i, k)$ , which determine its *neighborhood*  $(i, k) \in \mathbf{M}(m, n)$ , geometrically uniform  $\mathbf{M}(m, n) \equiv \mathbf{M}, \forall \mathbf{M}(m, n) \in \mathbf{Z}_{sp}^2$ . The neighborhood of the cell  $\mathbf{Aut}(n, m)$  (including the cell itself or not, in accordance with convention) is the set of all the cells  $\mathbf{Aut}(i, k)$ ,  $(i, k) \in \mathbf{M}(m, n)$  of the network which will locally determine the evolution of  $\mathbf{Aut}(n, m)$ . This local communication, which is *deterministic, uniform* and *synchronous* determines a *global evolution* of the cellular automaton, along *discrete time steps*  $t_{k+1} = t_k + \tau$ .

In the case of  $\mathbf{Z}_{sp}^2$ , the classical neighborhoods are the von Neumann and Moore ones. They are known as the nearest neighbors neighborhoods, and defined according to the usual norms and the associated distances. More precisely, for  $(i, j) \in \mathbf{Z}_{sp}^2$ ,  $\|(i, j)\|_1 = |i| + |j|$  and  $\|(i, j)\|_\infty = \max(|i|, |j|)$  will denote  $l_1$ - and  $l_\infty$ -norm respectively. Let  $\rho_1$  and  $\rho_\infty$  be the associated distances. Then Von Neumann and Moore neighborhoods (Fig.2) are  $\mathbf{M}_+(m, n) := \{(i, k) \mid \rho_1((m, n), (i, k)) \leq 1\}$  and  $\mathbf{M}(m, n) := \{(i, k) \mid \rho_\infty((m, n), (i, k)) \leq 1\}$ , respectively. To each cell  $\mathbf{Aut}(n, m)$  we assign an  $A_2(\mathbf{R}^i)$ -valued state  $\varphi(n, m, k) = \varphi(x_n, y_m, t_k)$  (i.e., the media's excitement). The dynamics of the cellular automaton are determined by a local transition rule, which specifies the new state  $\varphi(n, m, k+1) = \varphi(x_n, y_m, t_{k+1})$  of a cell as a function of its interaction Von Neumann neighborhood configuration, according to (5), i.e.,

$$\varphi(n, m, k+1) = \varphi(n, m, k) + D \cdot [\varphi(n+1, m, k) + \varphi(n-1, m, k) + \varphi(n, m+1, k) + \varphi(n, m-1, k) - 4\varphi(n, m, k)]. \quad (6)$$

This rule shows us the relation between a state  $\varphi(n, m, k+1)$  of the cell  $\mathbf{Aut}(n, m)$  at the current moment time  $k+1$  and the state  $\varphi(n, m, k)$  the same cell  $\mathbf{Aut}(n, m)$  and the states of the four neighboring cells  $\varphi(n+1, m, k)$ ,  $\varphi(n-1, m, k)$ ,  $\varphi(n, m+1, k)$ ,  $\varphi(n, m-1, k)$  at the previous moment time  $k$ .

## 4. Results and Discussion

### 4.1. The Schrodinger-Euclidean metamedium

We can write the Schrödinger equation (2) in the idempotent basis  $\{\mathbf{e}_{lum}, \mathbf{E}_{chr}\}$ . Because

$$\varphi(x, y, t) = \varphi_r(x, y, t) + \varphi_g(x, y, t)\varepsilon^1 + \varphi_b(x, y, t)\varepsilon^2 = \varphi_{lum}(x, y, t) \cdot \mathbf{e}_{lum} + \varphi_{chr}(x, y, t) \cdot \mathbf{E}_{chr},$$

$$f(x, y, t) = f_r(x, y, t) + f_g(x, y, t)\varepsilon^1 + f_b(x, y, t)\varepsilon^2 = f_{lum}(x, y, t) \cdot \mathbf{e}_{lum} + f_{chr}(x, y, t) \cdot \mathbf{E}_{chr},$$

$$D = r + g\varepsilon^1 + b\varepsilon^2 = D_{lum} \cdot \mathbf{e}_{lum} + D_{chr} \cdot \mathbf{E}_{chr},$$

the equation (2) breaks down onto two equations:

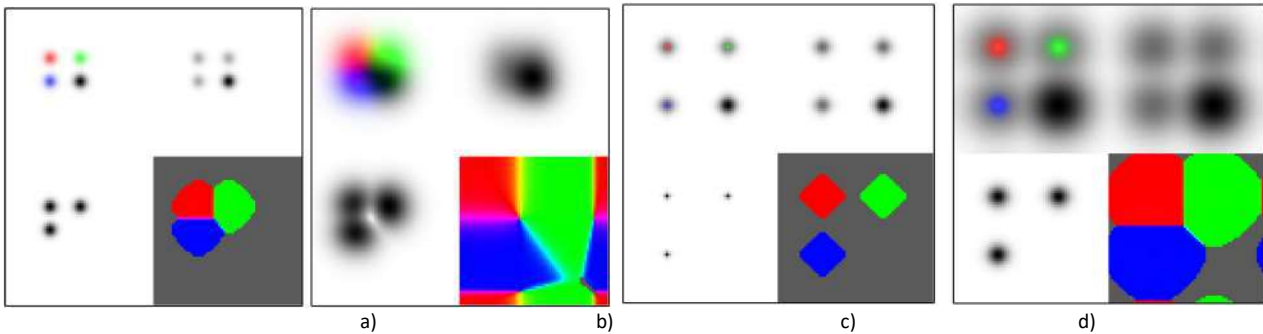


Fig. 3. The state of the color Schrodinger-Euclidean metamedium at moments a)  $t_k = 16$  and b)  $t_k = 210$ , when  $D_{lum} = S_{chr}$ ,  $\theta_{chr} = 0$  and c)  $t_k = 13$  and d)  $t_k = 120$ , when  $D_{lum} \leq S_{chr}$ ,  $\theta_{chr} = 0$ .



$$\begin{aligned}\frac{d}{dt}\varphi_{lum}(x,y,t) &= D_{lum} \cdot \left( \frac{d^2}{dx^2}\varphi_{lum}(x,y,t) + \frac{d^2}{dy^2}\varphi_{lum}(x,y,t) \right), \\ \frac{d}{dt}\varphi_{chr}(x,y,t) &= D_{chr} \cdot \left( \frac{d^2}{dx^2}\varphi_{chr}(x,y,t) + \frac{d^2}{dy^2}\varphi_{chr}(x,y,t) \right).\end{aligned}\quad (7)$$

one for a luminance and other - for a chromatic components.

Obviously,  $\varphi_{chr}(x,y,t) = |\varphi_{chr}(x,y,t)|e^{i\theta_{chr}(x,y,t)} = S(x,y,t)e^{iH_{chr}(x,y,t)}$ , where  $S(x,y,t) = |\varphi_{chr}(x,y,t)|$ ,  $H_{chr}(x,y,t) = \theta_{chr}(x,y,t)$  are the saturation and the hue of a wave function, respectively. The relation between an angle  $\theta_{chr}(x,y,t)$  and a color hue  $H_{chr}(x,y,t)$  is shown on a Fig. 2. The first expression in (7) is the equation of a heat conduction with a real-valued diffusion coefficient  $D_{lum} = r_D + g_D + b_D$ . It describes the time brightness evolution  $\varphi_{lum}(x,y,t)$  of a wave function  $\varphi(x,y,t)$ . The second expression appears to be the Schrodinger equation with a complex diffusion coefficient  $D_{chr} = \left( r_D - \frac{g_D + b_D}{2} \right) + i \frac{\sqrt{3}}{2}(g_D - b_D)$ . It describes the time hue evolution  $\varphi_{chr}(x,y,t)$  of the wave function.

For the modeling results representation we will use the cellular automaton, in which the cell's states are shown as color pixels (as triplet numbers). The sum of four Dirac's delta-functions (red, green, white and blue) as an input signal  $f(x,y,t)$ . On Fig. 3 these functions are represented as four points of the corresponding colors. Each figure consist of four parts: the top left quarter shows the resulting RGB picture (i.e. presents wave function  $\varphi(x,y,t)$  in the RGB format), the top right part shows the luminance component  $\varphi_{lum}(x,y,t)$  of a color wave function, the bottom left one shows the saturation  $|\varphi_{chr}(x,y,t)| = S(x,y,t)$  and the last one represents the color tone  $\theta_{chr}(x,y,t)$ .

Initially we will consider time evolution of the Schrodinger-Euclidean metamedium for the "balanced" chromatic and achromatic parameters  $D = D_{lum} \cdot \mathbf{e}_{lum} + S_{chr} \cdot e^{i\theta_{chr}} \cdot \mathbf{E}_{chr}$  where  $D_{lum} = S_{chr}$ ,  $\theta_{chr} = 0$ , i.e.  $D = D_{lum} \cdot (\mathbf{e}_{lum} + \mathbf{E}_{chr})$ . In this case we take the equal values of a diffusion coefficient's luminance and saturation, when the chromatic phase is equal to zero:  $D_{lum} = S_{chr}$ ,  $\theta_{chr} = 0$ . The results of a simulation for this case are shown on Fig. 3a ( $t_k = 16$ ) and Fig. 3b ( $t_k = 120$ ). Fig. 3c-d shows the process of a color excitement's propagation in a color metamedium, which diffusion coefficient has the low value of saturation ( $D_{lum} \leq S_{chr}$ ,  $\theta_{chr} = 0$ ). The achromatic components on all illustrations in this work are inverted to reduce the amount of dark colors for a better visual perception of pictures. Therefore, the darker colors mean higher values of excitement. Note that chromatic parts of all spots are spreading slower than achromatic ones: the size of spots in the top right quad (excitement's luminance representation) is bigger than in the bottom left one (excitement's saturation representation).

Results that are more interesting can be obtained when we increase the value of a color hue  $\theta_{chr}$  of a diffusion coefficient. As an input signal we use a single red-colored Dirac's delta-function that is affecting the central point of a cellular automaton. For a comparison, it is important to see the excitement of cellular automaton with a zero color hue  $\theta_{chr} = 0^\circ$  (when  $D_{lum} = S_{chr} = 0.11$ ). The results are presented on Fig. 4. This picture shows only resulting RGB images (the top part) and cells' chromatic phases (the bottom part). Also, note the Fig 5.

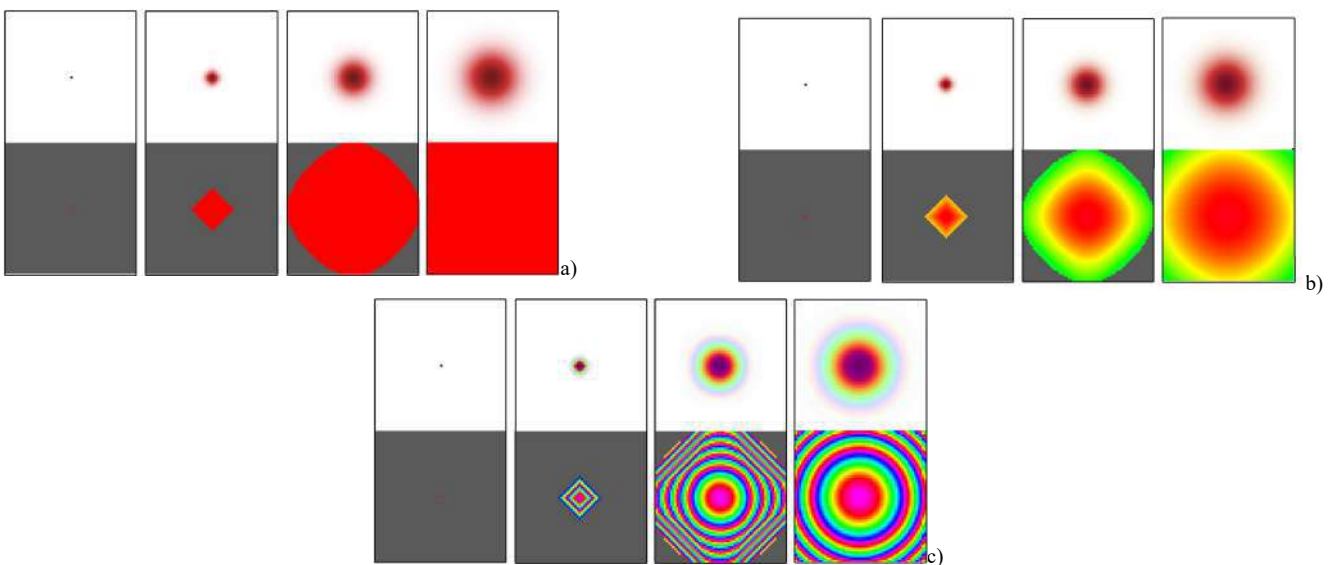


Fig. 4. The state of time evolution of the color Schrodinger-Euclidean metamedium at moments  $t_k = 0, 10, 70, 160$

a)  $\theta_{chr} = 0^\circ$ , b)  $\theta_{chr} = 5^\circ$ , c)  $\theta_{chr} = 60^\circ$ .

4.2. The Schrodinger-Yaglom color metamedia

The chromatic plane, in which  $D_{chr} = |D_{chr}| e^{i \cdot \theta_{chr}} = S_{chr} \cdot e^{i \cdot H_{chr}}$  lays, appears to be a classic complex algebra with  $i^2 = -1$ . It is interesting to study a color metamedium with a chromatic plane in the form of another two complex algebra with  $i^2 = +1$  and  $i_0^2 = 0$  [17], i.e. with the following chromatic components:

$$D_{chr} = |D_{chr}| e^{i \cdot \theta_{chr}} = S_{chr} \cdot e^{i \cdot H_{chr}} \quad \text{and} \quad D_{chr} = |D_{chr}| e^{j_0 \cdot \theta_{chr}} = S_{chr} \cdot e^{j_0 \cdot H_{chr}}.$$

We will call such media the color Schrodinger-Yaglom metamedia. The Fig. 6a contains excitements for color Schrodinger-Galilean metamedium at the moment of time  $t_k = 128$  (for the same input signal as in the previous case) for different values of the hue of the diffusion coefficient ( $\theta_{chr} = 5^\circ, 20^\circ, 40^\circ, 60^\circ$ ). As we can see on a Fig. 6b, the further increase of the hue (for diffusion coefficient)  $\theta_{chr} = 70^\circ, 80^\circ, 89^\circ, 90^\circ$  leads to the fast concentration and contraction of a phase circle in the middle of the bottom right square. In addition, our red-colored initial point completely turns into a spot with a pearl halo when the hue of the coefficient  $D$  reaches  $\theta_{chr} = 90^\circ$ . In addition, we should mention that values  $\theta_{chr} = \arg\{D_{chr}\}$  do not produce any new phenomena because of the periodic nature of trigonometric functions. Indeed, the color excitement with  $\arg\{D_{chr}\} = \theta_{chr} > 90^\circ$  turns out to be the inverted (by a color tone) excitement of a metamedium with  $\arg\{D_{chr}\} = \theta_{chr} - 90^\circ$  (see the Fig. 7a that is quite similar to Fig. 6b).

The example of the excitement of the Schrödinger-Minkowskian metamedium with a chromatic component of a diffusion coefficient in the form of a double number  $D_{chr} = |D_{chr}| e^{i \cdot \theta_{chr}} = S_{chr} \cdot e^{i \cdot H_{chr}}$  is shown on a Fig. 7b.

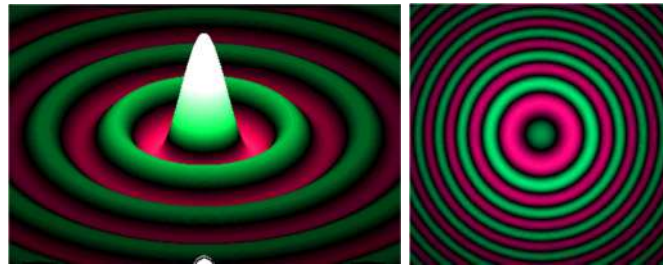


Fig. 5. The typical form of an excitement for Schrodinger-Euclidean metamedium under the impact of an input white Dirac's delta-impulse.

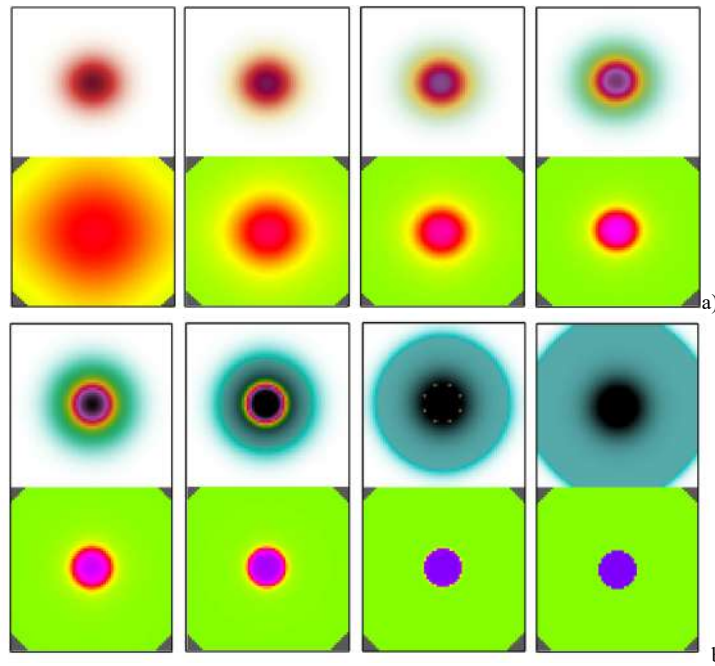


Fig. 6. The state of time evolution of the color Schrodinger-Galilean metamedium ( $i^2=0$ ) at the moment  $t_k=128$ : a)  $\theta_{chr} = 5^\circ, 20^\circ, 40^\circ, 60^\circ$ , b)  $\theta_{chr} = 70^\circ, 80^\circ, 89^\circ, 90^\circ$ .

4.3. The excitement of the color Schrödinger metamedium by a moving source

Let the excitement function  $f(x, y, t)$  in equation (2) be the Dirac delta-function that is moving on the circle with a radius  $R$  and the center at the point  $(x_0, y_0)$ . The source has an angular velocity  $\Omega$ :

$f(x, y, t) = \delta(x_0 + R \cdot \cos(\Omega \cdot t), y_0 + R \cdot \sin(\Omega \cdot t))$ , where  $(x_0 - x(t))^2 + (y_0 - y(t))^2 = R^2$ . It means that we have a moving quantum particle in a color metamedium. Firstly, we will research the color Schrödinger-Euclidean metamedium with the chromatic component in the form of a classical complex number  $D_{chr} = |D_{chr}| e^{i \cdot \theta_{chr}} = S_{chr} \cdot e^{i \cdot H_{chr}}$  that has a relatively low chromatic phase value  $\theta_{chr}$  of  $D_{chr}$ . The Fig.8a-b demonstrates the results of modeling for this case.

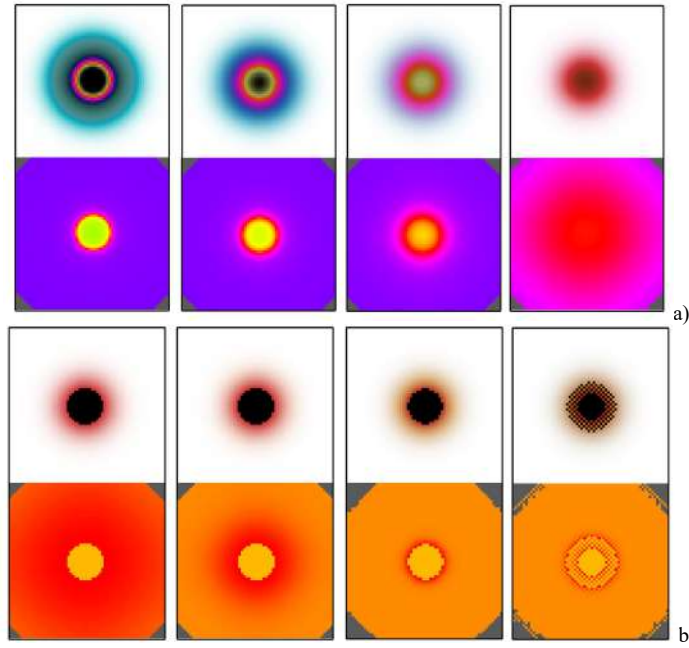


Fig. 7. a) The excitement of a color Schrödinger-Galilean metamedium ( $i^2=0$ ) at the moment  $t_k=128$  for the diffusion coefficients with hues  $\theta_{chr} = 100^\circ, 110^\circ, 130^\circ, 175^\circ$ ; b) The excitement of the color Schrödinger-Minkowskian metamedium ( $i^2=+1$ ) at the moment  $t_k=128$  for the diffusion coefficients with hues  $\theta_{chr} = 100^\circ, 110^\circ, 130^\circ, 175^\circ$ .

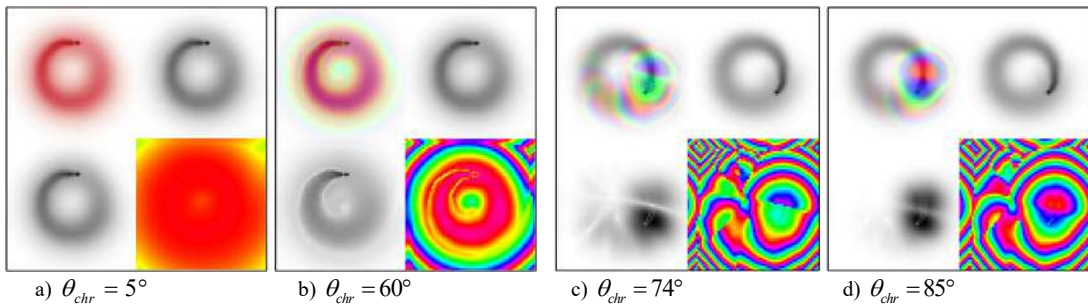


Fig. 8. The excitement of the color Schrödinger-Euclidean metamedium by a particle moving on a circular trajectory ( $t_k = 128$ ).

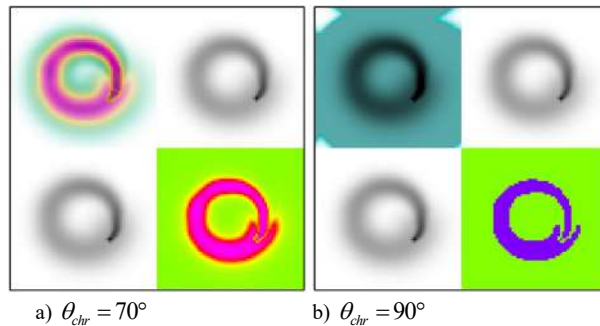


Fig. 9. The excitement of the Schrödinger-Galilean metamedium by a particle moving on a circle ( $t_k = 100, i_0^2 = 0$ ).

When the chromatic angle  $\theta_{chr}$  values are small (low color tone) then  $D_{chr}$ 's excitement fluctuation components, that are perpendicular to the movement trajectory, are almost absent. We only can see the parts of fluctuations that exist along the trajectory. When values of the chromatic angle  $\theta_{chr}$  (hue) are being increased, we can observe the excitement's fluctuations that are perpendicular to the trajectory of a movement. Also the interference of a "tail" and "head" parts becomes visible (see Fig. 8c-d).

Different results can be obtained for color media with a chromatic component in the form of a double  $D_{chr} = S_{chr} \cdot e^{i\theta_{chr}}$  and a dual  $D_{chr} = S_{chr} \cdot e^{i\theta_{chr}}$  number. For example, when we use a dual number  $D_{chr} = S_{chr} \cdot e^{i\theta_{chr}}$  the alteration of  $\theta_{chr} = \arg\{D_{chr}\}$  leads to some interesting and even more unusual consequences. The Fig. 9 shows the pictures of an excitement at the moment  $t_k = 100$  for the quite high values of a phase  $\theta_{chr}$  (the picture of an excitement changes weakly for the wide range of  $\theta_{chr}$ 's values). It can be seen that in the case of a Schrödinger-Galilean metamedium, the growth of a  $D_{chr}$ 's phase causes the increase of a violet and pearl color amounts (on condition that the moving particle has a red color). Particle trail's halo on the right part of Fig. 9 is quite bright, but there are no cells with high lightness and saturation parameter values in the investigated area. It is caused by irregular laws of the behavior of the chromatic component for this metamedia type.

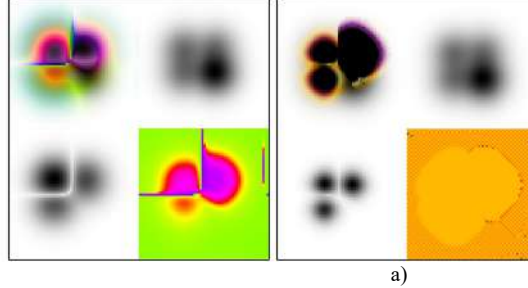


Fig. 10. The interference of four excitations at the moment  $t_k = 128$  in a) the Schrodinger-Galilean metamedium ( $i^2 = 0$ ) and b) the Schrödinger-Minkowskian metamedium ( $i^2 = 1$ ) metamedia ( $\theta_{chr} = 50^\circ$  in both cases).

#### 4.4. The interference of excitations in the color Schrödinger metamedia

The process of excitement's interference in Schrödinger-Euclidean metamedia has a classic character. The results of a simulation for Schrodinger-Galilean ( $i^2 = 0$ ) and Schrödinger-Minkowskian ( $i^2 = 1$ ) metamedia are shown on Fig. 10a and Fig. 10b, respectively. It can be seen on Fig. 10a that in the Schrödinger-Galilean metamedium the collision of different-colored excitations produces unusual rays in the areas where an occlusion happened. There are no such phenomena in the Schrodinger-Euclidean and in the Schrödinger-Minkowskian metamedia.

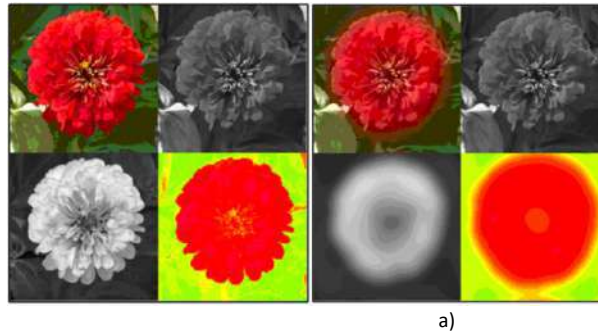


Fig. 11. a) The excitement function (input image)  $f(x, y, t_0 = 0)$  of the color Schrödinger-Euclid metamedium at the initial moment  $t_0 = 0$ ; b) The excitement of this metamedium at the moment  $t_k = 128$ . The metamedium has broken the image onto the areas of uniformity with respect to brightness and hue at this time.

#### 4.5. Some applications of the color Schrödinger metamedia

Let the excitement function  $f(x, y, t_0 = 0) = f_r(x, y, 0) + f_g(x, y, 0)\varepsilon + f_b(x, y, 0)\varepsilon^2 = f_{lum}(x, y, 0) \cdot \mathbf{e}_{lum} + f_{chr}(x, y, 0) \cdot \mathbf{E}_{chr}$  in (2) represents a color RGB image at the moment  $t_0 = 0$ . Then the color wave function

$$\varphi(x, y, t) = \varphi_r(x, y, t) + \varphi_g(x, y, t)\varepsilon + \varphi_b(x, y, t)\varepsilon^2 = \varphi_{lum}(x, y, t) \cdot \mathbf{e}_{lum} + \varphi_{chr}(x, y, t) \cdot \mathbf{E}_{chr} \quad (8)$$

shows us the time evolution of initial image  $f(x, y, t_0 = 0) = \varphi(x, y, 0)$ . As an example of such image, we take a flower in an RGB format (see Fig. 11b, top left quarter). The luminance component  $\varphi_{lum}(x, y, t)$  of wave function (8) represented in the bottom left part of Fig. 11b, the saturation component  $|\varphi_{chr}(x, y, t)|$  - in the top right quarter and the hue  $\theta(x, y, t) = \arg\{\varphi_{chr}(x, y, t)\}$  - in the bottom right part.

One of the most important tasks in the digital processing of color images [18] is the distinguishing of image's parts, where some of its components have uniform values. It is the uniformity areas detection, for example, we can detect the areas with a similar brightness, saturation or color tone, etc. Usually one has to perform such operation before starting the image segmentation by some parameter. It turns out that color Schrödinger metamedia are able to implement such operations. Fig. 11b



shows the excitement of a Schrödinger-Euclidean metamedia at the moment  $t=128$  after an impact that is represented as an image, which was described previously. It is easy to see that by this time the metamedia has broken the initial image onto areas of uniformity by luminance and by color tone. Fig. 12 and Fig. 13 shows the excitements of the Schrödinger-Galilean and Schrödinger-Minkowskian metamedia at the moments  $t_k = 0, 32, 64, 128, 160$  and  $t_k = 0, 84$ , respectively, after an input impact in the form of an initial image.

## 5. Conclusion

The metamedia with triplet (color) diffusion coefficients were first studied. Their laws of functioning are described by color Schrödinger equations. Simulation of these equations in the form of quantum cellular automata was considered. The results of modeling that were shown in this work demonstrate the complex character of the time evolution of such metamedia. Our future work will be focused on using commutative and Clifford algebras for hyperspectral image processing and pattern recognition.

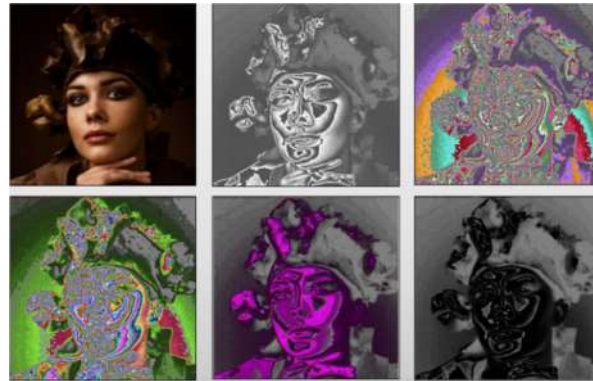


Fig. 12. The excitement of a Schrödinger-Galilean metamedia at moments of time  $t_k = 0, 32, 64, 128, 160$  when an input signal had the form of the initial image in the top left quarter.

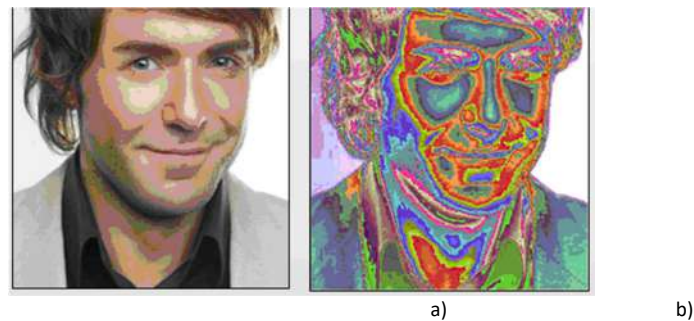


Fig. 13. The excitement of a Schrödinger-Minkowskian at the moment of time  $t_k = 0,84$  when an input signal had the form of an image.

## Acknowledgements

This work was supported by grants the RFBR № 17-07-00886, № 17-29-03369 and by Ural State Forest University Engineering's Center of Excellence in "Quantum and Classical Information Technologies for Remote Sensing Systems".

## References

- [1] Nagasawa M. Schrödinger equations and diffusion theory. Monographs in mathematics. Birkhauser Verlag, Basel, Switzerland 1993; 86: 238 p.
- [2] Lou L, Zhan X, Fu Z, Ding M. Method of Boundary Extraction Based on Schrödinger Equation. Proceedings of the 21th Congress of the International Society for Photogrammetry and Remote Sensing – ISPRS. Beijing, China 2008; B5(2): 813–816.
- [3] Hagan S, Hameroff SR, Tuzynski JA. Quantum Computation in Brain Microtubules. Decoherence and Biological Feasibility, Physical Review E, American Physical Society 2002; 65: 1–11.
- [4] Perus M, Bischof H, Caulfield J, Loo CK. Quantum Implementable Selective Reconstruction of High Resolution Images. Applied Optics 2004; 43: 6134–6138.
- [5] Rigatos GG. Quantum Wave-Packets in Fuzzy Automata and Neural Associative Memories. International Journal of Modern Physics C, World Scientific 2007; 18(9): 209–221.
- [6] Greaves Ch. On algebraic triplets. Proc. Irish Acad. 1847; 3: 51–54, 57–64, 80–84, 105–108.
- [7] Wolfram S. Cellular automata as models of complexity. Reprinted from Nature. Macmillan Journals Ltd 1985; 311(5985): 419–424.
- [8] Labunets V. Excitable Schrödinger metamedia. 23rd International Crimean Conference. Microwave and Telecommunication Technology. Conference proceedings 2013; I: 12–16.
- [9] Obeid I, Morizi J, Moxon K, Nicoletis MA, Wolf PD. Two Multichannel Integrated Circuits for Neural Recording and Signal Processing. IEEE Trans Biomed. Eng. 2003; 50: 255–258.

- [10] Harrison R, Watkins P, Kier R, Lovejoy R, Black D, Normann R, Solzbacher F. A Low-Power Integrated Circuit for a Wireless 100-Electrode Neural Recording System. International Solid State Circuits Conference 2006; Session 30.
- [11] Ruedi PF, Heim P, Kaess F, Grenet E, Heitger F, Burgi P-Y, Gyger S, Nussbaum P. A 128 128, pixel 120-dB dynamic-range vision-sensor chip for image contrast and orientation extraction. IEEE J. Solid-State Circuits 2003; 38: 2325–2333.
- [12] Lichtsteiner P, Posch C, Delbruck T. A 128 128 120 dB 30mW asynchronous vision sensor that responds to relative intensity change. IEEE J. Solid-State Circuits 2008; 43: 566–576.
- [13] Zaghoul K, Boahen K. A. Optic nerve signals in a neuromorphic chip: Part 1. IEEE Trans. Biomed Eng. 2004; 51: 657–666.
- [14] Zaghoul K, Boahen K. A. Optic nerve signals in a neuromorphic chip: Part 2. IEEE Trans. Biomed Eng. 2004; 51: 667–675.
- [15] Mojarradi M, Binkley D, Blalock B, Andersen R, Uslhoefer N, Johnson T, Del Castillo L. A miniaturized neuroprosthesis suitable for implantation into the brain. IEEE Trans. Neural Syst. Rehabil. Eng. 2003; 11: 38–42.
- [16] Labunets V, Artemov I, Chasovskikh V, Ostheimer E. Retinomorphic bichromatic Schrödinger metamedia. CEUR Workshop Proceedings 2017; 1901: 140-148. DOI: 10.18287/1613-0073-2017-1901-140-148.
- [17] Yaglom I. Complex numbers in geometry. New York: Academic Press 1968; 242: 203–205.
- [18] Rosin P, Adamatzky A, Sun X. Cellular Automata in Image Processing and Geometry. Switzerland: Springer International Publishing 2014; 65–80.

# Automated pathological growths parametrization based on computed tomography layer segmentation

N.I. Limanova<sup>1</sup>, S.G. Ataev<sup>1</sup>

<sup>1</sup>*Volga State University of Telecommunications and Informatics, Lev Tolstoy str. 23, 443010, Samara, Russia*

---

## Abstract

Nowadays the multi-layer computed tomography (CT) shots with contrast dye used for detection of small pathological growths regions are widespread in medical clinics. Although, CT study with contrast dye use has contraindications and is much more expensive for patients than study without use of dye. This article proposes the algorithm based on segmentation of multi-layer CT (without dye use). The algorithm allows to automatically identify pathological growths and to detect studying object regions. Algorithm also evaluates parameters of small objects and sinuses, which occupy only small part of initial CT layers, with high accuracy.

*Keywords:* Image analysis; computed tomography; computer-aided diagnosis; parameterization; software development

---

## 1. Introduction

X-ray computer tomography (CT) was developed in the 1970s and mass adopted in 1980s. Now CT is widely used for meeting different diagnostic needs in outpatient and inpatient medical care. Accessibility of CT scan devices led to the development of computer-aided diagnosis (CAD) systems, whose task is to increase the accuracy and speed of diagnosis process. CAD systems provide medics an additional data extracted from CT scans. Currently, medics are just beginning to use CADs in their everyday medical practice, but CAD development fundamentals is actively developing field of research [1, 2]. Some cases require to evaluate numerical parameters of small anatomical structures (such as measuring the volume of maxillary sinus and examining it if there are the pathological growths). This article proposes the algorithm that allows to calculate automatically the parameters of CT scan objects, including sinuses and small structures taking a small part of the picture.

The development of multi-layer CT has led to increased speed and effectiveness of medical examination. Array of multi-layer CT layers represents comprehensive information about inner structure of the examining object. Nowadays, there is a lot of CT visualization improvement [3, 4] and 3D-reconstruction methods are developed [5, 6, 7]. They can simplify CT scan reading process, but can't lead to automatic parameter evaluation. Also, there are some methods that analyze CT layers separately, for example pulmonary nodes classification methods [8, 9, 10]. This kind of methods is suitable for object classification problems, which can be solved with the use of neural networks [11]. All these methods can't be used for automated parameterization of small anatomic structures situated in small part of CT layers.

Nowadays the CT study with the use of contrast dye is widely used for detection of small pathological growths areas. The CT of maxillary sinus with contrasting should be performed after studying without contrasting. The purpose of this additional study is the differentiation of the maxillary sinus growth (such as cyst, polypoid growths, etc.). Contrast dye accumulates in tumor cells and marks them at CT scan layers, as a result soft tissues become distinctly visible. However, the CT study with use of contrast has contraindications. It is significantly more expensive and requires additional CT session. The article proposes CT scan segmentation algorithm that implements detection of the anatomic structure area and allows automatic detection of the pathological growths on CT scans, which are taken without contrast dye use.

## 2. The structure of multi-layer CT scan

A multi-layer CT scan structure gives a possibility for medics to receive comprehensive information about inner structure of scanned objects. Thickness of a single layer can vary (within tomograph capabilities) and is selected depending on specifics of the studied organs. Distinctive feature of multi-layer CT is placement of the studied objects on several layers at once. Each layer of a tomogram projection gives exact value of tissue density displayed by the corresponding pixels in a picture. The pixels have gradation of color from light to dark shades of gray. The gray shade is lighter, the tissue within pixel is denser. Thus, corresponding to studying object pixel set contains an object structure comprehensive data and allows its automated analysis and study [12].

Because of low brightness resolution of monitors that can't represent whole range of possible densities, every CT scan has its own density range called "Hounsfield scale". Density of any tissues shown on CT layer is measured in Hounsfield units. This range is selected depending on type of studying object. The tissues inside the "Hounsfield scale" are represented by shades of gray, which brightness is in direct ratio of the tissue density, while tissues outside the range are representing by black or white color (less than minimum or more than maximum, respectively). This range is set by a two numbers - center and width. CT scan layer example is shown on fig. 1. Values W: 90 and C: 40 (Hounsfield units) are width and center of Hounsfield scale.



Fig. 1. CT layer example, which contains Hounsfield Scale information.

### 3. The algorithm for automated small object parameterization

Consider the proposed automated object parameterization algorithm and the exemplary problem of maxillary sinus study. Task of automated object parameterization has been divided into two subtasks:

- CT layers segmentation (finding an object region and corresponding subset of pixels related to this object);
- Object pixels analysis (evaluating required parameter values of the object of study).

Proposed CT layer segmentation algorithm starts from receiving from user a set of CT layers. Then user should specify starting point of studying object search and set the criterion value, required for checking the pixels belonging to this object. This criterion applies as a threshold value in further algorithm performance and can be specified as a brightness value for layer pixels in range of 0 (black) to 1 (white) or as density value for tissues (in Hounsfield Units). Fig. 2 represents developed software graphic user interface with loaded set of CT layers.

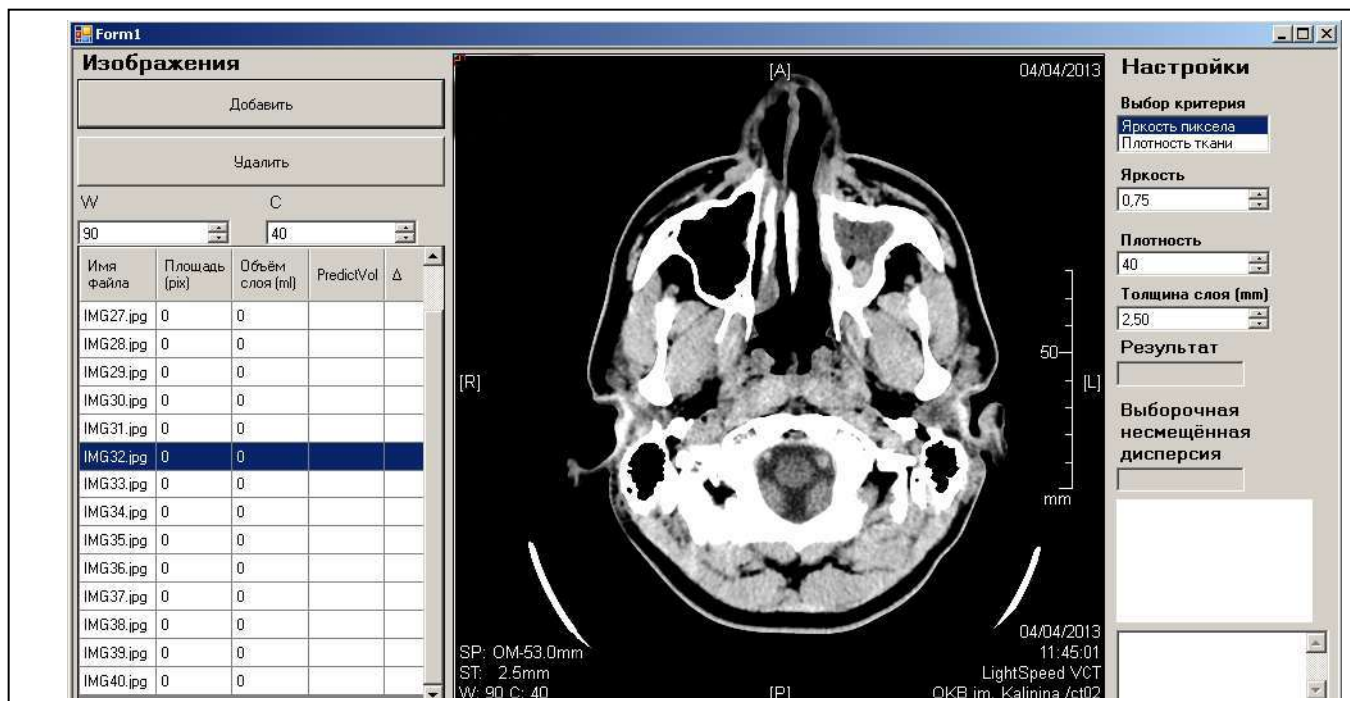


Fig. 2. Developed software graphic user interface with CT layers loaded.

At first, object related pixel subset contains only one starting pixel specified by user by clicking on the CT layer. Then every other pixel adjacent to object region is compared with object belonging criterion. Pixel is added to object related pixel set if it fits the criterion, otherwise it is marked as an object region border pixel. These pixels don't extend the object related region, but they can be useful later.



In case the object border on a separate layer of a picture isn't closed, and it needs to be completed, the algorithm performs the following operations. Set of existing border pixels allows doing an object border restoration. Although algorithm puts a restriction for the starting layer to have a closed borderline. Pixel search process on the starting layer ends when all object's adjacent pixels are checked and there is no possibility to extend studying object region, then algorithm performs possible extensions of object region to adjacent layers while it's possible. Every new layer should be processed the same way as first layer, but starting pixel set is defined by object region on previous layer. If adjacent layer hasn't got any pixels that fit the criterion, region extension in this direction stops. Segmentation algorithm completes when there are no further possibilities for region extension.

CT layers without closed object region borderline require additional borderline finding, which allows to strictly identify which part of layer is refers to studying object. Before examining next layer, algorithm performs extrapolation for area value prediction. If current region area significantly exceeds predicted value, it's borderline got unclosed character and region extension to this layer cancels. Set of pixels referring to external borderline has a shape of curved line. Algorithm searches for center of borderline, which divides its pixels to two subsets, then checks all pixel combinations from first and second subset. Borderline closes with line built through pair of points, which have highest  $w$  value, which shall be determined from the following equation:

$$w = \frac{l^\alpha}{d^\beta},$$

where  $l$  – length of borderline piece between two points,  $d$  – distance between those points.

$\alpha$  и  $\beta$  rates are needed to receive a few possible optional border closings and to choose the one that leads to area value which better fits the prediction. Fig. 3 shows fragment of initial CT layer (left) and borderline detection and closing (right). Squares mark the center of borderline and points, which has been chosen for borderline closing.

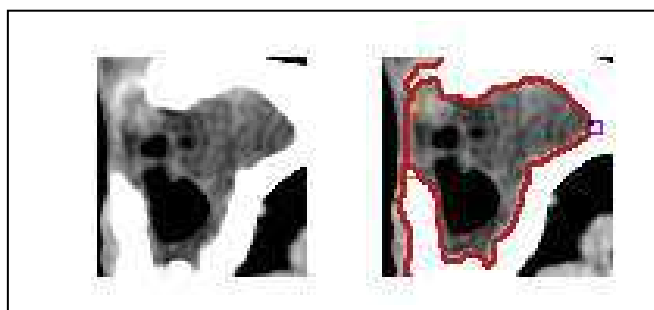


Fig. 3. CT layer example (left) with automated borderline closing (right).

#### 4. Results

Result of described CT layer segmentation algorithm is set of pixels related to studying object, which allows to do an analysis of this object and evaluate its parameters, such as volume, density distribution, density variance, layer areas and layer volumes. Software realization of developed algorithm allows browsing CT layers with marked object area and provides density distribution histogram. Studying object histogram can be used to identify its inner growths amount and character. Therefore it allows checking the possible pathological character of those growths.

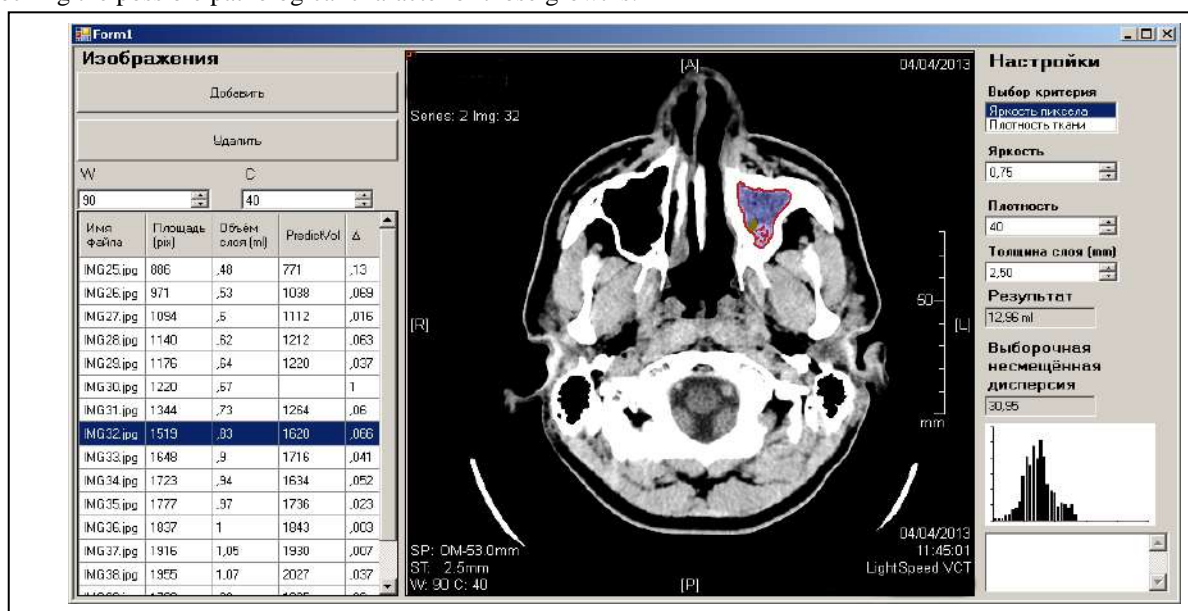


Fig. 4. Developed software user interface with object parameterization results.

Maxillary sinus study with application of developed algorithm has led to following results. Maxillary sinus total volume has been calculated as sum of its volume on each CT layer, and volume for each layer is product of object area value and layer

thickness. The volume value for studied maxillary sinus equals 12,96ml, and fig. 4 represents its analysis results. Developed algorithm and software application allows possible pathological growths visualization without use of contrast dye.

For example, fig. 4 shows CT layer with studied object area, with dark green color for object part which filled with air and shades of blue for tissues with density higher than density of air (which fills healthy maxillary sinus). Initial CT layers contain density information that allows doing this division after each layer segmentation. This kind of tissue localization can be alternative to contrast dye use.

Fig. 5 shows density distribution histograms of two maxillary sinuses. Left one mainly contains air, you can clearly see it by 0.92 value of its histogram left column, which corresponds to density less than lower boundary of Hounsfield range (i.e. air). Right one contains big amount of possibly pathological growths, so its histogram shows the densities distribution of those growths. Variance value of density can be used as distribution heterogeneity mark. For right maxillary sinus (marked at fig. 4) this value equals 30.95, while left one's equals 8.01. With a density distribution histogram, it illustrates heterogeneity degree of inner growths and demonstrates existence possibility of pathological growths.

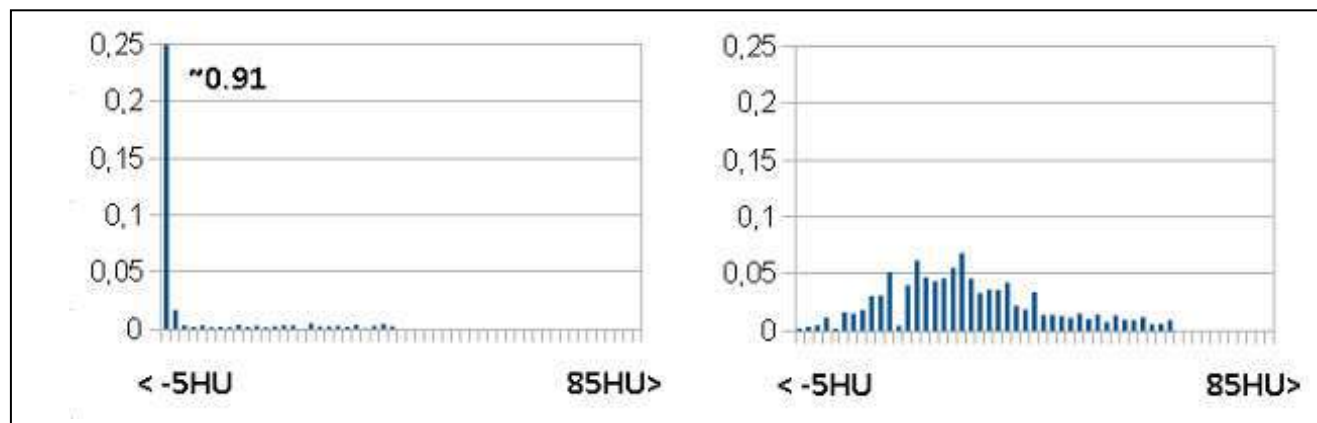


Fig. 5. Density distribution histogram of maxillary sinus (left - maxillary sinus with normal anatomical structure, filled with air, right - maxillary sinus with possible pathological growths).

## 5. Conclusion

Developed algorithm application leads to multi-layer CT automated analysis through object parameters calculation and region visualization. Additional region division allows finding specific types of tissues, detecting possible pathological tissue regions (the same way as it happens with contrast dye use). Density distribution histogram shows inner growth properties of studying object. Density distribution variation indicates inhomogeneity grade of its inner growth.

Automated parameters evaluation allows extending diagnostic data available to medics. Developed software passed approbation in the V.D. Seredavin Samara regional clinic and helps to improve diagnosis study of CT shots without use of contrast dye.

## References

- [1] Van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology* 2011; 261(3): 719–732.
- [2] Doi K. Current status and Future Potential of Computer-Aided Diagnosis in Medical Imaging. *The British Journal of Radiology*, 2005; 78: 3–19. DOI:10.1259/bjr/82933343.
- [3] Seletchi E, Duliu O. Image Processing and Data Analysis in Computed Tomography. *Romanian Journal of Physics*, 2007; 1: 764–774. URL: [http://www.nipne.ro/tjp/2007\\_52\\_5-6/0667\\_0667.pdf](http://www.nipne.ro/tjp/2007_52_5-6/0667_0667.pdf) (19.06.2015).
- [4] Bousson N, Fayad H, Le Pogam A, Pradier O, Visyikis D. Image Processing Methods in CT for Radiotherapy Applications. Theory and Applications of CT Imaging and Analysis. InTech 2011: 127–142. URL: <http://cdn.intechopen.com/pdfs-wm/14772.pdf> (20.06, 2015).
- [5] Maher M, Kalra M, Sahani D, Perumpillichira J, Rizzo S, Saini S, Mueller P. Techniques, Clinical Applications and Limitations of 3D Reconstruction in CT of the Abdomen. *Korean Journal of Radiology* 2004; 5(1): 55–67. DOI: 10.3348/kjr.2004.5.1.55.
- [6] Kim HC, Park SH, Park HC, Shin HC, Park SJ, Kim HH, Kim YT, Bae WK, Kim IY. Three-dimensional reconstructed images using multidetector computed tomography in evaluation of the biliary tract. *Abdominal Imaging* 2014; 29(4): 472–478. DOI: 10.1007/s00261-003-0123-x.
- [7] Xu F, Mueller K. Real-time 3D computed tomographic reconstruction using commodity graphics hardware. *Physics in Medicine and Biology* 2007; 52(12): 3405–3419. DOI:10.1088/0031-9155/52/12/006.
- [8] El-Baz A, Beache G, Gimel'farb G, Suzuki K, Okada K, Elnakib A, Soliman A, Abdollahi B. /Computer-aided diagnosis systems for lung cancer: challenges and methodologies. *International Journal of Biomedical Imaging* 2013; 2013: 1–46. DOI: 10.1155/2013/942353.
- [9] Chen H, Xu Y, Ma Y, Ma B. Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: Clinical evaluation. *Academic Radiology* 2010; 17(5): 595–602. DOI: 10.1016/j.acra.2009.12.009.
- [10] Montejo L, Jia J, Kim H, Netz U, Blaschke S, Müller G, Hielscher A. Computer-aided diagnosis of rheumatoid arthritis with optical tomography. Part1: feature extraction. *Journal of Biomedical Optics* 2013; 18(7): 123–137. DOI: 10.1117/1.JBO.18.7.076001.
- [11] Chen H, Wang XH, Ma DQ, Ma BR. Neural network-based computer-aided diagnosis in distinguishing malignant from benign solitary pulmonary nodules by computed tomography. *Chinese Medical Journal* 2007; 120(14): 1211–1215. URL: [http://124.205.33.103:81/ch/reader/create\\_pdf.aspx?file\\_no=200771851284230&year\\_id=2007&quarter\\_id=14&falg=1](http://124.205.33.103:81/ch/reader/create_pdf.aspx?file_no=200771851284230&year_id=2007&quarter_id=14&falg=1) (20.06.2015).
- [12] Hofer M. CT Teaching manual: a systematic approach to CT reading. Thieme 2007: 24–25.

# Color discrimination thresholds and the Einstein's field equations

L.D. Lozhkin<sup>1</sup>, O.V. Osipov<sup>1</sup>

<sup>1</sup>*Povolzhskiy State University of Telecommunications and Informatics, Lev Tolstoy street 23, 443010, Samara, Russia*

---

## Abstract

The paper discusses on the development of strictly equal-contrast color scale. It is known that in the CIE 1931 ( $x, y$ ) thresholds for color discrimination are represented by ellipses, which are used to characterize the equal-contrast of a color system. Different color systems are described by different values of ellipticity (the size of the MacAdam ellipses). The paper proposes a new approach to the creation of strictly equal-contrast color systems based on similar concepts of «the color horizon» in the colorimetry and «the event horizon» in general relativity (GR). It is proposed to use an equation similar to Einstein's field equations for the transformation of ellipses in a two-dimensional color pattern into circles and ellipsoids of rotation in three-dimensional model – in equal diameter balls. It is introduced a concept of a curvature of a color discrimination space and a tensor of a «color power». It is obtained a matrix equation from which the coefficients of transformation into strictly equal-contrast color system are determined.

*Keywords:* image processing; strictly equal-contrast color space; color locus; CIE colorimetric system; the metric tensor; curvature tensor; color tensor; basis of the moving frame

---

## 1. Introduction

Development of perceivable equal-contrast three-dimensional color scale would represent not only a great scientific achievement, but also proved to be helpful anyway. Its application would simplify the definition of colors and setting of color tolerances, bring clarity to the question of interpretation of the one-dimensional color scales for identification of some different colors, serve as a guide in the production of standard color patterns and assist in the selection of harmonious color combinations. Unfortunately, attempts to establish such a scale have not led to a significant success yet. On the contrary, they have confirmed the assumption that it is impossible to create this strictly equal-contrast three-dimensional scale. However, these attempts at least indicate there may be good enough approximation of an ideal equal-contrast color space. In this article there is will be continued the development of strongly equal-contrast color scales, and special attention will be given to the conclusion of numerical expressions for such scales.

If the observer is offered white, black and a set of gray color samples and asked to choose the one that equally differs from white and from black samples, he will face with a little difficulty because the assessment of the relative value of the two big color differences, eventually, based only on subjective impression. This is a special case of the color discrimination definition, which Newhall called by the method of color discrimination ratios [1]. However, the desired gray color can be determined based on the average assessment of several observers, the desired accuracy depends only on the number of observers and the number of assessments made by them. Then, the color range, which is between black and mid-gray, can be bisected, corresponding to you can do with an interval between white and mid-gray. Thus, the range from black to white forms equal-contrast lightness scale consisting of five equally spaced colors according to the subjective sensation. It was one of the methods used to determine the Munsellgray scale [2]. The method of converting the color locus, proposed in [3], is also of great interest.

## 2. The object of the study

In the early 40s of the last century it were appeared the publications of the results of experiments MacAdam carried out to ascertain the color discrimination thresholds [4]. Thresholds of color discrimination were graphically displayed in the form of ellipses on the color chart CIE 1931 ( $x, y$ ). Fig. 1 shows the results of MacAdam experiments which later became known as the MacAdam ellipses. In CIE 1931 ( $x, y$ ) color discrimination thresholds are shown by ellipses that can be used as a characteristic of the equal-contrast of a specific color system. It is sufficient to introduce the concept of color surface ellipticity defined as the ratio of the major to the minor axes of the ellipse. For various CIE colorimetric systems values of ellipticity are different [5].

Carry out a thought experiment. Reducing the brightness of the stimulus, at some point retinal cones with lower sensitivity, compared with sticks, switch off and there comes a scotopic vision in shades of gray, which means that the color discrimination threshold is increased and at zero brightness the color discrimination threshold is equal to infinity, exactly limited by color locus. It is obvious that an increase in the emission brightness to very high values leads to color sensitivity eyes will also decrease. This phenomenon can be explained by the fact that the collapse of iodopsin (photosensitive material of retinal cones) will be faster than his recovery. So the eye will be color-blind, i.e. the value of color discrimination threshold will increase with the brightness of radiation and at very high brightness the color discrimination threshold is equal to infinity. Graphically, this is shown in fig. 2b as a hyperboloid of one sheet.

First of all, introduce some concepts:

1. The color horizon (similar to «the event horizon» in general relativity) is a volume in which, from the point of view of an eye color discrimination, color is homogeneous.

2. Infinity is the area bounded by the color locus.

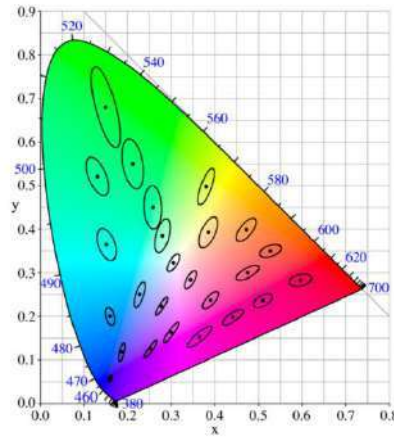


Fig. 1. The MacAdam ellipses (ellipses dimensions for clarity increased 10 times).

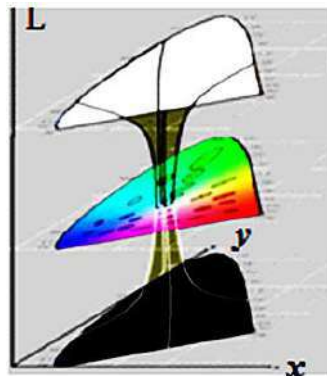


Fig. 2. Dependence of the color discrimination thresholds on the brightness.

Then consider some areas of the hyperboloid. The upper and lower portions, colored respectively in white and black colors, it can be said that the horizon of color extend to infinity. These surfaces are linear, flat and have a Euclidean geometry. A more complex structure is space located in the central part of the hyperboloid.

The horizon of color has a small radius and consequently the space enclosed by the horizon color is curved and closed, similarly as it occurs in the fundamental theory of stellar evolution. In connection with this it is possible to solve the problems of color discrimination thresholds as well as to create new equal-contrast color systems, in which all MacAdam ellipses would be transformed into equal circles and in three dimensions - in equal diameter balls, using Einstein's field equations [6].

### 3. Methods

Einstein field equations for relativistic gravitation looks as follows [6]:

$$[\hat{\mathbf{R}}_{ij}] - \frac{R}{2}[\hat{\mathbf{g}}_{ij}] + \Lambda[\hat{\mathbf{g}}_{ij}] = \frac{8\pi G}{c^4}[\hat{\mathbf{T}}_{ij}], \tag{1}$$

where  $[\hat{\mathbf{R}}_{ij}]$  is the Ricci curvature tensor obtained from the Riemannian curvature tensor  $[\hat{\mathbf{R}}_{ijkl}]$  by means of convolving with its pair of indices;  $R$  is the scalar curvature, i.e. the convolved Ricci curvature tensor;  $[\hat{\mathbf{g}}_{ij}]$  is the metric tensor;  $\Lambda$  is cosmological constant;  $[\hat{\mathbf{T}}_{ij}]$  is the stress-energy-momentum tensor;  $c$  is the speed of light in vacuum;  $G$  is the gravitational constant.

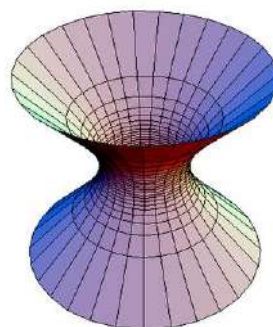


Fig. 3. Graphical interpretation of the Schwarzschild solution.

Einstein's equations do not impose any constraints on use to describe the «space-time» coordinates, i.e. have the property of general covariance and limit the choice of only 6 of the 10 independent components of a symmetric metric tensor. Therefore, their decision is ambiguous without introducing some constraints on the metric components, called the coordinate conditions [7]. Solving Einstein's equation (1) in conjunction with properly chosen coordinate conditions, you can find all the 10 independent components of the symmetric metric tensor.

**The metric tensor** makes it possible to determine the square of the interval in curved space that defines the distance in a metric space:

$$\delta S^2 = [\hat{\mathbf{g}}_{ij}](x) \delta x^a \delta x^b. \tag{2}$$

Consider separately the components of the equation (1). This equation assumes four-dimensional space-time, so it is considered its components in four-dimensional space. According to [8], in the so-called Schwarzschild coordinates  $t, r, \theta, \varphi$ , the last 3 of which are similar to spherical, the metric tensor is of the form:

$$[\hat{\mathbf{g}}_{ij}] = \begin{bmatrix} 1 - \frac{r_s}{r} & 0 & 0 & 0 \\ 0 & -\left(1 - \frac{r_s}{r}\right)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2 \theta \end{bmatrix}, \tag{3}$$

Where  $r_s$  is the Schwarzschild radius equal to the gravitational radius.

Expression (2) in this metric is written as follows:

$$\delta S^2 = \left(1 - r_s r^{-1}\right) c^2 \delta t^2 - \frac{\delta r^2}{1 - r_s r^{-1}} - r^2 \left(\sin^2 \theta \delta \varphi^2 + \delta \theta^2\right). \tag{4}$$

In the fig. 3 there is a graphical interpretation of the space by Schwarzschild. On the basis of similarity in fig. 2 and fig. 3, it can be concluded on the applicability of the foregoing mathematical apparatus for the construction of strictly equal-contrast color space.

The next element of the equation (1) is the stress-energy tensor, which in our case will be replaced in the future on the color energy tensor.

The curvature of color discrimination space. Now consider the concept of the curvature of color discrimination space, which can be described by the Ricci tensor. A scalar value can be built from it by the following formula:

$$R = \sum_{i=1}^2 \sum_{j=1}^2 [\hat{\mathbf{R}}_{ij}] [\hat{\mathbf{g}}_{ij}]. \tag{5}$$

The transition from the components of the Ricci tensor to the scalar curvature  $R$  is, at first sight, loss of information, i.e. nine variables are replaced by one. However, in the two-dimensional case, no data loss occurs. Indeed, the components of the curvature tensor are skew-symmetric both in the upper pair of indices and in the lower. In the paper there is proved that in the case of the spherical surface with radius  $r_0$  the scalar curvature is calculated as  $R = 2 / r_0$ .

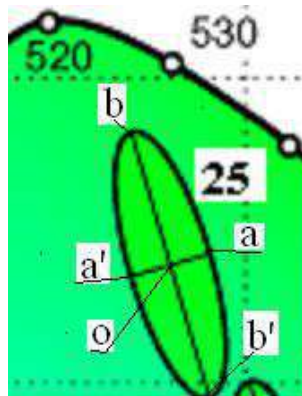


Fig. 4. Separate MacAdam ellipse.

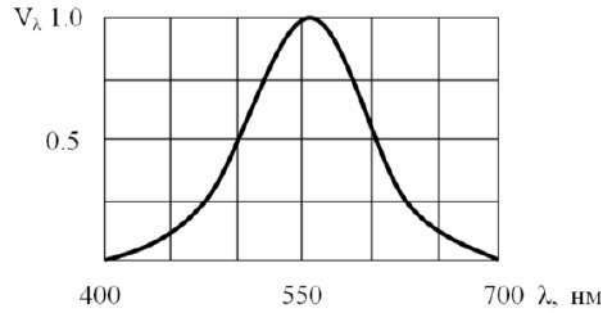


Fig. 5. Visibility curve.

**Color energy tensor.** Introduce the concept of color energy tensor  $[\hat{\mathbf{C}}_{ij}]$ . For new components of this tensor, turn to fig. 4. Evidently, the energy density in the stress-energy tensor in general relativity will correspond to the value of the MacAdam ellipse brightness density. According to MacAdam, the ellipse and in view of brightness – an ellipsoid is a threshold of color discrimination and brightness. Thus, from the viewpoint of the eye, this ellipsoid will be perceived as a geometrical point, there will be no color and brightness differences in the field (and also inside it). Therefore, the light energy density is equal to the brightness of the point, for example, in the center of the ellipse (fig. 4). Since these ellipses MacAdam received in his experiments (measuring color coordinates) in the first half of the last century, and in fact nowadays it is rather difficult to repeat these experiments, so the application has been developed, that allows determining the coordinates of any point and its brightness by the image of the ellipses on the color locus. To determine the brightness of these points we proceed as follows. Assume that the color of every point of the ellipse is created using monochrome emitters (two spectral colors). In calculating the brightness of the spectral colors a visibility curve can be taken (fig. 5) and in accordance with the spectral color wavelength of the color locus it can be determined the relative brightness sensation value on this curve.

Passing the math, due to their bulkiness, give the final expression for the color energy tensor:

$$[\hat{\mathbf{C}}_{ij}] = \begin{bmatrix} L_0 & L_d & L_a & L_b \\ L_d & (L_d - L_0)S^{-1} & (L_a - L_0)S^{-1} & (L_b - L_0)S^{-1} \\ L_a & (L_d - L_0)S^{-1} & (L_a - L_0)S^{-1} & (L_b - L_0)S^{-1} \\ L_b & (L_d - L_0)S^{-1} & (L_a - L_0)S^{-1} & (L_b - L_0)S^{-1} \end{bmatrix}, \quad (6)$$

where  $S$  is the ellipse area;  $L_0, L_a, L_b, L_d$  are the radiance of the ellipsoid center and the points on the ellipsoid surface corresponding to the main axes.

Also, introducing the amount of color discrimination threshold  $r_0$  to the elements of the metric tensor (3), get:

$$[\hat{\mathbf{g}}_{ij}] = \begin{bmatrix} 1 - \frac{r_0}{r} & 0 & 0 & 0 \\ 0 & -\left(1 - \frac{r_0}{r}\right)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2 \theta \end{bmatrix}. \quad (7)$$

Thus, the transition to strictly equal-contrast color system reduces to solving equations of the form similar to equation (1):

$$\frac{2}{r_0}[\hat{\mathbf{I}}] - \frac{1}{r_0}[\hat{\mathbf{g}}_{ij}] = [\hat{\mathbf{k}}_{ij}][\hat{\mathbf{C}}_{ij}], \quad (8)$$

where  $[\hat{\mathbf{k}}_{ij}]$  is the diagonal matrix of constants that are proportional to coefficients of frame «mobility» in the space of Riemann geometry;  $\hat{\mathbf{I}}$  is the unit diagonal matrix, 4x4 dimension; color energy tensor and the metric tensor are defined by equations (6) and (7).

Thus, the transition to a strictly equal-contrast color space system is reduced to the determination of the tensor elements  $[\hat{\mathbf{k}}_{ij}]$ . Of the article scope limitation there is omitted the math and written the expressions for the coefficients:

$$\begin{aligned} k_{11} &= \frac{(2 - r_0 b^2 d^2) \zeta_T}{r_0}; \\ k_{22} &= \frac{(2 - r_0^3 a^2 d^2) \zeta_H}{r_0}; \\ k_{33} &= \frac{[2 - r_0^3 a^2 b^2 (1 - d^2)] \sin^2 \theta L_S}{r_0}, \end{aligned} \quad (9)$$



where  $a, b, d$  are the ellipsoid semi-axes;  $\zeta_H$  is the color saturation;  $\zeta_T$  is the color hue;  $L_s$  is the brightness on the ellipsoid surface.

Thus, the relations (9), as measured the color saturation, the hue and the linear ellipsoid dimensions, allow the transition to a strictly equal-contrast color space.

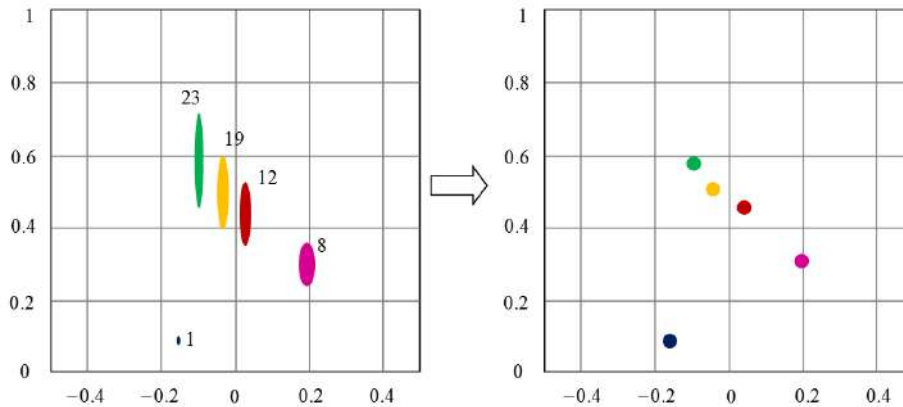


Fig. 6. Cross-section of ellipsoids and their transformation into equal size balls. The size of the ellipsoids and spheres is increased 10 times.

#### 4. Results and Discussion

In fig. 6 and fig. 7 there are presented results of the transformation using the relations (9), taking into account the above mathematics. This transformation was carried out for the CIE 1931 colorimetric system ( $x', y'$ ).

Obviously, using the above method it may be converted the color space of the Riemann space. For example, if it is depicted a sphere with radius equal to the maximum brightness (white color) for the CIE 1931 system ( $x', y'$ ), and the color locus is drawn on the sphere surface, keeping the transition from the metric flat space to the Riemann space, using the basis of the moving frame, you can get a curved color space, which can be depicted in the same basis of the moving frame of color discrimination space (MacAdam balls).

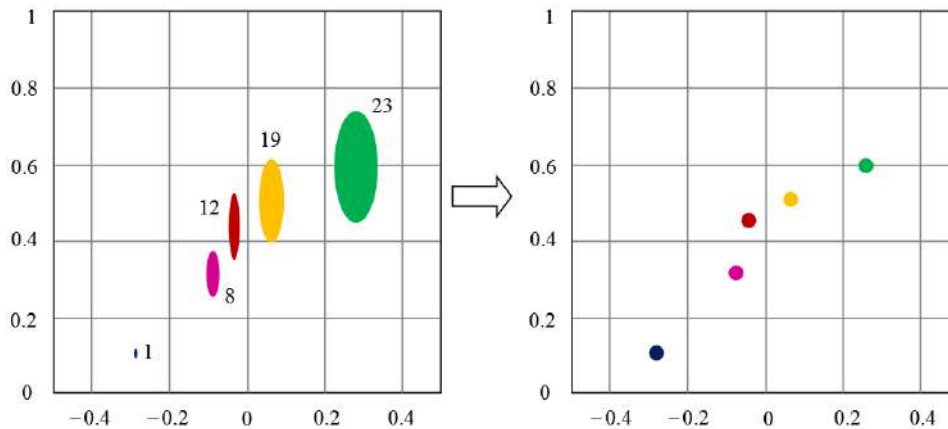


Fig. 7. Cross-section of ellipsoids and their transformation into equal size balls. The size of the ellipsoids and spheres is increased 10 times.

With this mathematical approach the input data can be presented in any of the existing CIE colorimetric systems such as CIE 1960 ( $u, v$ ) and (or) in CIE 1976 (Lab).

#### 5. Conclusion

In summary, there are formulated the main conclusions of the work.

1. It has been found the similarity between space-time state and the color space.
2. On the basis of this similarity for developing the strictly equal-contrast color space it were used the Einstein's field equations.
3. The solution of these equations was carried out in four-dimensional space, which used three-dimensional metric space ( $u, v, w$ ) in the CIE 1960 system and the fourth dimension was brightness, thereby creating the strictly equal-contrast four-dimensional color space. Similarly, the calculation for the three-dimensional color space was made.
4. It is shown that as the input color space can be used any of the known color spaces and it will be obtained the equal spheres of color discrimination thresholds at the output.
5. The resulting color body is a sphere which radius depends on the exact point representing the color.

6. The color difference between two colors resulting in the strictly equal-contrast color system is determined by the length of the arc, linking two colors that lie on different surfaces of concentric spheres.

## References

- [1] Newhall SM. The ratio method in the review of the Munsell colors. *Am. J. Psychol.* 1939; 52: 394 p.
- [2] Munsell A, Sloan T, Godlove I. Neutral value scales I. Munsell neutral value scale. *Opt. Soc. Am.* 1933; 23: 394 p.
- [3] Jimenez JR, Hita E, Romero J, Jimenez L. Scalar curvature of space as a source of information of new uniformity aspects concerning to color representation systems. *J. Optics* 1993; 24(6): 243–249.
- [4] Mac Adam DL. Visual sensitivities to color differences. *Josa* 1943; 33(18).
- [5] Mac Adam DL. Color essays. *Josa* 1975; 65(5): 463–485.
- [6] Pauli W. Theory of relativity. Ed. by Ginzburg VL and Frolova VP. M.: Nauka, 1991; 328. (in Russian)
- [7] Fock VA. Theory of space, time and gravity. M.: GITTL, 1955; 504 p. (in Russian)
- [8] Schwarzschild K. About the gravitational field of a mass point in the Einstein's theory. In: *Albert Einstein and the gravitational theory*. M.: Mir, 1979; 199–207. (in Russian)



# Method for identification of perlite-class steel microstructure parameters using metallographic images

R.G. Magdeev<sup>1</sup>, A.G. Tashlinskiy<sup>1</sup>

<sup>a</sup> Ulyanovsk State Technical University, Severnii Venets, 32, 432027, Ulyanovsk, Russia

---

## Abstract

A method for finding the microstructural parameters of low-carbon steel using its metallographic images is proposed. It allows to determine the following parameters: the ratio of perlite to ferrite phases, the parameters of grains of crystallites and their mutual arrangement; the degree of granularity of the pearlite phases. The method is aimed at predicting the strength characteristics of steel samples and consists of several stages. The preprocessing step involves color reduction, refinement of the area of interest, noise filtering, illumination refinement and histogram equalization. The segmentation of the image is associated with the search for the size, area and convex shell of grains. The stage of finding the microstructural parameters is based on the stochastic gradient-based estimation of the parameters of the segmented objects. Examples of analysis of steel oil pipeline samples with a forecast of their strength characteristics are given.

*Keywords:* metallographic image; digital image processing; stochastic gradient-based estimation; convex shell; steel microstructure parameters; grains of crystallites; perlite; ferrite

---

## 1. Introduction

Low-carbon (with a carbon content of less than 0.8%), low-alloy (less than 5% of alloying elements) steels, pearlite hypereutectoid steels are the main products of ferrous metallurgy. They are used for manufacturing a wide range of tools and parts with increased strength and elastic properties, pipelines, steel trusses, etc. [1,2].

One of the key problems in manufacturing and operation of steel products is the control of the compliance of these products with the required characteristics (strength, residual life, possibility of use under certain conditions, etc.), which is primarily ensured by the characteristics of the steel itself. The mechanical properties include hardness, strength, viscosity, elasticity, plasticity, etc. These properties are determined by chemical composition, macro- and microstructure, production and processing methods [3].

Steel microstructure is a combination of a large number of grains in the form of adjacent crystallites differing in size, shape, and spatial orientation. All microstructures of low-carbon and low-alloy steels contain a pearlite-eutectoid mechanical mixture of ferrite and cementite [2]. Pre-eutectoid steel has a lamellar structure consisting of alternating plates of ferrite and cementite [4]. An essential feature of the microstructure is the presence of internal boundaries separating the grains in the metal.

Usually, images of microstructures are obtained by means of a digital metallographic microscope at various magnifications [5]. In the images one can see different phases, outlines of the grains and their mutual arrangement. The images used in this work were obtained from steel pipelines for transferring petroleum products using the LOMO BIOLAM M-1 laboratory research microscope with a special nozzle and installed MC-3 digital camera with resolution 1200x900 pixels. Fig.1 shows an example of an image of the microstructure of a metal of a 17GS pipeline at 200x magnification.

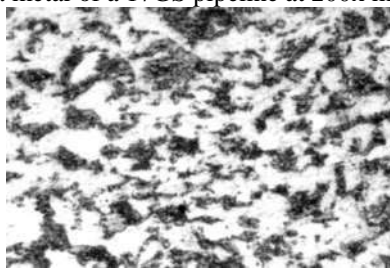


Fig. 1. Example of steel 17GS microstructure.

Metallographic methods for detecting and determining the grain size of steels and alloys are established by GOST 5639 [6]. These methods are: visual comparison of grains with scale templates, counting the number of grains per unit surface of the section, counting the intersections of grain boundaries by straight lines, measuring the length of the chords with the determination of the relative fraction of grains of a certain size, and the ultrasonic method based on the dependence of the attenuation of ultrasonic oscillations on grain sizes.

Numerous studies have shown the relationship between the parameters of the microstructure and the mechanical properties of metals and alloys [5, 7-10, etc.]. The dependence of the strength characteristics of cold-rolled steels at the production stage on the change in the microstructure was investigated in [11, 12, 13]. In practice, most microstructural studies are carried out visually by experts, which does not allow for an objective assessment of their reliability. Therefore, the development of a technique capable of automating the process of obtaining microstructural characteristics on the basis of digital image processing methods and algorithms seems to be an urgent task. Unfortunately, a small number of papers have been devoted to solutions of this problem, in particular [14, 15]. However, they do not consider the evaluation of the properties of steels based on the

microstructural characteristics found, which is in demand both at the production stage and after long-term operation. In addition to the basic microstructural characteristics, parameters such as the degree of ordering of orientations of the pearlite grains and the degree of granularity of the pearlite phases are also important for determining the degree of metal fatigue.

## 2. Basic microstructural parameters of steel

Shape and arrangement of grains in low-carbon and low-alloy steels are subjected to certain regularities associated with the solidification of the metal and its transformation during processing and operation. In accordance with GOST 8233 [11], the basic parameter is the ratio of the pearlite and ferritic phases of the microstructure of metal. For a ferrite-pearlite the ratio can vary between 0 and 95%. This range of ratios can be covered using templates by various image processing means.

One of the main geometric parameters of the microstructure of a metal is the grain size, which is its mean diameter. The size  $d_i$  of the  $i$ -th grain is characterized by the arithmetic average of the longitudinal (maximum) size  $W_i$  and the transverse (minimum) grain size  $H_i$ . The average size  $\bar{d}$  is found as the average of all the changed grains.

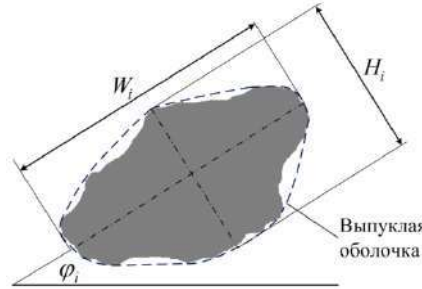


Fig. 2. Metal grain geometric parameters.

A correlation was found between the average grain size  $\bar{d}$  and the yield and strength limits, which are described by the Hall-Petch dependence [2]:  $\sigma_S = \sigma_{S_0} + g\sqrt{\bar{d}}$ , where  $\sigma_S$  - the yield strength;  $\sigma_{S_0}$  - the yield strength of the initial (in the production of steel);  $g$  - constant coefficient determined by the steel grade.

Large dispersion of grain sizes adversely affects the uniformity of mechanical and operational properties of products. To take into account this circumstance, the spread parameter of the grains is used [2]. It was shown in [5] that the grain size is a random variable that has a normal distribution law, accordingly the spread of grain sizes corresponds to the mean square deviation of the Gaussian distribution.

Over time, with sufficient external forces, the plastic deformation covers the entire volume of the polycrystalline. As a result, the grains get an elongated shape in the direction of the most intense flow of the metal. Simultaneously with the change in the shape of the grains during the deformation, the crystallographic axes of the individual grains rotate in the direction of the greatest deformation, which leads to anisotropy of the properties of the metal.

To control the microstructural characteristics at the stages of metal rolling, the Tretyakov method [12] based on empirical formulas for determining the mechanical characteristics of steels and alloys depending on the degree of deformation is used. In particular, for cold rolling, the conditional yield strength  $\sigma_{CS}$  is calculated by the formula:  $\sigma_{CS} = \sigma_{CS_0} + A\varepsilon^b$ , where  $\sigma_{CS_0}$  is the conditional yield strength in the initial state;  $\varepsilon$  - degree of metal deformation, %;  $A$  and  $b$  are constant coefficients determined by the steel grade. However, the above relation does not take into account the anisotropic properties of microstructures, for which the average coefficients of anisotropy of the grain shape are:  $\bar{k}_{begin} = \bar{W}_{begin}/\bar{H}_{begin}$  before and  $\bar{k}_{end} = \bar{W}_{end}/\bar{H}_{end} = (\bar{k}_{begin} + \varepsilon)/(1 - \varepsilon)$  after deformation.

The degree of orderliness of orientations of the pearlite grains in the investigated region of the steel microstructure is characterized by the directivity vector and the ordering coefficient. For a particular grain, the vector of orientation  $\vec{K}_i^{dir} = d_i \cdot \exp(-j\phi_i)$ , i.e. the direction of the vector coincides with the direction of the longitudinal axis of the pearlite grain (fig. 2). The general vector of the grain orientation is:  $\vec{K}_2^{dir} = \sum_{i=1}^n \vec{K}_i^{dir}$ , and the ordering coefficient is:  $k_{ord} = |\vec{K}_2^{dir}|/(n\bar{d})$ , where  $n$  - the number of grains is.

Other important characteristics of the metal microstructure affecting the mechanical properties of steel are the ratio of pearlite to ferrite  $k_{PF}$ , %, and the degree of granularity of the pearlite phases  $k_{grit}$ , %.

## 3. Stages of metallographic image processing method

The proposed technique for identifying microstructure parameters can be divided into three main stages.

*Preliminary processing of images under study aimed at increasing the accuracy and reliability of finding the microstructural parameters of pearlite grains.* It consists of the following operations: color reduction of the image in order to simplify subsequent calculations area of interest extraction is aimed at excluding low-information areas of the image, filtering the image to compensate for high-frequency distortions caused by the peculiarities of the metallographic microscope path, brightness refinement compensating for uneven illumination of the microsection, and histogram equalization.

*Segmentation on metallographic images of areas corresponding to pearlite grains according to which their microstructural parameters are further located.* It is achieved by the following procedures: segmentation, aimed at identifying areas of pearlite

grains, mathematical morphology for eliminating internal discontinuities in grain images and excluding from the further analysis of small objects, isolating external boundaries and constructing convex shells of grains for the subsequent calculation of microstructural parameters.

*Estimation of microstructural parameters of perlite grains*, including the formation of adaptive templates for finding object parameters, Gaussian filtration of convex shells of isolated grains and formed templates for the purpose of expanding the working range of stochastic gradient descent procedures [13, 14] used to estimate the microstructural parameters of grains, finding particular and integral microstructural parameters of perlite grains and the degree of granularity of pearlite phases.

Below the implementation of the above steps of the methodology is briefly considered.

### 3.1. Image preprocessing

*Image color reduction.* In the formulated problem, the informative component of color is small, so vector-based (color) images are advisable to translate into levels of gray [15]. It should be noted that there are color models of images in which the luminance component is already separated into a separate stream: HSV, HSL, YUV, etc. The YUV model in the recommendation ITU-R BT.601, whose brightness component is calculated according to the formula [16]:

$$z_{\text{grey}}(x, y) = 0.299z_r(x, y) + 0.587z_g(x, y) + 0.114z_b(x, y),$$

where  $z_r(x, y)$ ,  $z_g(x, y)$ ,  $z_b(x, y)$  – values of the red, green, and blue components of the pixel with the coordinates  $(x, y)$ ;  $z_{\text{grey}}(x, y)$  pixel brightness value obtained as a result of monochromization. This procedure is performed for all sample counts.

*Area of interest extraction* is aimed at excluding from the further processing of low-information areas, the presence of which is related to the specificity of the imaging by a metallographic microscope, which causes the image to be formed approximately in the form of a circle. The analysis [9, 17] showed that the brightness of the non-informative fragment differs significantly from the brightness of the informative fragment, and is usually 2.5-3% of the maximum brightness. With this in mind, the center and the radius of the circle are sought. Further processing of only the highly informative image area reduces the requirements for computational resources. In a particular implementation of the technique, after finding the image processing area for visual convenience, the brightness of the samples is inverted. Fig. 3,a shows examples of the selected processing area of two metallographic images.

*Image filtering* is aimed at elimination of brightness distortions caused by the imperfections of the optical detectors of the metallographic microscope. In images they appear as small (one - two pixels), but significant (up to 45% to neighboring pixels) brightness increase. The nature of the appearance of distortions is due to optics and reflections during photography. To eliminate their influence on the final result, nonlinear median filtration was used [19]. The size of the median filter window for metallographic images is usually 3x3 or 5x5.

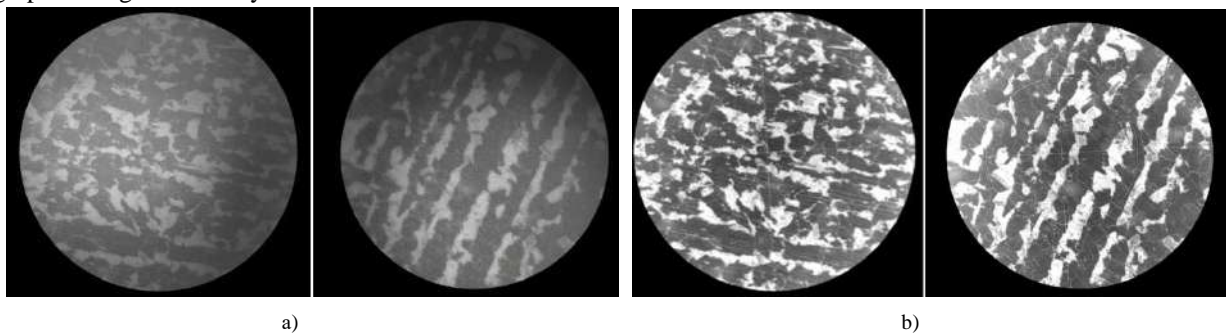


Fig. 3. Examples of two metallographic images: a) after area of interest extraction, б) after image equalization.

*Illumination refinement* is used for lighting unevenness compensation. A typical example of distortion of this kind in metallographic images is a shadow. In this case, the image  $Z$  can be represented as:  $Z = X \cdot \gamma$ , where  $X$  - an undistorted image,  $\gamma$  – illumination coefficient. One can obtain an approximate illumination map by applying a Gaussian filter with a large blur radius (about 5% of the highly informative image area). The restored image is looked for as:  $\hat{X} = \exp(\log(Z) - \log(\hat{\gamma}))$ , which allows not only to align the image with the level of illumination, but also to perform gradient transformations [22].

*Histogram equalization* aligns the intensities with the aim of improving the quality of the display. To carry out the equalization, the conversion is performed:  $z_{\text{ekv}}(x, y) = f(z(x, y))$ , where  $z(x, y)$  – brightness value in the pixel with coordinates  $(x, y)$  of the original image,  $z_{\text{ekv}}(x, y)$  – brightness value of the converted image, and  $f(z)$  is the single-valued monotonically increasing conversion function.

Fig. 3b shows examples of images after the preprocessing phase.

### 3.2. Pearlite grains segmentation

At this stage, the problem of isolating individual perlitic spots is solved with a view to their further analysis. To solve this problem, a growing areas method is used [16] with preliminary procedures for binary segmentation and mathematical morphology of the resulting binary image.

The problem of segmentation of the perlite spots is solved using the image binarization procedure based on the histogram analysis of the image, taking into account the sizes of the desired regions.

Morphological closing (in particular, with 5x5 kernel) is aimed at eliminating objects less than the specified window and filling the gaps that are inside the images of pearlite spots.

Individual perlite grain segmentation is based on the method of growing areas [19] as follows: on a binary image is a pixel belonging to perlite. If the neighboring pixel of the image also belongs to the perlite - the decision is made whether this pixel belongs to this spot, and it is appropriately marked. The procedure continues until all adjacent pixels are labeled or belong to ferrite.

Pearlite grain boundaries estimation. The boundaries of the selected objects are found via algorithms for the sequential construction of contours, which are characterized by high speed, absence of discontinuities and "extra" boundaries with low computational complexity. In particular, the recursive algorithm of the "beetle" was used [10, 20]. Its computational complexity is determined by two main components: the search for the first point of the object and the sequential search for objects. The advantage of the algorithm with respect to the problem under consideration is that it isolates only the outer boundary of the object, without separating the internal ones.

Construction of convex shells of perlite grains. There are many algorithms for extracting a convex hull, for example, the algorithm of Chan, Kirkpatrick, Melkman [21], but Graham [22], Jarvis [23] and the so-called "quickhull" (QH) algorithms were most widely used [24]. Their effectiveness has been investigated for the problem on binary images of simple figures and pearlite spots [25]. On simple figures, all algorithms showed an adequate result with a slight difference in speed. On binary images of real objects - pearl spots obtained from images of microstructures of metal pipelines, the Jarvis algorithm and the QH algorithm distinguish the convex envelopes of the spot correctly, unlike the Graham error algorithm. In this case, the average time of the Graham algorithm was approximately 1.1 times less than the QH algorithm and 1.9 times less than the Jarvis algorithm (the experiment was performed on PC with AMD Athlon II X2 3GHz and 3GB RAM). Therefore, the method included the QH algorithm. Note that the computational complexity of the Graham algorithm does not depend on the number of vertices found and is proportional to  $q \log(q)$ , where  $q$  is the number of external points of the polygon (spot). The complexity of the Jarvis algorithm depends on the number of vertex spots and is proportional to  $qh$ , where  $h$  is the number of common spot points and its convex hull, which in the worst case is  $q^2$ . The computational costs of the QH algorithm is compounded by the complexity of constructing all subsets. At best, the problem is divided into two equally powerful subtasks, then the complexity of the algorithm is from  $2q$  to  $q^2$ . The advantage of the QH algorithm is also the possibility of parallel computations for all subsets.

Examples of isolated convex shells of pearlite spots are shown in Fig. 4, a. After the convex hulls are selected, their linear characteristics are calculated, which are then used to estimate the microstructural parameters.

### 3.3. Estimation of microstructural parameters

In order to find the microstructural parameters of the spots, stochastic gradient descent-based estimation was used [14]. The point is that the parameters of templates are adaptive and adapt to the parameters of the spots represented by convex hulls. The initial approximations of the parameters of the templates are chosen taking into account the working range of the stochastic gradient descent optimization procedures. As an a priori information for finding the initial approximations of the parameters of the templates, the area of each selected convex hull in the image is used, which is related to the area of the desired ellipse by the obvious relation:  $S = \pi ab$ . Studies have shown that three initial approximations of ellipticity (semi-axis relations) are sufficient:  $c = (\sqrt{3})^{-1}$ , 1 and  $\sqrt{3}$ . In this case, taking into account that the ellipse at  $c = (\sqrt{3})^{-1}$  differs from the ellipse  $c = \sqrt{3}$  only by a rotation by  $90^\circ$ , we obtain as the initial approximations the circle ( $c_0=1$ ,  $\phi_0=0^\circ$ ) and two ellipses ( $c_0=1/3$ ,  $\phi_0=0^\circ$ ) and ( $c_0=1/3$ ,  $\phi_0=90^\circ$ ).

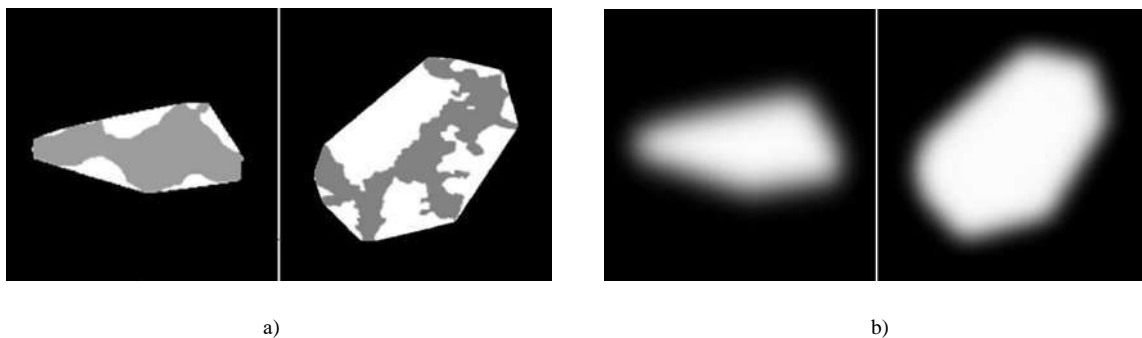


Fig. 4. Examples of convex shells of pearlitic spots and the results of their filtering.

In order to increase the working range of the evaluation, the templates and convex hulls obtained for each object under study are subjected to an approximate procedure [18] of Gaussian filtering with a filter radius of 15% of the linear dimension of the object. Examples of filtered convex shells of pearlite spots are shown in Fig. 4, b.



As a model of possible deformations of a customized template, when it is adjusted to a convex hull, a model similar to the similarity model is used:

$$x = x_0 + \kappa \left( (x - x_0) \cos \phi - k(y - y_0) \sin \phi \right) + h_x, \quad y = y_0 + \kappa \left( (x - x_0) \sin \phi + k(y - y_0) \cos \phi \right) + h,$$

where as the following adaptive parameters are used: scale factor  $\kappa$ , ellipticity coefficient  $k$ , parallel shift  $\vec{h} = (h_x, h_y)$ ,  $\phi$ - directional angle and the coordinates of the rotation center of the spot  $(x_0, y_0)$  [26]. It should be noted that to ensure the work of stochastic gradient descent optimization procedures, it is necessary to estimate all the coefficients of the model, and to calculate the microstructural parameters it is sufficient to use just  $k$  and  $\phi$ . For each  $i$ -th spot with respect to the parameters of the adapted template, which has the maximum correlation with its convex hull, the following parameters are calculated: longitudinal size  $W_i$ ; transverse size  $H_i$ ; average size  $d_i$ ; directional vector  $\vec{K}_i^{\text{dir}}$  and the form anisotropy coefficient  $k$  (which coincides in this technique with the ellipticity coefficient). For example, Fig. 5 shows the histograms of the distribution of spots by the coefficient of anisotropy  $k$  of the shape (fig. 5,a) and the directional angle (fig. 5,b). The left figure corresponds to the left microstructure of fig. 3,b, and the right one - the right. The selected type of histogram of grain directivity (from 0 to 180 degrees), due to the specific nature of the problem, makes it possible to identify the directions of growth of pearlite spots close to 0 (180) degrees.

Then the integral parameters of the grains are found: the number of grains, the ratio of pearlite to ferrite, the granularity of the pearlite phases, the general grain orientation vector, the average grain size, the grain size distribution, the degree of ordering of the grain orientations, and the mean value of the anisotropic shape coefficient. In particular, for the left image of fig. 3,b, we obtain:  $n=41$ ,  $k_{\text{PF}}=31,3\%$ ,  $k_{\text{grit}}=26,5\%$ ,  $\bar{d}=43$ ,  $\delta_d=37$ ,  $\bar{k}_{\text{end}}=0,44$ ,  $|\vec{K}_{\Sigma}^{\text{dir}}|=696$ ,  $\phi=168$ ,  $k_{\text{ord}}=0,39$ , and for the right one:  $n=30$ ,  $k_{\text{PF}}=32,7\%$ ,  $k_{\text{grit}}=28,5\%$ ,  $\bar{d}=34,2$ ,  $\delta_d=23$ ,  $\bar{k}_{\text{end}}=0,40$ ,  $|\vec{K}_{\Sigma}^{\text{dir}}|=718$ ,  $\phi=70$ ,  $k_{\text{ord}}=0,70$ . From the obtained strength characteristics, taking into account the data from [12], it can be concluded that the metal structure shown in fig. 3,b on the left is equivalent to cold rolling with a coefficient of deformation  $\varepsilon = 0,3$ , and in fig. 3,b on the right  $\varepsilon = 0,47$  with  $\varepsilon_{\text{crit}} = 0,5$ . Due to minor deviations from the factory parameters, the product with the microstructure shown in the left image in fig. 3b can be allowed for further operation, but with the microstructure shown in the right-hand image of fig. 3b, taking into account the grain parameters, orientation and anisotropy of the grain shape, requires an additional, more in-depth analysis of the metal.

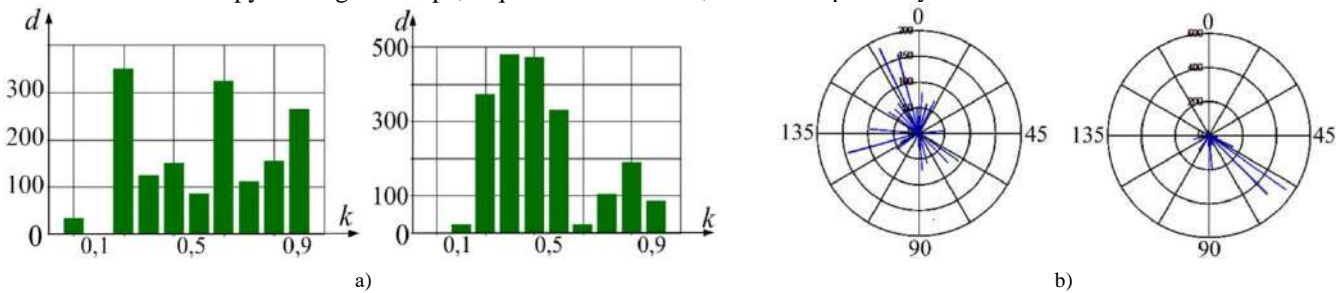


Fig. 5. Histograms of the shape anisotropy coefficient and the grain directivity angle.

It should be noted that the conducted studies showed that the average grain sizes, automatically found using the method examined, and calculated according to the traditional methods of GOST 5639 [6] differ by no more than 5%.

As already noted, the practical actual task is to determine the changes in the strength characteristics of metals after a long period of operation. A particularly topical task is to determine the changes in the strength characteristics of metals, after prolonged use. The processes of changing the microstructure of metal structures during long-term operation are similar to the processes of cold rolling metals with a certain degree of deformation. We will give an example of the change in the microstructural characteristics of the steel of the 12GS pipeline after 40 years of explantation with its outer (fig. 6,a) and internal (fig. 6,b) surfaces. Histograms of the directivity vectors of microstructures are also shown there.

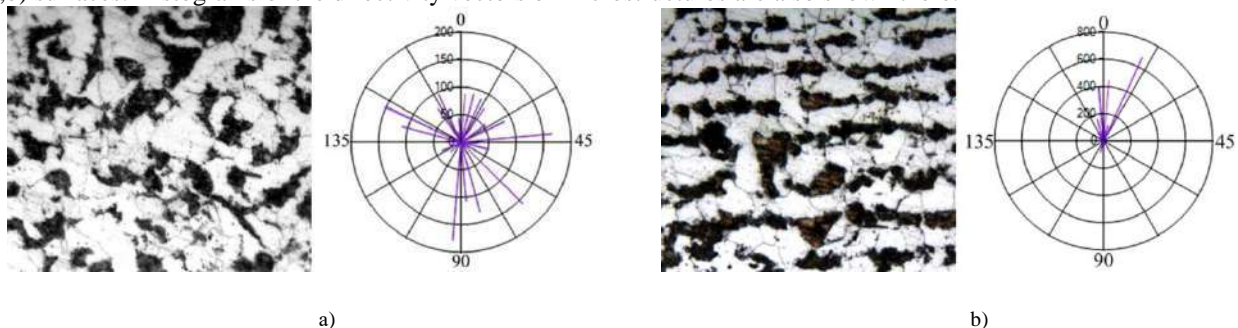


Fig. 6. The microstructure of 12GS steel after 40 years of operation and its histogram of directional vectors external (a) and internal (b) pipeline surfaces.

Analysis of the data shows that all the parameters of the microstructure, other than the directivity, as well as the strength characteristics of the outer and inner surfaces of the pipeline metal differ slightly. It is essential to differentiate the orderliness of the microstructure grains, which on the one hand leads to hardening, but on the other hand to the brittleness of the metal and the

increase in the probability of the appearance of microcracks. Therefore, the forecast of the strength characteristics of the pipeline metal must take into account the directivity of the grains of both the external and internal surfaces.

#### 4. Conclusion

A technique for determining the microstructural parameters of low-carbon steel is proposed. It allows to determine from metallographic images in real time the perlite to ferrite ratio, perlite grain, the general grain orientation vector, the average grain size, the grain size distribution, the degree of ordering of the grain orientations, and the mean value of the anisotropic shape coefficient.

The technique can be arbitrarily divided into three stages: preliminary processing of the images under study, aimed at increasing the accuracy and reliability of finding microstructural parameters, determining the areas corresponding to perlite grains on the images, and actually evaluating the microstructural parameters of the grains. Preprocessing includes operations: color reduction, selection of the working area of processing, image filtering to compensate for high-frequency distortion caused by the peculiarities of the metallographic microscope path, compensation of uneven illumination of the microsection, histogram equalization. Segmentation of perlite grains on the images is achieved by the following procedures: segmentation, aimed at identifying areas of pearlite grains, mathematical morphology for eliminating internal discontinuities in grain images and excluding from the further analysis of small objects, delineating outer boundaries and constructing convex shells of grains. Estimation of microstructural parameters includes the formation of adaptive templates for finding object parameters, Gaussian filtering of convex shells of segmented grains and formed templates for the purpose of expanding the working range of the required parameters, finding the microstructural parameters of the perlite grains and the degree of granularity of the pearlite phases.

The peculiarity of the technique is that the parameters of the templates are adaptive and adapt to the parameters of the spots represented by convex hulls. As an a priori information for finding the initial approximations of the parameters of the templates, the area of the selected convex envelope of the spot is used.

The proposed technique can be used to determine the strength characteristics of the metal at different stages of production and operation: from quality control at the plant to determining the remaining resource. Approbation of the technique on images of microsections of oil and water pipelines of different service life has shown its high efficiency.

#### Acknowledgement

The reported study was supported by RFBR and Government of Ulyanovsk region (Russia), project № 16-41-732053.

#### References

- [1] Garber EA. Cold rolling mills: (theory, equipment, technology). Cherepovets: CSU HPE, 2004; 416 p. (in Russian)
- [2] Arzamasov BN, Makarova VI, Muhin GG, Rizhov MN, Silaeva VI. Material Science: A Textbook for Universities. 8 th ed. A stereotype. Moscow: Publishing House of the Moscow State Technical University, 2008; 648 p. (in Russian)
- [3] GOST 380-2005 Carbon steel of ordinary quality. Stamps. Moscow: "Standardinform" Publisher, 2008; 5 p. (in Russian)
- [4] GOST 5640-68 Steel. Metallographic method for evaluating the microstructure of sheets and ribbons. Moscow: Publishing house of Standards, 1988; 17 p. (in Russian)
- [5] Starodubov DN. Methods and algorithms for processing and analysis of flaw detection and metallographic images: cand. tech. sciences thesis: 05.13.01. Vladimir, 2008; 183 p. (in Russian)
- [6] GOST 5639-82 Steels and alloys. Methods for the detection and determination of grain size. Moscow: IPK Publishing House of Standards, 2003; 20 p. (in Russian)
- [7] Gumerov AG, Zainullin RS, Yamaleev KM, Roslyakov AV. The aging of the pipes of oil pipelines. Moscow: "Nedra" Publisher, 1995; 218 p. (in Russian)
- [8] Malygin GA. Plasticity and Strength of Micro- and Nanocrystalline Materials (Review). Physics of the solid body 2007; 49(6): 961–982. (in Russian)
- [9] Vinogradova LA, Magdeev RG, Kurganova YuV. Algorithm for the determination of the ratio of phase shapes in the perlite of tubular steels with the structure of ferrite and perlite. RVM 2012; 6: 41–45. (in Russian)
- [10] Magdeev RG, Vinogradova LA, Dementiev VE. Method for determination of graininess in the pearlitic constituent of metals using imaging methods. Bulletin of Ulyanovsk State Technical University 2010; 4: 40–42. (in Russian)
- [11] Tretyakov AV, Trofimov GK, Guryanova MK. Mechanical properties of steels and alloys during plastic deformation. Pocket Guide. Moscow: "Mechanical Engineering", 1971; 63 p. (in Russian)
- [12] Vinogradov AI, Traino AI, Sarycheva IA. On the transformation of the grain structure of a metal during plastic deformation. Metals 2009; 2: 54–60. (in Russian)
- [13] Sarycheva IA. Method and algorithms for processing information for evaluating the mechanical characteristics of cold-rolled carbon steels: cand. tech. sciences thesis: 05.13.01. Cherepovets, 2012; 130 p. (in Russian)
- [14] Sadykov SS, Starodubov DN. Algorithms for determining the length and width of discrete area objects. Automation and modern technologies. Moscow: Mechanical Engineering, 2007; 10: 8–12. (in Russian)
- [15] NEXSYS ImageExpert™ Pro 3: Program for quantitative analysis of images. URL: [http://www.nexsys.ru/nexsys\\_iepro3x.htm](http://www.nexsys.ru/nexsys_iepro3x.htm) (14.02.2017).
- [16] GOST 8233-56 Steel. Standards of microstructure. Moscow: IPK Publishing House of Standards, 2004; 10 p. (in Russian)
- [17] Tashlinsky AG. Estimation of the parameters of spatial deformations of image sequences. Ulyanovsk: UISTU, 2000; 132 p. (in Russian)
- [18] Tashlinskii AG. Pseudogradient Estimation of Digital Images Interframe Geometrical Deformations. Vision Systems: Segmentation & Pattern Recognition. Vienna, Austria: I Tech Education and Publishing 2007: 465–494. DOI: 10.5772/4975.
- [19] Gonzalez RC, Woods RE. Digital Image Processing. 3rd Edition. New Jersey: Prentice-Hall Inc., 2006; 921 p.
- [20] Wizil'ter YuV, Zheltov SYu, Bondarenko AV, Ososkov MV, Morzhin AV. Processing and analysis of images in computer vision problems: A course of lectures and practical exercises. Moscow: "Fizmatkniga" Publisher, 2010; 672 p. (in Russian)

- [21] Dementiev VE, Magdeev RG, Dementiev EG. Use of image processing algorithms for inspection of steel pipelines. Survey of buildings and structures: problems and solutions: materials VI int. Scientific-practical. Conf., October 15-16 – St. Petersburg: Publishing house Polytechnic. University, 2015; 58–68. (in Russian)
- [22] Gruzman IS, Kirichuk VS, Kosykh VP, Peretyagin GI, Spektor AA. Digital image processing in information systems: Proc. Allowance. Novosibirsk: Publishing house of NSTU, 2002; 352 p. (in Russian)
- [23] Baatz M, Schäpe A. Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. Journal of Photogrammetry and Remote Sensing 2004; 58(3-4): 239–258.
- [24] Magdeev RG, Tashlinsky AG. Efficiency of object identification on binary images using pseudo-gradient adaptation procedures. Radio engineering 2014; 7: 96–102. (in Russian)
- [25] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 3rd Edition. Boston: MIT Press, 2009; 1312 p.
- [26] Williams DJ, Shas M. Edge Contours Using Multiple Scales. Computer Vision, Graphics and Image Processing 1990; 256–274. DOI: 10.1016/0734.
- [27] Graham RL. An efficient algorithm for determining the convex hull of a finite planar set. Information Processing Letters 1972; 1: 132–133. DOI: 10.1016/0020.
- [28] Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software 1996; 22(4): 469–483. DOI: 10.1145/235815.
- [29] Magdeev RG, Biktimirov LSh. Application of algorithms for constructing a convex hull in the analysis of images of a metal microstructure. News of the Samara Scientific Center of the Russian Academy of Sciences 2014; 16(2): 496–500. (in Russian)
- [30] Magdeev RG, Tashlinskii AG. A comparative analysis of the efficiency of the stochastic gradient approach to the identification of objects in binary images. Pattern recognition and image analysis 2014; 24(4): 535–541. DOI: 10.1134/S1054661814040130.

# A fast one dimensional total variation regularization algorithm

A. Makovetskii<sup>1</sup>, S. Voronin<sup>1</sup>, V. Kober<sup>1</sup>

<sup>1</sup>Chelyabinsk State University, ul. Bratiev Kashirinykh, 129, 454001, Chelyabinsk, Russia

## Abstract

Denosing has numerous applications in communications, control, machine learning, and many other fields of engineering and science. A common way to solve the problem utilizes the total variation (TV) regularization. Many efficient numerical algorithms have been developed for solving the TV regularization problem. Condat described a fast direct algorithm to compute the processed 1D signal. In this paper, we propose a variant of the Condat's algorithm based on the direct 1D TV regularization problem. The usage of the Condat's method with the taut string approach leads to a clear geometric description of the extremal function.

*Keywords:* Image restoration; total variation; denoising; exact solutions

## 1. Introduction

One of the most known techniques for denosing of noisy signals and images was proposed by Rudin, Osher, and Fatemi [1]. This is a total variation (TV) regularization problem. Let  $J(u)$  be the following functional in the functional space  $L_2$ :

$$J(u) = \|u - u_0\|_{L_2}^2 + \lambda TV(u), \quad (1)$$

where  $\|u - u_0\|_{L_2}^2$  is called a fidelity term and  $\lambda TV(u)$  is called a regularization term. Here  $u_0$  is an observed signal that is distorted by additive noise  $n$ ,

$$u_0 = v + n. \quad (2)$$

Consider the following variational problem:

$$u_* = \arg \min_{u \in BV(\Omega)} J(u). \quad (3)$$

where  $u_*$  is an extremal function for  $J(u)$ . Numerical results have shown that TV regularization is quite useful in image restoration [2-4]. Here we consider a one dimensional TV (1D TV) regularization problem. In [5,6] Strong and Chan considered the behavior of explicit solutions to the 1D TV problem when the parameter  $\lambda$  in Eq. (1) is sufficiently small. The exact solutions to one dimensional TV regularization problem and to two dimensional radial symmetric TV regularization problem were considered in [7-10]. Recently, Condat [11,12] proposed explicit solutions to the 1D TV problem as well as a direct fast algorithm for the case of discrete functions. The algorithm is very fast and has complexity of  $O(n)$  for typical discrete functions. In contrast, the proposed approach for finding exact solutions has a clear geometrical meaning.

In this paper, we propose a variant of the Condat's algorithm based on the direct 1D TV regularization problem. The usage of the Condat's method with the taut string method [12] leads to a clear geometric description of the extremal function.

## 2. Formulation of 1D TV regularization as a discrete problem

Let  $u_0$  be a discrete function  $u_0 = \{u_0^1, \dots, u_0^n\}$ . For the function  $u_0$  the problem in Eq. (1) takes following form:

$$J(u) = \sum_{i=1}^n (u^i - u_0^i)^2 + \lambda \sum_{i=1}^{n-1} |u^{i+1} - u^i|. \quad (4)$$

The functional  $J(u)$  is convex. Thus for the extremal (minimum) function  $u_*$  the subgradient  $\nabla J(u)$  satisfies the condition:

$$0 \in \nabla J(u_*). \quad (5)$$

**Remark.** The subgradient  $\nabla f(x)$  of the function  $f(x) = |x|$ :

$$\nabla f(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \\ [-1; 1], & \text{if } x = 0 \end{cases}. \quad (6)$$

### 2.1. Computation of the subgradient

Consider subgradient  $\nabla J(u)$ :

$$\nabla J(u) = \sum_{i=1}^n \nabla (u^i - u_0^i)^2 + \lambda \sum_{i=1}^{n-1} \nabla |u^{i+1} - u^i|. \quad (7)$$

$$\sum_{i=1}^n \nabla (u^i - u_0^i)^2 = (u^1 - u_0^1, u^2 - u_0^2, \dots, u^{n-1} - u_0^{n-1}, u^n - u_0^n). \quad (8)$$

In a similar manner with Eq. (6) the subgradients  $\nabla |u^{i+1} - u^i|$ ,  $i = 1, \dots, n-1$ , can be written as



$$\nabla|u^2 - u^1| = \begin{cases} (-1, 1, 0, 0, 0, \dots, 0, 0), & \text{if } u^2 > u^1 \\ (1, -1, 0, 0, 0, \dots, 0, 0), & \text{if } u^2 < u^1 \\ \{(\delta^1, -\delta^1, 0, 0, 0, \dots, 0, 0) | \delta^1 \in [-1; 1]\}, & \text{if } u^2 = u^1 \end{cases}, \quad (9)$$

$$\nabla|u^3 - u^2| = \begin{cases} (0, -1, 1, 0, 0, \dots, 0, 0), & \text{if } u^3 > u^2 \\ (0, 1, -1, 0, 0, \dots, 0, 0), & \text{if } u^3 < u^2 \\ \{(0, \delta^2, -\delta^2, 0, 0, \dots, 0, 0) | \delta^2 \in [-1; 1]\}, & \text{if } u^3 = u^2 \end{cases}, \quad (10)$$

...

$$\nabla|u^{n-1} - u^{n-2}| = \begin{cases} (0, 0, 0, 0, 0, \dots, -1, 1, 0), & \text{if } u^{n-1} > u^{n-2} \\ (0, 0, 0, 0, 0, \dots, 1, -1, 0), & \text{if } u^{n-1} < u^{n-2} \\ \{(0, 0, 0, 0, 0, \dots, \delta^{n-2}, -\delta^{n-2}, 0) | \delta^{n-2} \in [-1; 1]\}, & \text{if } u^{n-1} = u^{n-2} \end{cases}, \quad (11)$$

$$\nabla|u^n - u^{n-1}| = \begin{cases} (0, 0, 0, 0, 0, \dots, 0, -1, 1), & \text{if } u^n > u^{n-1} \\ (0, 0, 0, 0, 0, \dots, 0, 1, -1), & \text{if } u^n < u^{n-1} \\ \{(0, 0, 0, 0, 0, \dots, 0, \delta^{n-1}, -\delta^{n-1}) | \delta^{n-1} \in [-1; 1]\}, & \text{if } u^n = u^{n-1} \end{cases}, \quad (12)$$

$$\sum_{i=1}^{n-1} \nabla|u^{i+1} - u^i| = \{(\delta^1, \delta^2 - \delta^1, \delta^3 - \delta^2, \delta^4 - \delta^3, \dots, \delta^{n-1} - \delta^{n-2}, -\delta^{n-1}) | \delta^i = -1, \text{if } u^{i+1} > u^i, \delta^i = 1, \text{if } u^{i+1} < u^i, \delta^i \in [-1; 1], \text{if } u^{i+1} = u^i, i = 1, \dots, n-1\}. \quad (13)$$

From expressions (8) and (13) we get the following parameterization of the subgradient:

$$\begin{cases} (\nabla J(u))^1 = (u^1 - u_0^1) + \lambda \delta^1 \\ (\nabla J(u))^2 = (u^2 - u_0^2) + \lambda \delta^2 - \lambda \delta^1 \\ (\nabla J(u))^3 = (u^3 - u_0^3) + \lambda \delta^3 - \lambda \delta^2 \\ \dots \\ (\nabla J(u))^{n-1} = (u^{n-1} - u_0^{n-1}) + \lambda \delta^{n-1} - \lambda \delta^{n-2} \\ (\nabla J(u))^n = (u^n - u_0^n) + \lambda \delta^{n-1} \end{cases}. \quad (14)$$

where

$$\delta^i = \begin{cases} -1, & \text{if } u^{i+1} > u^i \\ 1, & \text{if } u^{i+1} < u^i \\ \in [-1; 1], & \text{if } u^{i+1} = u^i \end{cases}. \quad (15)$$

Since  $(\nabla J(u_*))^i = 0, i = 1, \dots, n-1$  for some values of the parameters  $\delta^i$  satisfying Eq. (15) we get:

$$\begin{cases} u_*^1 = u_0^1 - \lambda \delta^1 \\ u_*^2 = u_0^2 - \lambda \delta^2 + \lambda \delta^1 \\ u_*^3 = u_0^3 - \lambda \delta^3 + \lambda \delta^2 \\ \dots \\ u_*^{n-1} = u_0^{n-1} - \lambda \delta^{n-1} + \lambda \delta^{n-2} \\ u_*^n = u_0^n + \lambda \delta^{n-1} \end{cases}. \quad (16)$$

Consider the sequence of the cumulative sums:

$$\begin{cases} u_*^1 = u_0^1 - \lambda \delta^1 \\ u_*^2 + u_*^1 = u_0^2 + u_0^1 - \lambda \delta^2 \\ u_*^3 + u_*^2 + u_*^1 = u_0^3 + u_0^2 + u_0^1 - \lambda \delta^3 \\ \dots \\ u_*^{n-1} + \dots + u_*^1 = u_0^{n-1} + \dots + u_0^1 - \lambda \delta^{n-1} \\ u_*^n + \dots + u_*^1 = u_0^n + \dots + u_0^1 \end{cases}. \quad (17)$$

Consider such variables  $U^1, \dots, U^n$  and  $U_0^1, \dots, U_0^n$ , that

$$\begin{cases} U^1 = u_*^1, U_0^1 = u_0^1 \\ U^2 = u_*^2 + u_*^1, U_0^2 = u_0^2 + u_0^1 \\ \dots \\ U^{n-1} = u_*^{n-1} + \dots + u_*^1, U_0^{n-1} = u_0^{n-1} + \dots + u_0^1 \\ U^n = u_*^n + \dots + u_*^1, U_0^n = u_0^n + \dots + u_0^1 \end{cases}. \quad (18)$$

So the solution to the problem in Eq. (3) is reduced to the solution of the problem:

$$\begin{cases} U^1 = U_0^1 - \lambda\delta^1 \\ U^2 = U_0^2 - \lambda\delta^2 \\ U^3 = U_0^3 - \lambda\delta^3 \\ \dots \\ U^{n-1} = U_0^{n-1} - \lambda\delta^{n-1} \\ U^n = U_0^n \end{cases}, \quad (19)$$

with given discrete function  $U_0$  and unknown discrete functions  $U$  and  $\delta$  satisfying to the conditions in Eq. (15).

Consider additional variables  $U^0 = U_0^0 = 0$ . Note that then for any  $i = 1, \dots, n - 1$  the condition  $u^{i+1} > u^i$  is equivalent to the condition  $U^{i+1} - 2U^i + U^{i-1} > 0$ , the condition  $u^{i+1} < u^i$  is equivalent to the condition  $U^{i+1} - 2U^i + U^{i-1} < 0$ , the condition  $u^{i+1} = u^i$  is equivalent to the condition  $U^{i+1} - 2U^i + U^{i-1} = 0$ .

Then the set of equations in Eq. (19) can be rewritten taking into account additional variables:

$$\begin{cases} U^0 = U_0^0 = 0 \\ U^1 = U_0^1 - \lambda\delta^1 \\ U^2 = U_0^2 - \lambda\delta^2 \\ U^3 = U_0^3 - \lambda\delta^3 \\ \dots \\ U^{n-1} = U_0^{n-1} - \lambda\delta^{n-1} \\ U^n = U_0^n \end{cases}, \quad (20)$$

where

$$\delta^i = \begin{cases} -1, \text{ if } U^{i+1} - 2U^i + U^{i-1} > 0 \\ 1, \text{ if } U^{i+1} - 2U^i + U^{i-1} < 0 \\ \in [-1; 1], \text{ if } U^{i+1} - 2U^i + U^{i-1} = 0 \end{cases}. \quad (21)$$

### 2.2. Construction the „tube”

The values  $U_0^0, U_0^1, \dots, U_0^n$  of the discrete function  $U_0$  defines a piecewise linear curve, which is an axial line of the tube. The values  $U_0^0, U_0^1 + \lambda, \dots, U_0^{n-1} + \lambda, U_0^n$  form the upper piecewise linear border of the tube, the values  $U_0^0, U_0^1 - \lambda, \dots, U_0^{n-1} - \lambda, U_0^n$  form the bottom piecewise linear border of the tube. Figure 1 shows an example of a tube.

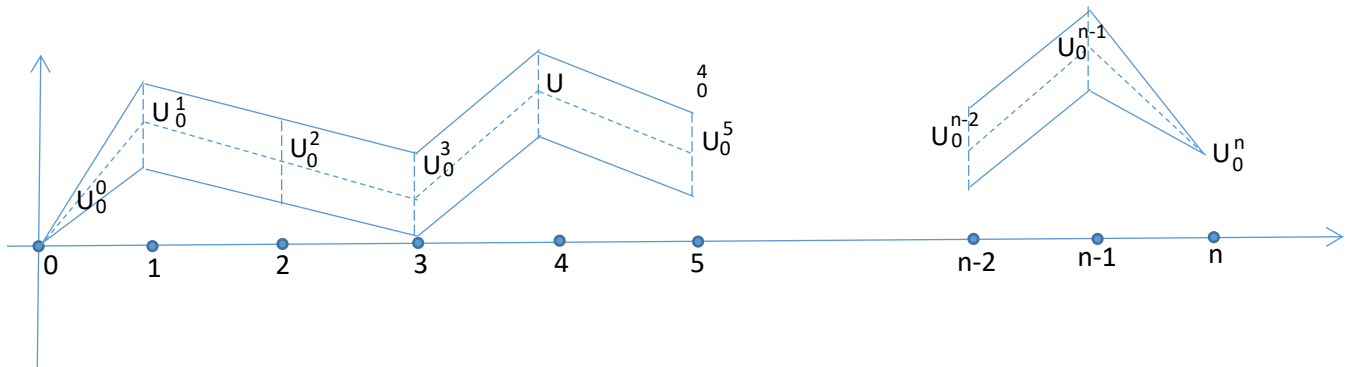


Fig. 1. Example of a tube.

### 2.3. Description of the extremal function $U$

Since  $\delta^i, i = 1, \dots, n - 1$ , take values in the segment  $[-1; 1]$ , a piecewise linear curve defined by the values  $U^1, \dots, U^n$  of a discrete function  $U$  (i.e. solution to the problem in Eq. (20)) entirely belongs to the tube.

If the second discrete derivative equals zero,  $U^{i+1} - 2U^i + U^{i-1} = 0$  then the piecewise linear curve defined by the values  $U^1, \dots, U^n$  of a discrete function  $U$  in the neighborhood of the point  $i$  is a straight line.

If the second discrete derivative is positive,  $U^{i+1} - 2U^i + U^{i-1} > 0$  then from Eq. (21) we see that  $\delta^i = -1$  and Eq. (20) shows us that  $U^i = U_0^i + \lambda$ , i.e.  $U^i$  belongs to the upper border of the tube.

If the second discrete derivative is negative,  $U^{i+1} - 2U^i + U^{i-1} < 0$  then from Eq. (21) we see that  $\delta^i = 1$  and Eq. (20) shows us that  $U^i = U_0^i - \lambda$ , i.e.  $U^i$  belongs to the lower border of the tube.

It means that a piecewise linear curve defined by the values  $U^0, \dots, U^n$  of a discrete function  $U$  exactly coincides with so called „taut string” connecting the endpoints of the tube.

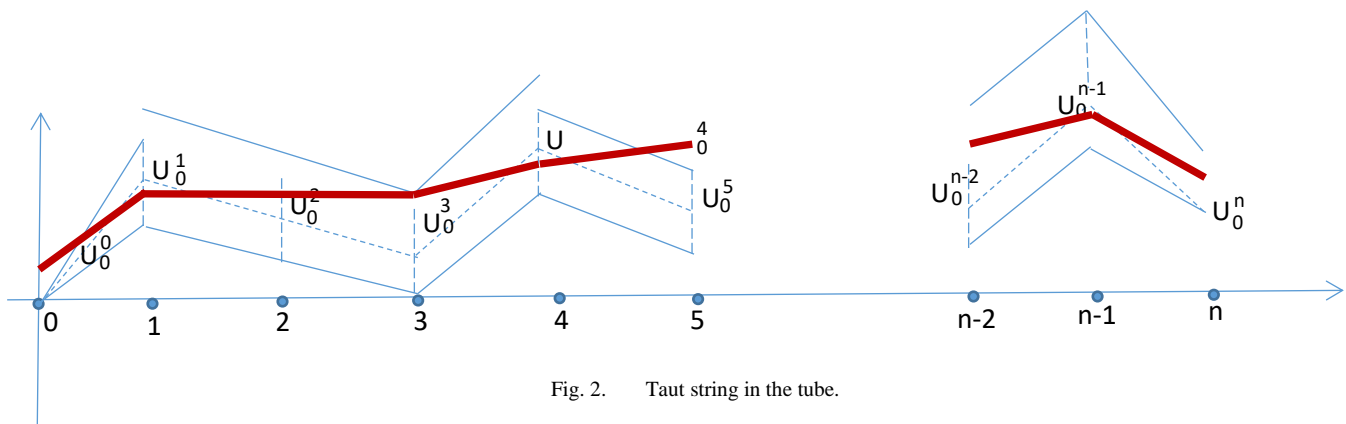


Fig. 2. Taut string in the tube.

## Conclusion

In this paper, we propose a variant of the Condat's method based on the direct 1D TV regularization problem. The usage of the Condat's method with the taut string method leads to a clear geometric description of the extremal function.

## Acknowledgements

The work was supported by Russian Science Foundation grant №15-19-10010.

## References

- [1] Rudin L, Osher S, Fatemi E. Nonlinear total variation based noised removal algorithms. *Phys. D* 1992; 60: 259–268.
- [2] Chambolle A, Lions PL. Image recovery via total variational minimization and related problems. *Numer. Math.* 1997; 76: 167–188.
- [3] Osher S, Burger M, Goldfarb D, Xu J, Yin W. An iterative regularization method for total variation based image restoration. *Multiscale Modelling and Simulation* 2005; 4: 460–489.
- [4] Chambolle A. An Algorithm for Total Variation Minimization and Applications. *Journal of Mathematical Imaging and Vision* 2004; 20: 89–97.
- [5] Strong DM, Chan TF. Exact Solutions to Total Variation Regularization Problems. *UCLA CAM Report*, 1996.
- [6] Strong DM, Chan TF. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems* 2003; 19: 165–187.
- [7] Voronin S, Makovetskii A, Kober V, Karnauhov V. Properties of exact solutions of the total variation regularization functions of one variable. *Journal of Communications Technology and Electronics* 2015; 60: 1356–1359.
- [8] Voronin S, Makovetskii A, Kober V. Explicit solutions of one-dimensional total variation problem. *Proc. SPIE's 60 Annual Meeting: Applications of Digital Image Processing XXXVIII* 2015; 9599: 959926–1.
- [9] Voronin S, Makovetskii A, Kober V. An efficient algorithm for total variation denoising. *Proc. Int. Conference of Analysis of Images, Social Networks, and Texts (AIST)* 2016; 236–248.
- [10] Voronin S, Makovetskii A, Kober V. Total variation regularization with bounded linear variations. *Proc. SPIE's 61 Annual Meeting: Applications of Digital Image Processing XXXIX* 2016; 9971: 99712T–9.
- [11] Condat L. A Direct Algorithm for 1-D Total Variation Denoising. *IEEE Signal Processing Letters* 2013; 20(11): 1054–1057.
- [12] Davies PL, Kovac A. Local extremes, runs, strings and multiresolution. *Ann. Statist.* 2001; 29(1): 1–65.

# Method of analysis of geomagnetic data based on wavelet transform and threshold functions

O. Mandrikova<sup>1,2</sup>, I. Solovev<sup>1,2</sup>, S. Khomutov<sup>1</sup>, K. Arora<sup>3</sup>, L. Manjula<sup>3</sup>, P. Chandrasekhar<sup>3</sup>

<sup>1</sup>Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS, 684034, Paratunka, Russia

<sup>2</sup>Kamchatka State Technical University, 683003, Petropavlovsk-Kamchatsky, Russia

<sup>3</sup>CSIR-National Geophysical Research Institute, 500007, Hyderabad, India

---

## Abstract

The suggested method is aimed at studying the dynamics of the magnetospheric current systems during magnetic storms. The method is based on algorithmic solutions for processing of geomagnetic field variations, detection of local increases in geomagnetic disturbance intensity and estimation of their dynamic characteristics. Parameters of the algorithms allow us to evaluate the characteristics of small-scale local features emerging during geomagnetic activity slight increases and large-scale variations observed during magnetic storms. To evaluate the method, geomagnetic data from the stations located in the north-east of Russia and equatorial India were used. The method testing showed the possibility to apply it for the detection of pre-storm anomalous effects in geomagnetic data.

*Keywords:* Wavelet transform; geomagnetic data processing; magnetic storm

---

## 1. Introduction

The work is devoted to the creation of methodical and software means for the analysis of recorded geomagnetic data. At present, theoretical and experimental bases of construction of systems for data processing and analysis are intensively developing, in particular in geophysics (for example, <http://www.cosmos.ru/magbase>; <http://matlab.izmiran.ru/magdata/>; <https://www.ngdc.noaa.gov/>; <http://smdc.sinp.msu.ru/>). It is caused by the increase of human society demands in automation of data flow processing. The subjects of this investigation are complex dynamic processes in the Earth magnetosphere and ionosphere determined by the phenomena and processes of solar origin. Solar activity impact on the Earth magnetosphere has quite a complicated character. Many aspects of it are still under-investigated [1]. As long as the state of magnetospheric-ionospheric system is an important factor of space weather which affects many aspects of our life, works in this area are of high scientific interest [1].

The Earth magnetic field variations reflect different geophysical processes in the Earth near space. During magnetic storms they contain uneven local features occurring at random times and carrying important information on the processes in the magnetosphere [2-14]. Traditional methods for data analysis applying basic models of time series, different techniques of smoothing and Fourier analysis methods are not effective enough to investigate fast unsteady processes. As it was noted in the papers [15-18], they do not allow one to identify thin local features characterizing short-period oscillations during increased geomagnetic activity and to estimate their dynamic characteristics before and during storms. At present, modern mathematic methods and technologies are intensively developing in this area [3-9, 14, 18-25]. Based on Data Mining application in order to improve the processes of geophysical data recording and organization of world data centers, methods for automation of expert work in this area have been developed (creation of so called "artificial experts") to solve geomagnetic data analysis problems, to detect noise at the stage of their preliminary processing, to identify anomalies during magnetic storms, to process magnetograms etc. [19-25]. To solve these tasks, the authors apply of the papers [9] a new approach, «discrete mathematical analysis» (DMA), which includes fundamental notions of mathematical analysis and modern approaches based on L. Zadeh's logic. A group of scientists from India (Kaleekkal Unnikrishnan) developed a technique for modeling of geomagnetic field variations for low-latitude stations. It is based on neural networks. The developed approach allowed them to improve the quality of the forecasting technique for magnetic storms (in 86% of cases) in comparison to the nearest method based on logistic regressive model obtained in 2005 (in 77% of cases) [23]. A group of scientists from Egypt (Space Weather Center, Faculty of Science, Helwan University, Cairo, Egypt) suggested to apply neural networks to predict the time of interplanetary shock wave propagation [24]. Based on the neural networks, the authors [25] suggested an algorithm for interplanetary magnetic field data processing and Dst-index calculation. The authors of that paper suggest an approach based on the combination of neural networks with wavelet transform and show the efficiency of joint application of mathematical apparatus data in comparison with a neural network in the problems of analysis of natural time series with complicated structures [26-29], in particular, for geomagnetic data analysis [28, 29]. It is shown in these papers that wavelet transform allows us to investigate the data structure in detail and to detect informative components which, in their turn, improve the procedure of neural network training and its performance efficiency. At present time, wavelet transform is widely used in the problems of analysis of the Earth magnetic field variations [2-5, 10, 14]. In the paper [2] wavelet transform is applied to investigate the relations between short-period oscillations of the geomagnetic field, solar wind parameters and interplanetary field during geomagnetic storms. Based on the wavelet transform, we solve such problems as denoising and elimination of a periodic component from geomagnetic field variations which is caused by the Earth rotation [4]. Applying the discrete wavelet transform, the authors of the paper [3] suggested an algorithm for automatic detection of magnetic storm initial stage periods. On the basis of the analysis of geomagnetic field variation wavelet spectrum, a method for forecast of strong geoeffective solar flares was proposed [13]. In

terms of wavelets, the authors of this paper suggested a new model of geomagnetic field variations [14, 30] and developed automatic algorithms do detect calm diurnal variation and to estimate disturbance intensities [30]. This approach allowed us to automate the procedure for calculation of geomagnetic activity index  $K$ , close to  $J$ . Bartels method, and to decrease the calculation error in comparison to the current methods [30]. In this paper we continue the investigation in this direction where a special emphasis is placed on the development of calculation solutions to detect and to estimate short-time anomalous increases in geomagnetic disturbance intensity which may occur before magnetic storms and have applied significance. The important thing in this approach is the possibility to apply the geomagnetic field data recorded on the ground, the analysis methods of which may significantly contribute to the current forecast methods. Taking into account incomplete prior knowledge on the dynamics of magnetospheric current systems and the limited scope of the obtained information on the processes in the near Earth space, noises, possible equipment failures etc., successful solution of the problem of space weather forecast requires a complex of methods and technologies. The confirmation of it is the large number of papers and scientific groups which aim their efforts at creating methods for recognition and classification of the effects in geophysical observation time series with applications in space weather problems.

## 2. Description of the method

In the papers [14, 30] the authors propose geomagnetic field variation representation based on multiscale wavelet decompositions:

$$f_0(t) = \sum_n c_{-6,n} \varphi_{-6,n}(t) + \sum_{j \in D} \sum_n d_{j,n} \Psi_{j,n}(t) + \sum_{j \notin D} \sum_n d_{j,n} \Psi_{j,n}(t) = f_{trend}(t) + f_{dist}(t) + e(t), \quad (1)$$

where  $\Psi_j = \{\Psi_{j,n}\}_{n \in \mathbb{Z}}$  is the wavelet-basis,  $\varphi_j = \{\varphi_{j,n}\}_{n \in \mathbb{Z}}$  is the basis, obtained from a scaling function, coefficients  $c_{j,n}$  and  $d_{j,n}$  are defined from the equations:  $c_{j,n} = \langle f, \varphi_{j,n} \rangle$ ,  $d_{j,n} = \langle f, \Psi_{j,n} \rangle$ ,  $D$  is a set of indices of the disturbed components,  $j$  is the scale, the inferior index «0» denotes that the initial discrete data belong to a domain of scale «0».

Component  $f_{trend}(t) = \sum_n c_{-6,n} \varphi_{-6,n}(t)$  describes the undisturbed level of the horizontal component of the Earth magnetic field during quiet geomagnetic field, and the component  $f_{dist}(t) = \sum_{j \in D} g_j(t)$  where  $g_j(t) = \sum_n d_{j,n} \Psi_{j,n}(t)$  describes perturbations, arising during increasing geomagnetic activity. Component  $e(t) = \sum_{j \notin D} \sum_n d_{j,n} \Psi_{j,n}(t)$  is the noise.

Minimizing the errors in the class of orthonormal functions, the Daubechies basis of order 3 was determined as the wavelet basis [30].

The set of indices  $D$  can be determined on the basis of the following criteria [14, 30]:

$$j \in D, \text{ if } m(A_j^v) > m(A_j^k) + \varepsilon, \quad (2)$$

where  $m$  is the sample mean,  $v$  is the index of disturbed field variation,  $k$  is the index of calm field variation, and  $\varepsilon$  is a positive number.

Assuming that  $A_j^v$  and  $A_j^k$  are normally distributed with mean  $\mu^v$ ,  $\mu^k$ ,  $\mu^v > \mu^k$  and variances  $\sigma^{2,v}$ ,  $\sigma^{2,k}$ , it is possible to estimate  $\varepsilon_j$  as  $\hat{\varepsilon}_j = x_{1-a/2} \frac{\sigma_j^k}{\sqrt{n^k}}$ , where  $\sigma_j^k$  is the variance of the greatest wavelet coefficients (for scale  $j$ ) for quiet days (this variance is determined as a result of multiple measurements);  $x_{1-a/2}$  is the  $1-a/2$  quintile of the standard normal distribution;  $n^k$  is the number of analyzed quiet-field variations. If  $a=0.1$ , the confidence probability is  $1-a/2 = 0.95$ , the quintile is  $x_{1-a/2} = 1.96$ , and  $\varepsilon_j = 1.96 \frac{\sigma_j}{\sqrt{n}}$ .

The measure of geomagnetic disturbance of the component  $g_j(t)$  on the scale  $j$  is [14, 30]:

$$A_j = \max_n (|d_{j,n}|). \quad (3)$$

Taking into account that the component  $f_{dist}(t) = \sum_{j \in D} g_j(t)$ , where  $g_j(t) = \sum_n d_{j,n} \Psi_{j,n}(t)$  describes the disturbances (see relation (1)), and the equivalence of discrete and continuous wavelet decompositions, in order to obtain more detailed information on the properties of the function  $f$  under analysis, continuous wavelet transform may be applied [31, 32]

$$(W_\Psi f)(b, a) := |a|^{-1/2} \int_{-\infty}^{\infty} f(t) \Psi\left(\frac{t-b}{a}\right) dt, \quad \Psi \text{ is the wavelet, } f \in L^2(\mathbb{R}), a, b \in \mathbb{R}, a \neq 0, \quad (4)$$

In this case, when a scale  $a$  vanishes, the wavelet coefficients  $(W_\Psi f)(b, a)$  characterize the local properties of the function  $f$  in the vicinity of the instant time  $t = b$  [31, 32].

Following the relation (3) as a measure of geomagnetic disturbance intensity, it is logical to consider the wave coefficient amplitude

$$i_{b,a} = |(W_\Psi f_{b,a})|.$$

The intensity of field multi-scale disturbances at an instant time  $t = b$  is estimated on the basis of the value [14, 30]

$$I_b = \sum_a (W_\Psi f_{b,a}). \quad (5)$$

In the case of field positive disturbances (current variation increase relatively the characteristic level),  $I_b$  value is positive. In the case of field negative disturbances (variation decrease relatively the characteristic level),  $I_b$  value is negative.

To distinguish the periods of increased geomagnetic activity, the following threshold function is applied:

$$P_{T_a}(W_\Psi f_{b,a}) = \begin{cases} W_\Psi f_{b,a}, & \text{если } (W_\Psi f_{b,a}) \geq T_a \\ 0, & \text{если } |W_\Psi f_{b,a}| < T_a \\ -W_\Psi f_{b,a}, & \text{если } (W_\Psi f_{b,a}) < -T_a \end{cases}, \quad (6)$$

where  $T_a = U * St_a^l$  is the threshold function where  $St_a^l = \sqrt{\frac{1}{l-1} \sum_{k=1}^l (W_\Psi f_{b,a} - \overline{W_\Psi f_{b,a}})^2}$ , is the standard deviation,  $l$  is the time window length,  $\overline{W_\Psi f_{b,a}}$  is the average value,  $U$  is the threshold coefficient.

It is obvious that the parameters of function (5), the window length  $l$  and the threshold coefficient  $U$ , are adjustable and determine the size of a time window within which geomagnetic disturbances are estimated and the level of determined geomagnetic disturbances (we applied the window length of  $l = 1440$  that corresponds to 24 hours and the threshold coefficient  $U = 7$ ).

To estimate the intensity of the detected disturbances at an instant time  $t = b$  according to the paper [30], we apply the value of

$$Y_b = \sum_a P_{T_a}(W_\Psi f_{b,a}) \quad (7)$$

We make wavelet transform of value  $Y_b$  (see (4))

$$(W_\Psi Y_{c,d}) := |d|^{-1/2} \int_{-\infty}^{\infty} f(t) \Psi\left(\frac{t-c}{d}\right) dt, \quad d, c \in R, \quad d \neq 0, \quad (8)$$

and taking into account that a wavelet is a window function [31], we obtain a dynamic spectrum of geomagnetic disturbance intensity.

### 3. Processing results of geomagnetic data during the magnetic storms on January 7, 2015 and March 17, 2015

Based on the suggested method, we processed and analyzed the data from the sites in the north-eastern segment of Russia (Table 1). To analyze the processes in the magnetosphere at the near equatorial latitudes, the data of the Indian HYB ‘‘Hyderabad’’ and CPL ‘‘Choutuppal’’ sites were used. The considered events and the results of application of the developed method for detection of anomalous increases of geomagnetic disturbance intensity before magnetic storms is shown in Table 2. Analysis of the results of Table 2 indicates the possibility of occurrence of weak geomagnetic disturbances before magnetic storms, which was first mentioned in the papers [33, 34]. It shows high sensitivity and the efficiency of the method suggested in the paper. In what follows are the detailed results of geomagnetic data processing during geomagnetic storms which occurred on 07.01.2015, 17.03.2015, 21.06.2015, 15.08.2015, 19.12.2015.

Table 1. Sites of the north-eastern segment of Russia.

Observatory	Code	Geographical latitude	Geographical longitude	Geomagnetic latitude	Geomagnetic longitude	Local time (LT)
Magadan (1)	MGD	59°33.1'	150°48.3'	51°32.4'	146°2.4'	UTC+11
Paratunka (1)	PET	52°58.3'	158°15.0'	45°51.6'	137°57.6'	UTC+12
Khabarovsk (1)	KHB	48°29.0'	135°04.0'	39°15'	156°48.6'	UTC+10
Choutuppal (2)	CPL	17°17.33'	78°55'	8°37.2'	152°34.8'	UTC+5:30

Note: site affiliation is indicated in brackets (1) – IKIR FEB RAS, (2) – CSIR-National Geophysical Research Institute.

Table 2. Results of detection of anomalous increases of geomagnetic activity before storms.

Date of a storm	Storm source	Time of storm beginning (UT)	max Kp	max Dst	Anomalies detected before a magnetic storm	Maximum value of the detected disturbance intensities		
						Time interval before a magnetic storm beginning	Scales	KHB site
07.01.2015	CME	6:15	6	-103	4 hours 55 minutes	4-10	20	0
					12 hours 10 minutes	2-8	12	27
					21 hours	14-20	1085	328
17.03.2015	CME	4:45	8	-233	11 hours 15 minutes	5-40	523	225
					12 hours 10 minutes	20-34	232	0
					16 hours 30 minutes	6-10	3	2
21.06.2015	CIR/CME	16:55	8	-111	7 hours	10-16	2	53
					10 hours	0-8	2	0
15.08.2015	CME/CIR	8:30	6	-64	5 hours	8-12	0	209
					5 hours 20 minutes	40-60	2	415
					5 hours 30 minutes	8-20	65	35
					7 hours	0-8	3	0
					10 hours	16-32	1	0
					13 hours 30 minutes	8-22	18	19
					18 hours 40 minutes	0-4	1	0
19.12.2015	CME	16:18	7	-170	1 hour	8-32	1	0
					7 hours 20 minutes	6-12	4	0
					15 hours 30 minutes	2-8	5	0
					25 hours 30 minutes	0-4	1	1

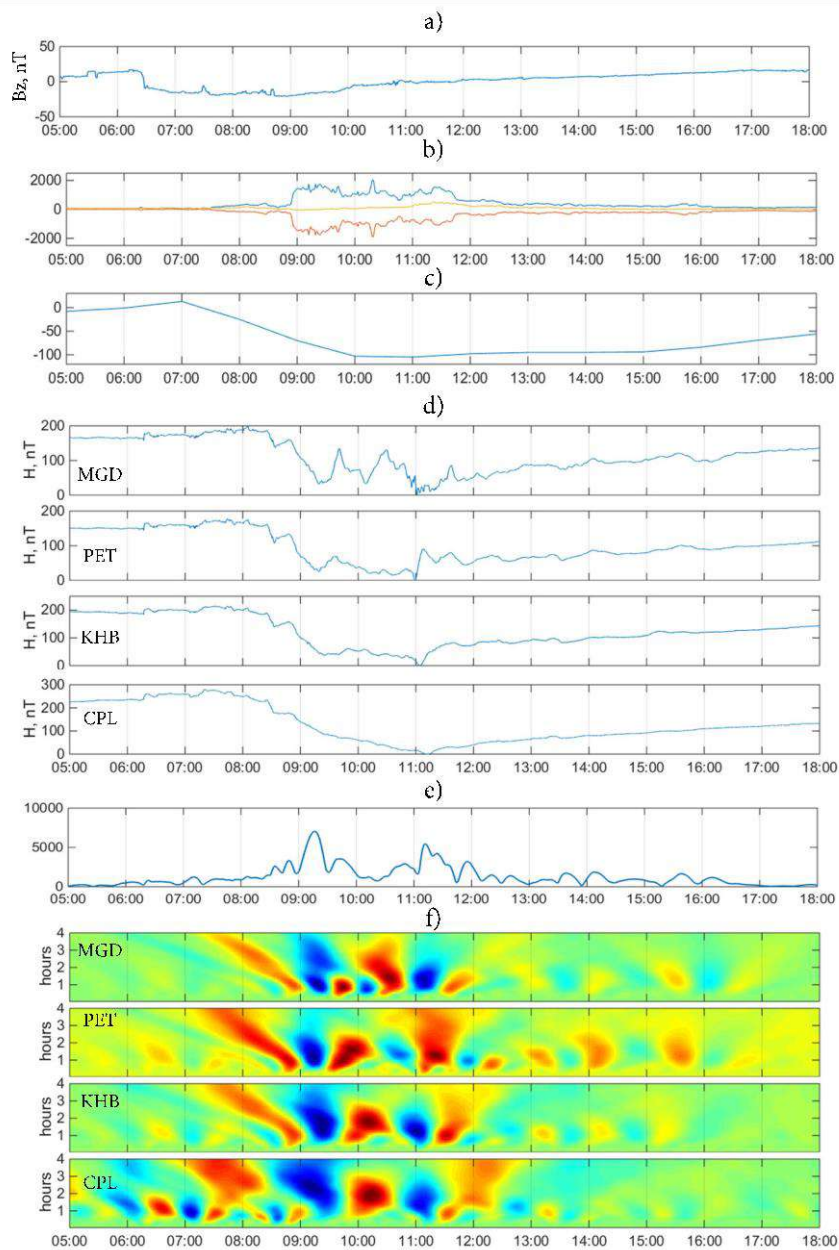


Fig 1. Processing results of the data for January 7, 2015; a)  $B_z$  component of the Interplanetary Magnetic Field; b) AE-index (yellow line) and AL-index (red line) and AU-index (blue line); c) Dst-index; d) H-component of the magnetic field; e) geomagnetic disturbance intensity (relation (5)); f) dynamic spectrum of geomagnetic disturbance intensity.



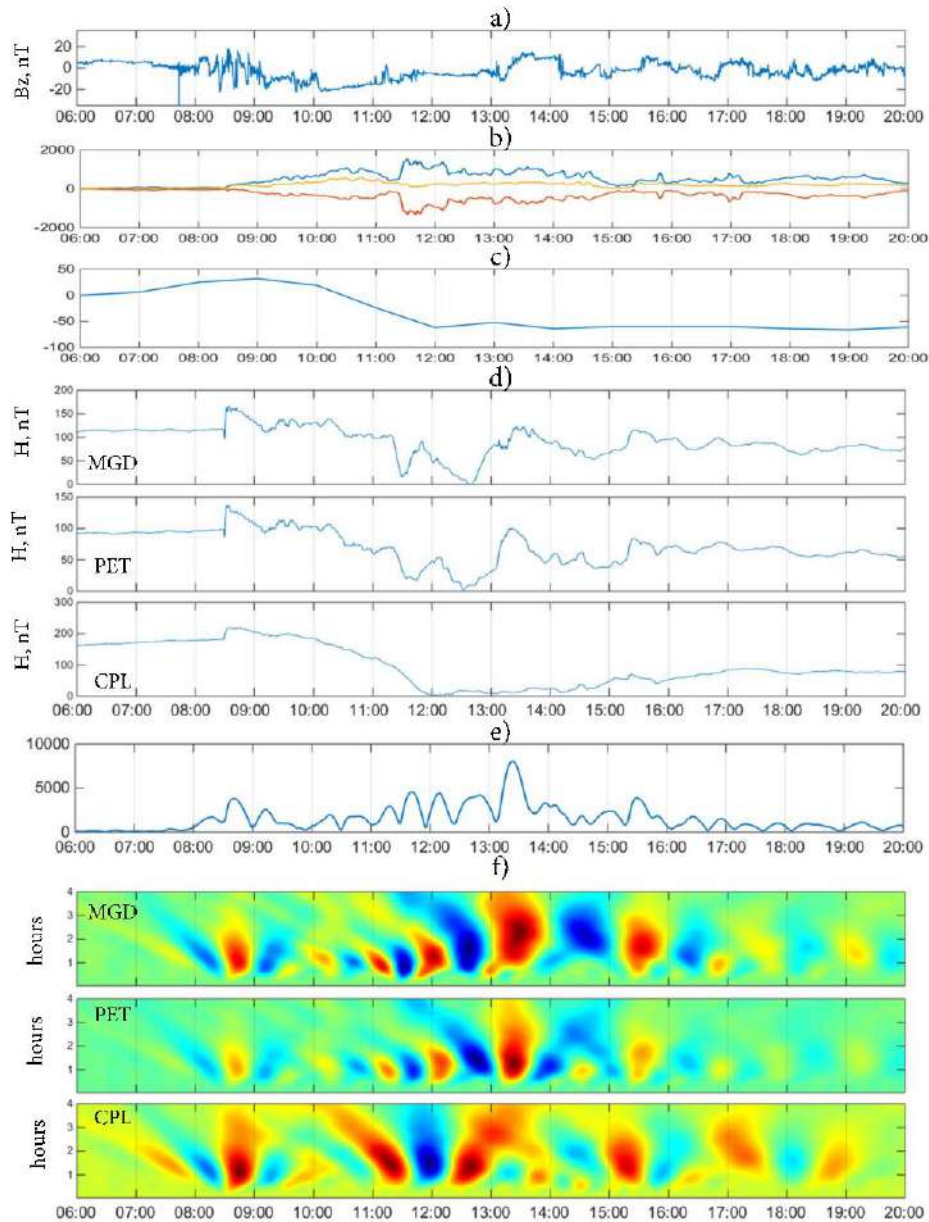


Fig. 2. Processing results of the data for August 15, 2015; a)  $B_z$  component of the Interplanetary Magnetic Field; b) AE-index (yellow line), AU-index (blue line) and AL-index (red line); c) Dst-index; d) H-component of the magnetic field; e) geomagnetic disturbance intensity (relation (5)); f) dynamic spectrum of geomagnetic disturbance intensity.

Fig. 1 shows the results of processing of geomagnetic data during the magnetic storm on January 7, 2015. This event was caused by coronal ejection of solar material (CME on January 4, <http://ipg.geospace.ru/space-weather-review-07-01-2015.html>). Its dynamics was of classical character with clearly defined major phases of a storm in Dst-variation (Fig. 1c). The results of estimation the geomagnetic disturbance intensity shows that during the initial stage of the storm, from about 07:00 UT, geomagnetic activity gradually increased and the Dst-index had positive values. Maxima of disturbance intensity (Fig. 1e) are observed during Dst-index decrease and AE-index increase characterizing the occurrence of an intensive substorm in the auroral zone. The dynamic spectrum of geomagnetic disturbance intensity (relation (5)) illustrated in Fig. 1f shows the regions of disturbance concentration and propagation in the areas under analysis. During the event, a general picture of the dynamics of magnetospheric current systems is observed. The beginning of the storm from 6:00 to 08:00 UT was the most clearly defined at the near equatorial site (India). During the main phase of the storm, activation areas are observed in the dynamic spectra of all the sites (Fig. 1f; red color is the intensity increase; blue color is the intensity decrease). They are likely to characterize large-scale processes in the magnetosphere probably associated with energy accumulation and release during the event. At the most northern site Magadan, local regions (from 10:00 to 11:00 UT) are distinguished. They are likely to be associated with auroral processes.

Fig. 2 shows the data processing for the magnetic storm on August 15, 2015 which was caused by a solar medium coronal mass ejection (CME on August 12) and high-velocity flows from a coronal hole (CIR). The magnetic storm began at 08:30 UT during a sharp increase of solar wind velocity and increase of the magnetic field horizontal component at all the sites under analysis. Short-period anomalous increases of geomagnetic activity began about 12 hours before the magnetic storm (Fig. 2e). The highest values of geomagnetic disturbance intensity are observed at the sites during the main phase of the storm (the period of significant decrease of Dst-indexes). The wavelet spectrum of geomagnetic disturbance intensity shows that geomagnetic field



disturbances at all the sites increased in the vicinity of special points (points of local extreme periods, function inflection) of Dst-variation (08:00-10:00; 11:00-13:00 UT). It indicates active processes in the magnetosphere at these time periods. The disturbances at the near equatorial site CPL had the most clearly defined character.

The results of application of the developed method for detection of pre-storm anomalies of geomagnetic disturbance intensity increases are illustrated on the example of the event on August 15, 2015 in Fig. 3. Analysis of Fig. 3e shows synchronous anomalous increases of geomagnetic activity 18 hours before the magnetic storm. Several minutes before registration of the event at the sites, short-time geomagnetic disturbance intensity significantly increased and reached the maximum values during the initial phase.

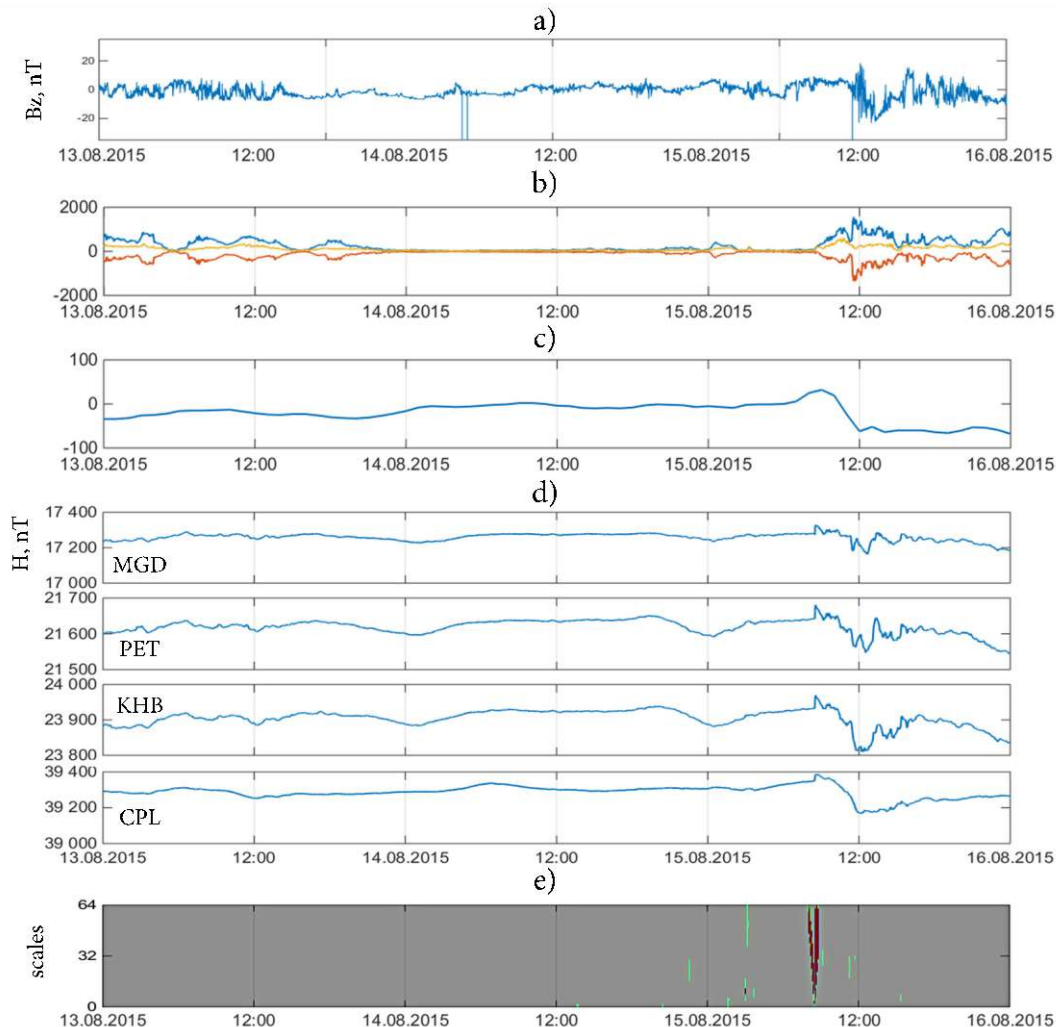


Fig. 3. Processing results of the data for August 13-15, 2015; a) Bz component of the Interplanetary Magnetic Field; b) AE-index (yellow line), AU-index (blue line) and AL-index (red line); c) Dst-index; d) H-component of the magnetic field; e) detection of the periods of increased geomagnetic activity (relation (6)).

#### 4. Conclusions

A detailed analysis of geomagnetic data during strong magnetic storms in 2015 was carried out by the suggested method. The dynamic spectrum of geomagnetic disturbance intensity showed spatial pattern of the events and allowed us to analyze geomagnetic disturbance propagation along the observation meridian and at the near equatorial sites. During the main phases of the storms, activation areas were detected. They have large spatial scales and are likely to be associated with the processes of energy accumulation and release in the magnetosphere. Before the events, synchronous local increases of geomagnetic activity were observed at the sites under analysis. They are likely to be associated with nonstationary effect of solar wind plasma on the Earth magnetosphere in the course of an interplanetary disturbance approaching. Such anomalous pre-storm effects are mentioned in the papers [33, 34]. According to the processing results of large experimental material and joint analysis of geomagnetic field H-component oscillations with the oscillating processes on the Sun, the authors [13, 33] showed that the success rate of the suggested forecast method for the geoeffective flare events is 90%. This result indicates high probability of possible occurrence of pre-storm anomalous features in geomagnetic data. The possibility of automatic registration of anomalous feature data is an important aspect of the suggested method for space weather forecast.

#### Acknowledgments

The development of the method for geomagnetic data analysis was supported by RSF Grant № 14-11-00194. The data primary analysis was supported by RFBR Grant № 16-55-45007. The authors are grateful to the Institutes supporting the magnetic observatories which data were used in the investigation.

## References

- [1] Yermolaev YuI, Yermolaev MYu. Solar and Interplanetary Sources of Geomagnetic Storms: Space Weather Aspects. *Izvestiya, Atmospheric and Oceanic Physics* 2010; 46(7): 799–819.
- [2] Nayar SRP, Radhika VN, Seena PT. Investigation of substorms during geomagnetic storms using wavelet techniques. *Proceedings of the ILWS Workshop Goa, India, 2006*: 328–331.
- [3] Hafez AG, Ghamry E, Yayama H, Yumoto K. Systematic examination of the geomagnetic storm sudden commencement using multi resolution analysis. *Advances in Space Research* 2013; 51: 39–49.
- [4] Xu Z, Zhu L, Sojka J, Kokoszka P, Jach A. An assessment study of the wavelet-based index of magnetic storm activity (WISA) and its comparison to the Dst index. *J. Atmos. Solar–Terr. Phys.* 2008; 70: 1579–1588.
- [5] Jach A, Kokoszka P, Sojka J, Zhu L. Wavelet-based index of magnetic storm activity. *J. Geophys. Res.* 2006; 111(A9). DOI:10.1029/2006JA011635.
- [6] Paschalis P, Sarlanis C, Mavromichalaki H. Artificial neural network approach of cosmic ray primary data processing. *Solar Physics* 2013; 182(1): 303–318.
- [7] Macpherson KP, Conway AJ, Brown JC. Prediction of solar and geomagnetic activity data using neural networks. *J. Geophys. Res.* 2001; 100: 735–744.
- [8] Woolley JW, Agarwal PK, Baker J. Modeling and prediction of chaotic systems with artificial neural networks. *International Journal for Numerical Methods in Fluids* 2010; 63. DOI:10.1002/flid.2117.
- [9] Soloviev A, Chulliat A, Bogoutdinov S, Gvishiani A, Agayan S, Peltier A, Heumez B. Automated recognition of spikes in 1 Hz data recorded at the Easter Island magnetic observatory. *Earth Planets Space* 2012; 64(9): 743–752.
- [10] Rotanova N, Bondar T, Ivanov V. Wavelet Analysis of Secular Geomagnetic Variations. *Geomagnetism and Aeronomy* 2004; 44: 252–258.
- [11] Rybák J, Antalová A, Storini M. The wavelet analysis of the solar and cosmic-ray data. *Space Science Reviews* 2001; 97: 359–362.
- [12] Zaourar N, Hamoudi M, Manda M, Balasis G, Holschneider M. Wavelet-based multiscale analysis of geomagnetic disturbance. *Earth Planets Space* 2013; 65(12): 1525–1540.
- [13] Smirnova AS, Snegirev SD, Sheyner OA. Sun ultraviolet radiation as a possible cause of preflare long-period oscillations of geomagnetic field horizontal component. *Vestnik of Lobachevsky University of Nizhni Novgorod* 2013; 6(1): 88–93.
- [14] Mandrikova OV, Solov'ev IS, Zalyaev TL. Methods of analysis of geomagnetic field variations and cosmic ray data. *Earth Planet Space* 2014; 66. DOI:10.1186/s40623-014-0148-0.
- [15] Golovkov VP, Papitashvili VO, Papitashvili NE. Automated calculation of the K indices using the method of natural orthogonal components, *Geomagn. Aeron* 1989; 29: 667–670.
- [16] Nowożyński K, Ernst T, Jankowski J. Adaptive smoothing method for computer derivation of K-indices. *Geophys. J. Int.* 1991; 104: 85–93.
- [17] Menvielle M, Papitashvili N, Hakkinen L, Sucksdorff C. Computer production of K indices: review and comparison of methods. *Geophys. J. Int.* 1995; 123: 866–886.
- [18] Mandrikova OV, Smirnov SE, Solov'ev IS. Method for Determining the Geomagnetic Activity Index Based on Wavelet Packets. *Geomagnetism and Aeronomy* 2012; 52(1): 111–120.
- [19] Bogoutdinov SR, Gvishiani AD, Agayan SM, Solovyev AA, Kihn E. Recognition of Disturbances with Specified Morphology in Time Series. Part 1: Spikes on Magnetograms of the Worldwide INTERMAGNET Network. *Izvestiya, Physics of the Solid Earth* 2010; 46(11): 1004–1016.
- [20] Sidorov RV, Soloviev AA, Bogoutdinov ShR. Application of the SP algorithm to the INTERMAGNET magnetograms of the disturbed geomagnetic field. *Izvestiya, Physics of the Solid Earth* 2012; 48(5): 410–414.
- [21] Krasnoperov RI, Soloviev AA. Analytical geoinformation system for integrated geological-geophysical research in the territory of Russia. *Gornyi Zhurnal* 2015; 10: 89–93. DOI: 10.17580/gzh.2015.10.16.
- [22] Soloviev A. et al. *Data Science Journal* 2013; 12. DOI:10.2481/dsj.WDS-019.
- [23] Uwamahoro J, McKinnell LA, Habarulema JB. Estimating the geoeffectiveness of halo CMEs from mass associated solar and IP parameters using neural networks. *Annales Geophysicae* 2012; 30: 963–972.
- [24] Mahrous A, Radi A, Youssef M, Faheem A, Ahmed S, Gopalswamy N. Prediction of the interplanetary Coronal Mass Ejection and its associated shock by using neural network. 38th COSPAR Scientific Assembly in Bremen, Germany 2010; D23-0052-10.
- [25] Pallochia G, Amata E, Consolini G, Marcucci MF, Bertello I. Geomagnetic Dst index forecast based on IMF data only. *Annales Geophysicae* 2006; 24: 989–999. URL: [www.ann-geophys.net/24/989/2006/](http://www.ann-geophys.net/24/989/2006/).
- [26] Mandrikova OV. Multicomponent model of a signal with a complicated structure. *Problems of the evolution of open systems* 2008; 2(10): 161–172. (in Russian)
- [27] Mandrikova OV, Polozov YuA. Approximation and analysis of ionospheric parameter based on the combination of wavelet transform and neural network collectives. *Information technologies* 2014; 7: 61–65. (in Russian)
- [28] Geppener VV, Mandrikova OV, Zhizhikina EA. Automatic method for estimation of the earth's magnetic field state. *Proceedings of international conference on soft computing and measurements, SCM* 2015; 18: 251–254. DOI: 10.1109/scm.2015.7190473.
- [29] Mandrikova OV, Zhizhikina EA. An automatic method for estimating the geomagnetic field. *Computer Optics* 2015; 39(3): 420–428. DOI: 10.18287/0134-2452-2015-39-3-420-428.
- [30] Mandrikova OV, Solov'ev I, Geppener V, Taha A-KR, Klionskiy D. Analysis of the Earth's magnetic field variations on the basis of a wavelet-based approach. *Digit Signal Process* 2013; 23: 329–339.
- [31] Chui CK. *An introduction in wavelets*. Academic Press, New York, 1992; 264 p.
- [32] Daubechies I. *Ten Lectures on Wavelets*. CBMS-NSF Lecture Notes 1992; 61: 377 p.
- [33] Sheiner OA, Fridman VM. The features of microwave solar radiation observed in the stage of formation and initial propagation of geoeffective coronal mass ejections. *Radiophysics and Quantum Electronics* 2012; 54(10): 655–666.
- [34] Mandrikova OV, Bogdanov VV, Solov'ev IS. Wavelet analysis of geomagnetic field data. *Geomagnetism and Aeronomy* 2013; 53(2): 268–273.

# Large scale networks security strategy

Ya. Mostovoy<sup>1</sup>, V. Berdnikov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The article deals with optimum two-phase planning of secure routs in large scale computer networks. Uncertainty of future needs is covered by extensive statistical modeling, which resulted in identification of statistical dependences and phenomena allowing for optimization of creation. To describe secure paths in random matrices the author uses programmable percolation apparatus. Tolerance of the created secure routes to failures in certain secure paths is demonstrated here.

*Keywords:* IT security; large scale networks; percolation; programmable percolation; two-phase operations

---

## 1. Introduction

Large scale (complex) networks are characterized by the large number of nodes, paths connecting them and mixed topology. There are a number of crucial research tasks pertaining to such networks, for example, analysis of dimensions and number of various-object clusters appearing in the networks; analysis of paths connecting nodes and clusters; analysis of nodes, removal of which may cause disintegration of the network into unlinked parts and etc.

The main task of the security strategy being a generalized long-term activity plan aimed at assuring security of large scale networks is effective use of limited resources.

In this case problem solving done in a responsive planning way basing on minimum aggregate expenditures allows succeeding.

Papers [10, 14, 17] tackle the issue of using the classical percolation theory for the applied research networks. However, the authors never address methods to study large scale networks based on programmable percolation theory explicated in [5, 6, 7, 8].

Habitually the percolation theory describes a grid of vertices and bonds or a square matrix of  $L$  lines, where the random number of cells is black allowing liquid, gas, traffic or data through, whereas the rest of cells are white or closed. If the concentration (probability of occurrence) of black cells increases, some of them randomly adjoin with the edges and merge. These black cells with adjoining edges make random open-bond clusters. These clusters appear and grow as the black cells concentration grows [1, 2].

The classical percolation theory describes randomly-filled matrix representing model of environment in direct geometrical interpretation [1, 3, 4, 9, 15, 16]. As it is, such an approach does not suit for large scale networks security analysis, because bonds among network nodes are diverse or even ill-defined, and though there are protective bonds among nodes they do not cover the majority of possible routes. It is necessary to migrate from network topology to that of the secure paths inter nodes matrix (SPNM). Such a matrix might ignite research of network security and availability of through paths using methods of the percolation theory.

## 2. Statement of the problem, method of analysis and computer experiment

A large scale network is considered here. Paths among certain nodes are secure. Resource scarcity makes it impossible to build all possible secure routes at once. One may make a secure route each time it is necessary, though it is time-consuming and costly, if compared to other routes incorporating available secure paths (or passing through clusters of such secure paths) provided the latter are abundant. In this case, to create the required route it is necessary to introduce few secure paths covering inter-cluster gaps.

Uncertainty in realization of the given secure route may be determined by statistical analysis based on a large number of random routes.

Thus, the network under consideration has a random number of securely connected nodes making up a somewhat stochastic secure basis. Now it is possible to build a completely secure route via any nodes of the network introducing additional secure segments where they are missing or necessary. As these additional secure segments are formed emergently, they require thorough positioning. Besides they are more expensive than secure paths from stochastic basis.

It is required to define probability of a secure path in the stochastic basis (in terms of the percolation theory – concentration of the open black cells) which minimizes overheads of building secure routs in the network.

In the classic percolation theory [1, 2, 12, 15, 16] they define  $K_{th}$  – the concentration of the open black cells or stochastic percolation threshold, when a random route passing through black cells from top to bottom of the matrix in the given direction, i.e. stochastic percolation cluster, appears. However, this stochastic percolation cluster has loose structure, considerable number of dead branches and is obviously redundant for real-world application.

With the programmable percolation [5, 6, 7, 8] at the first stage there is built a basis consisting of randomly distributed secure paths making clusters and having concentration well below the stochastic percolation threshold. At the second stage by inserting additional secure paths into existing inter-cluster gaps there is created a through percolation route. Here concentration of the stochastic basis is chosen to make cumulative cost of the two-phase operation minimal. Solving of this problem shows

that programmable percolation allows having concentration of objects (secure paths) more than twice as little as the stochastic percolation threshold. As well the concentration of objects is in the neighborhood of concentration typical to average maximal number of clusters appearing ( $K = 0.25$ ).

As far as targets of research are large scale networks and statistical phenomena of secure-path clusters, and the goal of research is long-term planning of optimal-cost secure routes in large scale networks, our theoretical considerations are verified by a computer experiment - the only possible way of application investigation.

The computer experiment for long-term planning of secure routes in large scale networks consisted of a number of consecutive stages, repeated for each of randomly filled matrices (SPNM) being models of operation environment.

Each time the following steps were made for each value of secure routs concentration:

- the matrix was randomly filled with objects in conformity with the predetermined probability law and concentration value;
- the resultant clusters and objects were identified and analyzed;
- measures of cluster distribution (average values, scatter and etc.), cluster size, inter-cluster gaps and etc. were calculated;
- gaps between stochastically-formed clusters were analyzed; shortest artificial percolation paths were formed; average length of the above mentioned path was measured and average number of additionally inserted secure segments covering inter-cluster gaps was calculated per totality of randomly-filled matrices.

In order to identify clusters and estimate their characteristics we used the Hoshen-Kopelman algorithm [11, 13]. To make paths through clusters we created a Lightning – Closest Point algorithm which is an adaptation of Lightning strike and Dijkstra's algorithms [5, 6, 7, 8].

### 3. SPNM properties

Classic percolation theory considers a randomly-filled matrix to be a model of environment in direct geometrical interpretation. Such an approach does not suit for analysis of network security, because network topology cannot be rendered by a two-dimensional array. It is necessary to migrate from network topology to that of the secure paths inter nodes matrix (SPNM). Such a matrix might ignite research of network security and availability of through paths using methods of the percolation theory.

Example of such a transmission is demonstrated below in Figures 1 and 2.

Black lines are data connections (paths) between network nodes. Yellow lines are secure connection (secure paths) between the nodes.

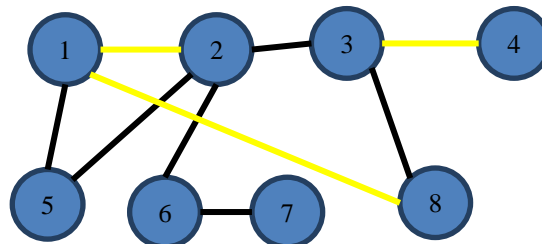


Fig. 1. Random network graph.

	1	3	6
2			
8			
5			
7			
4			

Fig. 2. SPNM for secure sections of the network? Demonstrated in Figure 1.

SPNM is filled according to the following rule: end-node names of interest are recorded in the vertical direction, start-node names of interest are recorded in the horizontal direction (in Figure 2 they are blue). **Note that nodes in the vertical and horizontal directions are not repeated.** The suggested research tool SPNM strict squareness is unimportant. Casual randomization of lines and columns is possible. SPNM filling algorithm is the following:

1. Repeat unless all nodes are done:

1.1. If node  $A$  is missing in the table, record the node name in the horizontal direction.

1.2. Record nodes, which are connected with the node  $A$  in the network in vertical direction.

1.3. Mark SPNM cells correspondingly: black if there is secure connection between the node  $A$  and other nodes from the table.

2. End of the loop.

Adjoining black cells make up a cluster. The point is that information can be securely transferred via this segment of the network. In the given example (see Fig.2) there are two clusters: the «1-2, 1-8» cluster and the «3-4» cluster.

If certain inter-cluster gaps in the SPNM are filled with secure connections (marked red), then a through non-stochastic but programmed percolation route is created in the SPNM. It means that all the nodes recorded in the vertical direction are available for secure connection with the nodes recorded in the horizontal direction.

Thus, secure interconnection of all the nodes recorded in the vertical direction is rendered on the SPNM as a programmable percolation vertical route (see Fig.3):

	1	3	6
2	■		
8			
5	■	■	
7		■	
4		■	

Fig. 3. Programmable percolation in SPNM.

It is obvious, that the number of such secure paths might be great. They might pass through one or several nodes located on the horizontal axis. There might be other percolation routes generated with directed percolation. To plot the shortest route in the given direction we used an adaptation of Dijkstra's algorithm. All programmable percolation routes are the subject of statistical modeling.

For statistical analysis we used different-size SPNM filled with the help of the random number generator.

Example of an SPNM randomly filled with secure paths (black cells) is given in Figure 4. Concentration of the black cells differs. SPNM size here is  $50 \times 50$ . Possible shortest routes of the programmable percolation in the bottom-top direction across the SPNM are plotted in red. Note greater tortuosity of the programmable percolation route across the matrix with  $K = 0.6$  concentration.

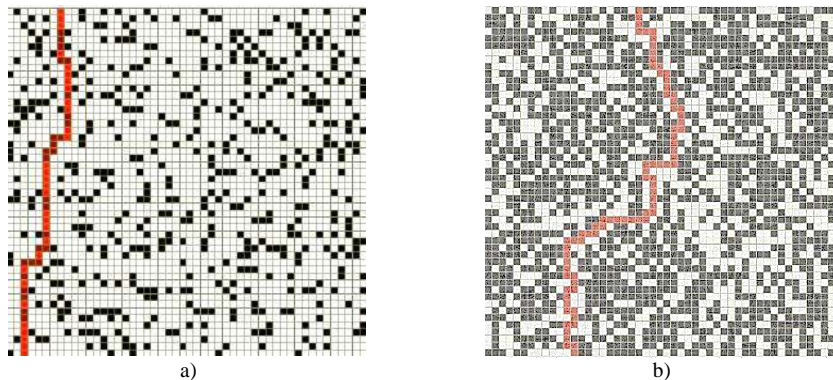


Fig. 4. Examples of percolation routes across matrices with a)  $K = 0.25$ , b)  $K = 0.6$  population concentration.

#### 4. Some statistical peculiarities of clusters' formation in large scale networks

Concentration  $K$  is a relative fraction of black nodes during random and homogeneous filling of the matrix. It makes black cells [2] likely to appear, when probability of their occurrence in the matrix is uniformly distributed. That is why here and elsewhere we use both: the expression "probability of the predefined object (secure path) in the matrix cell" and its epitomized version - "concentration".

Statistical modeling using square randomly filled matrices allows detection and analysis of cluster statistical phenomena (peculiarities) being of great practical consequence.

The first peculiarity is presence of the stochastic percolation threshold in the shape of matrix dissection by the open percolation cluster. It is guaranteed at  $K = 0.6$ .

The second peculiarity is such concentration of objects when average number of clusters is maximum [5, 6, 7, 8, 18]. A *ibid* is demonstrated that the value is robust, i.e. low responsive to the object presence in the matrix cell probability distribution law. This peculiarity manifests itself at  $K = 0.25$  (see Fig. 5).

The third statistical peculiarity is maximum average length of the shortest route through the stochastically formed clusters in the percolation direction. This value appears when the population of objects and route tortuosity grows. Average length of the programmable percolation  $L(K)$  shortest route grows up to the stochastic percolation threshold, and upon reaching it starts decreasing. The more tortuous is the percolation route (i.e. the longer it is), the more passing clusters it incorporates.

#### 5. Analysis of two-phase operations

During statistical modeling we considered several thousands of different size matrices. The cells of those matrices were randomly filled with provision for equal probability of objects distribution in the cells. In order to identify all the clusters in the received random matrices we used the Hoshen-Kopelman algorithm [5, 6, 7, 13]. Then we estimated their statistical characteristics and plotted curves of average values.

Dependence of the average number of clusters in the matrix from the probability of the object in the cell  $K$  is given in Figure 5. When the probability increases up to  $\sim 0.25$ , the matrix is being filled with the objects, and the number of cluster

grows. Further growth of concentration results in merging of the clusters. Their average number decreases while their size grows.

On several physical grounds we established that the number of clusters appeared in the matrix with the certain concentration depended on the matrix area size  $L^2$ , while length of routes depended on the linear dimension of the matrix  $L$ . Consequently, it is possible to save numerical results of statistical modeling from influence of the matrix size by dividing them by  $L$  or  $L^2$  correspondingly. Numerical computations verify the above said (see Fig.5).

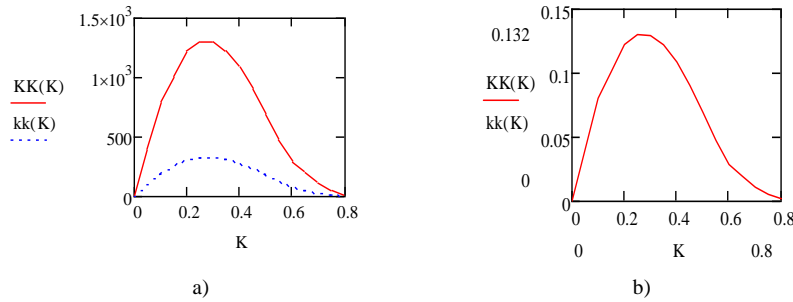


Fig. 5. Dependence of the average number of clusters on the object probability in the cell for a 50x50 matrix (in dots) and a 100x100 matrix – a) average number of clusters normalized by the matrix area size – b) for both cases.

We may decrease the appropriate concentration and consequently number of objects necessary for percolation, if we replace classical stochastic percolation with the suggested programmable percolation and apply the two-phase approach.

Taking into account different value of type I objects randomly distributed to form a stochastic basis (black cells) and type II objects inserted in certain places of the coverage area to get the shortest programmable percolation route (red cells), we are able to come at a such concentration of the stochastic basis when total cost of the created programmable percolation route is minimal.

Having said this it can be believed that each of the objects from the stochastic basis scattered in the operating environment is cheaper than an additional object inserted into a certain place of the same operating environment.

Figure 6 demonstrates processed results of two-phase operations computer experiment: average number of the inserted objects necessary for programmable percolation with various concentrations of objects in the stochastic basis and for different-size matrices. Figure 6a: in vertical direction is given the average number of the objects inserted in 50x50 matrix (dotted line) and 100x100 matrix. Figure 6b: the dependences are normalized according to the matrix size (whereupon the graphs coincided).

Stochastic percolation cluster is formed at concentration  $K = 0.6$  and the shortest percolation route passes through it. That is why in this case the average number of the added cells tends to zero. At this concentration tortuousness and length of the percolation route are maximal. Further growth of the concentration makes the shortest percolation route more straight and its length decreases (Fig. 6c) [8].

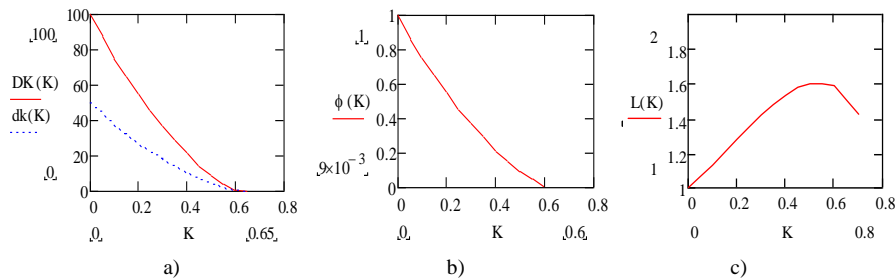


Fig. 6. Dependency of the average number of inserted objects  $\varphi(K)$  and the average normalized length of the programmable percolation route ( $K$ ) from the probability of the object in the cell  $K$ .

Let us calculate cost of the two-phase operation. The cost of finding (preparation) of each random secure path is designated as  $\alpha$ , the cost of a single additional secure path selected (prepared) in a certain place of a large scale network during the second phase is designated as  $\theta(K)$ .

Then the total cost of the two-phase operation P is:

$$P = \alpha * K * L^2 + \theta(K) * \varphi(K) * L \tag{1}$$

Where the first term is the cost of preparation of the operating environment stochastic basis,  $K * L^2$  – number of basic secure paths in the stochastic basis expressed as concentration function. The addend in (1) is the cost of the secure paths necessary to form the shortest programmable percolation route through stochastically generated clusters.  $\varphi(K) * L$  is average number of the added secure paths in SPNM of  $L$  size specified by stochastic computer experimental results and demonstrated in the normalized dependence (see Fig. 6c).

$\theta(K)$  function reflects cost of each secure path created and inserted in the large scale network variation versus stochastic basis concentration.

We assume that the cost of each additional secure path created in a certain SPNM cell is proportional to the size and number of inter-cluster gaps covered along the percolation route. In other words it is proportional to the number of the reds in



the route  $\varphi(K) * L / L(K) * L$  and inversely proportional to relative tortuosity of the route  $L(K) / L$  as maximal tortuosity

means absence of gaps to be covered (absence of the reds). Therefore, wealth of gaps in the route results in greater cost of each additional secure path,  $\theta(K) = \theta_0 * \varphi(K) * L / L(K)^2$ . With account of this equation, the cumulative costs formula (1) shall be

written as:

$$P = \alpha \times K \times L^2 + \theta_0 * \varphi(K)^2 * L^2 / L(K)^2 \quad (2)$$

Let us analyze relative cost of a two-phase operation. For this we divide the left-hand side and the right-hand side of the obtained equation (2) by  $P_n = \alpha * K_n * L^2$  – cost of a purely stochastic one-phase operation.

Then:

$$P_{OTH} = P / P_n = 1.7K + 1.7 * \left( \frac{(\theta_0 * \varphi(K)^2)}{(\alpha * L(K))} \right) = 1.7 \left( K + R * \varphi(K)^2 / L(K)^2 \right), \quad (3)$$

where  $R = \theta_0 / \alpha$  is ratio of the additional object cost to the stochastic basis object cost. Figure 7 demonstrates two-phase operation relative cost variation versus stochastic basis  $K$  obtained with the above equation taking into consideration  $\varphi(K)$  and  $L(K)$  variations (see Fig. 5) for  $R = 1$ .

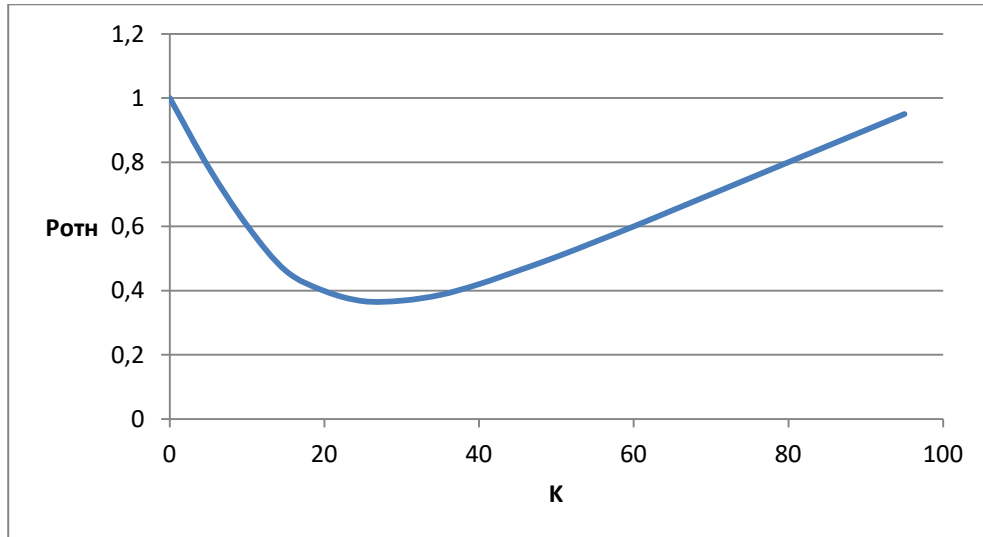


Fig. 7. Two-phase operation relative cost variation versus concentration of objects in the stochastic basis.

The plot in Figure 7 demonstrates that from the perspective of two-phase operations total cost minimization, optimal probability of a secure path in the stochastic basis cell shall be  $\sim 0.25$ , which corresponds to the maximal number of clusters in the stochastic basis (see Fig. 5). This remarkable point does not explicitly occur in the equations used for plotting of the graph (see Fig. 7). The result may be interpreted as validation of statistic computer experimental data and model of two-phase operation labor consumption.

## 6. Percolation route stability analysis

Complex technical systems rarely work as expected. But smart security strategy shall consider off-design operation and available redundancy. This is the only way to overcome failures and errors.

For this reason large scale network security model shall support safe operation of the network even in case of node faults. Statistical analyses of SPNM suggested in the article enables pointedly handle the issue.

We studied failures of secure bonds along a percolation route. Here failure means that protection in the SPNM cell is destroyed and it results in interruption of secure information stream along a chosen route. Supposing that, it is impossible to promptly recover the fault point, but to bypass. The question is: what is the cost of such a bypass, how much is the route lengthened? It is evident that the answer depends on the large scale network topology or on the concentration of the blacks if we consider our percolation model.

So, to answer the question we performed statistical computer experiment. Computer-aided experiment was conducted in the following way: first of all we randomly chose a cell on the percolation route and put a veto on the route passing through it.

Then, we let a new percolation route start from the preceding point in the same direction and on the same conditions of optimality bypassing the fault. Probability that the new percolation route reverts to the former one is demonstrated in Figure 8; optimal concentration of the stochastic basis  $K = 0.25$  was estimated in advance. The plot obtained in the cause of the experiment did not depend on the matrix size. The plot was normalized according to the following rule:  $L_n = i/L$ , where  $i$  is coordinate position of the failure in SPNM in the vertical direction,  $L$  is the SPNM height.

It is safe to say that the route is invariable up to  $L_n = 0.85$ , i.e. a new percolation route is likely to revert to the former one. All the upsurges seen on the plot are within statistical error.

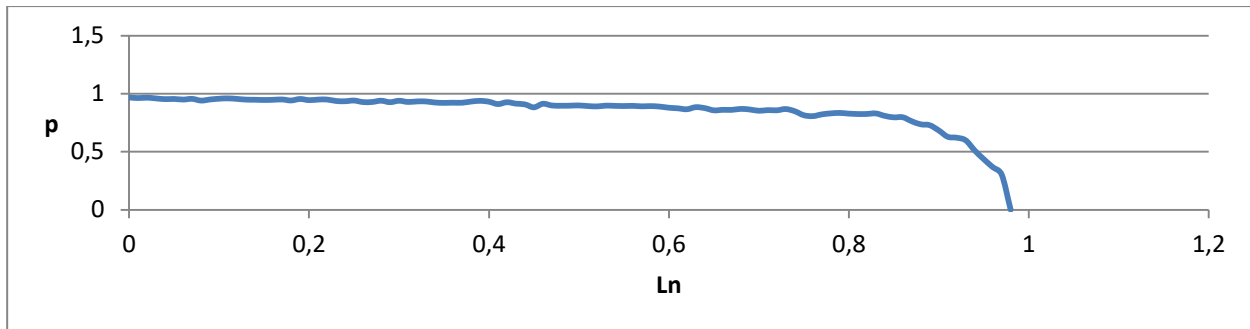


Fig. 8. Probability of a new route to revert to the formed one from the failure point.

As a part of the study were obtained variations of the green-cell number versus concentration for matrices of different  $L$  sizes. From now on green cells are the cells newly added to the percolation route with the purpose to bypass a banned cell denoting a failure.

Variation of the green cells number  $L_G$  versus concentration is given in Figure 9. Actually, the plot for matrices of different sizes is the same.

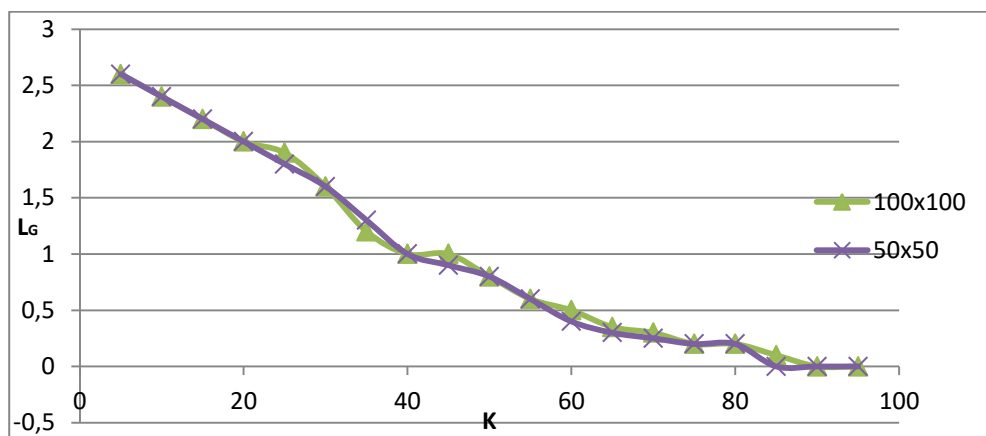


Fig. 9. Variation of the number of green cells added to a bypass route versus concentration for matrices of different sizes.

Note that all the deviations of the plot are within statistical error.

Based on the findings it follows that the number of the green cells independent from the matrix size, but depends on the concentration. Therefore, the fault phenomenon is of purely local nature.

## 7. Analysis of programmable percolation in SPNM

Besides, we studied programmable percolation, i.e. making routes from the given point to the target point. This problem might be highly topical for information networks outside statistical research, when, for example, it is required to establish secure communication between some given nodes of the network.

Let us locate point A anywhere in the first row of the matrix and point B – anywhere in the last row of the matrix. Percolation route created for points A and B goes at some angle, further on referred to as “angular displacement” relative to the matrix vertical line. In other words, this route goes along some centerline between points A and B. To avoid variation versus matrix size we shall normalize to the length of the guiding axis in the following way:

$$\varphi_T = \frac{L_R}{l}; l = \sqrt{(i_B - i_A)^2 + (j_B - j_A)^2}, \quad (4)$$

where  $L_R$  – the number of red cells added to the percolation route,  $l$  – geometric distance between points A and B calculated by the Pythagorean theorem, where  $(i_A; j_A)$  и  $(i_B; j_B)$  are coordinates of points A and B correspondingly.

Upon thorough study of various angular displacements it was found out that value  $\varphi_T$  was independent from the angular displacement. Then we constructed variation of the value  $\varphi_T$  versus concentration according to the following rule: points A and B should be located inside the clusters. We juxtaposed the obtained plot (see Fig. 10, plot b) with the plot in Figure 6b (see Fig.



10, plot a). The results agreed. Hence, the average number of objects (red cells) added to the percolation route would be the same irrespective to the direction of a percolation route.

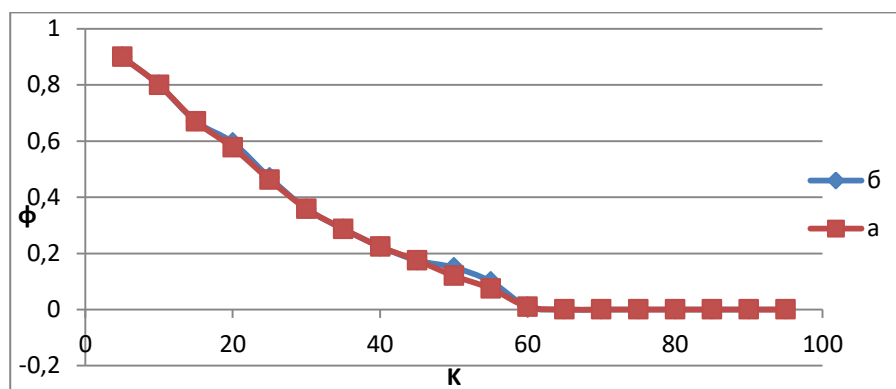


Fig. 10. a – variation of the average number of added objects  $\varphi(K)$  and average normalized length of a programmable percolation route versus concentration (Fig. 9); b – variation of the normalized length of a programmable percolation route versus concentration.

Note that graph 10a was plotted by averaging the number of the red cells in the situation of the percolation failure, whereas graph 10b was plotted by averaging the number of the red cells for programmable percolation between target points A and B.

## 8. Conclusion

1. In case of limited resources cost-effective planning of secure routes shall be two-phased: firstly is created a stochastic basis of secure though rather low-concentrated paths, and secondly are built secure routes via clusters of the stochastic basis with minimal insertion of secure paths in between the gaps of the clusters.

2. Concentration of secure paths in the stochastic basis shall be 0.25. At such concentration of secure nodes the number of the generated clusters is maximal. In this case any secure route built between the given nodes of the network has minimal average total cost.

3. Subsequent to the results of the percolation route stability analysis it was found that the fault phenomenon was of purely local nature and bypass routes were likely to revert to the original percolation route.

4. When the optimal concentration of secure paths is 0.25, the average number of additionally inserted secure paths to bypass the failed one is not more than 2.

## References

- [1] Moskalev P, Shitov V. Porous structures computer experiment. Moscow: Fizmatlit, 2007; 120 p. (in Russian)
- [2] Percolation: theory, application, algorithms: Reference Book. Edited by Tarasevich YuYu. Moscow: Editorial URSS, 2002; 109 p. (in Russian)
- [3] Golubev AS, Zvyagin MYu, Milovanov D. Percolation effect in information networks with unstable links. Bulletin of Lobachevsky State University of Nizhni Novgorod 2011; 2(3): 260–263.
- [4] Nekrasova AA Sokolov SS. Study of the possibility of percolation theory for flow control in information networks transport. Bulletin of Admiral Makarov State University of Maritime and Inland Shipping 2010; 32(4): 192–198.
- [5] Mostovoi YaA. Statistical phenomena in large-scale distributed clusters of nanosatellites. Vestnik of Samara University. Aerospace and Mechanical Engineering 2011; 26(2): 80–89.
- [6] Mostovoi YaA. Two-phase operation in large-scale networks of nanosatellites. Computer Optics 2013; 37(1): 120–130.
- [7] Mostovoi YaA. Programmable percolation and optimal two-phase operations in large-scale networks of nanosatellites. Infokommunikacionnye Tehnologii 2013; 11(1): 53–62.
- [8] Mostovoi YaA. Simulation of optimal two-phase operations in random operating environments. Avtometriya 2015; 51(3): 35–41.
- [9] Alexandrowicz Z. Critically branched chains and percolation clusters. Physics Letters A 1980; 80(4): 284–286.
- [10] Agrawal P, Redner S, Reynolds PJ, Stanley HE. Site-bond percolation: a low-density series study of the uncorrelated limit. J. Phys. A: Math. Gen. 1979; 12: 2073–2085.
- [11] Babalievski F. Cluster counting: the Hoshen-Kopelman algorithm vs. Spanning three approach. International Journal of Modern Physics 1998; 9(1): 43–61.
- [12] Galam S, Mauger A. Universal formulas for percolation thresholds. Phys. Rev. E 1996; 53(3): 2177–2181.
- [13] Hoshen J, Kopelman R. Phys. Rev. B 1976; 14: 3438–3445.
- [14] Sarshar N, Boykin PO, Roychowdhury VP. Scalable Percolation Search in Power Law Networks. Proceedings of the Fourth International Conference on Peer-to-Peer Computing. Zurich, 2004.
- [15] Stauffer D. Scaling theory of percolation clusters. Physics Reports 1979; 54: 1–74.
- [16] Stauffer D, Aharony A. Introduction to Percolation Theory. London: Taylor & Francis, 1992.
- [17] Vakulya G, Simon G. Energy Efficient Percolation-Driven Flood Routing for Large-Scale Sensor Networks. Proceedings of the International Multiconference on Computer Science and Information Technology. Wisla, Poland, 2008; 877–883.
- [18] Wilkinson D, Willemsen JF. Invasion percolation: A new form of percolation theory. J. Phys. A 1983; 16: 3365–3376.

# Computationally efficient methods of clustering ensemble construction for satellite image segmentation

I.A. Pestunov<sup>1</sup>, S.A. Rylov<sup>1</sup>, Yu.N. Sinyavskiy<sup>1</sup>, V.B. Berikov<sup>2</sup>

<sup>1</sup>*Institute of computational technologies SB RAS, 6 acad. Lavrentiev ave., 630090, Novosibirsk, Russia*

<sup>2</sup>*Sobolev institute of mathematics SB RAS, 4 acad. Koptug ave., 630090, Novosibirsk, Russia*

---

## Abstract

Combining multiple partitions into single ensemble clustering solution is a prominent way to improve accuracy and stability of clustering solutions. One of the major problems in constructing clustering ensembles is high computational complexity of the common methods. In this paper two computationally efficient methods of constructing ensembles of nonparametric clustering algorithms are introduced. They are based on the use of co-association matrix and subclusters. The results of experiments on synthetic and real datasets confirm their effectiveness and show the stability of the obtained solutions. The performance of the proposed methods allows to process large images including multispectral satellite data.

*Keywords:* ensemble; co-association matrix; nonparametric clustering algorithm; multispectral image segmentation

---

## 1. Introduction

Segmentation is one of the most important steps in the analysis of digital images [1]. It consists of dividing an image into non-overlapping regions based on similarity of their spectral and spatial characteristics (texture, size, shape, etc.). Segmentation methods have found wide application in many applied fields including Earth remote sensing [2], which is developing rapidly in recent years.

The most common approach to the segmentation of satellite images is based on data clustering algorithms [3]. Generally, clustering problem has to be solved without any *a priori* information about the number of clusters and their probabilistic characteristics. Under these conditions, the most attractive is nonparametric approach, which is famous for its ability to discover arbitrary-shaped clusters without hard assumptions on the density distribution function [4]. Nonparametric density-based algorithms define clusters as high density regions in the feature space separated by low density regions. Histogram and Parzen-window density estimates are the most popular for nonparametric clustering algorithms. The resulting estimate is highly dependent on the smoothing bandwidth parameter, which defines the scale of observation. Larger values result in smoother density estimate, while for smaller values the contribution of each sample to overall density has the emphasized local character, resulting in density estimate revealing details on a finer scale. Bandwidth selection in practical tasks is very complicated [5].

Ensemble approach is well known as a prominent method for improving robustness, stability and accuracy of clustering solutions [6-15]. Ensemble approach combines multiple partitions generated by clustering algorithms into a single consensual solution. Research on clustering ensembles focuses on two challenging aspects: how to generate different and diverse partitions and how to design consensus function. There are four common ways of generating multiple data partitions for clustering ensemble [6]: applying different clustering algorithms [7], applying the same clustering algorithm with different values of parameters or initializations [8], combining data representations (feature spaces) [9] and different subsets of initial dataset [10].

The optimal consensual result can be obtained by solving median partition problem [11]. In this case, one seeks to find the partition, which minimizes the sum of distances to all partitions in the ensemble with respect to a clustering distance function. The direct solution turns out to be NP-hard for any reasonable clustering distance function [12]. Therefore, in practice other ways of solving this problem are used (e.g. hypergraph partitioning, voting approach, mutual information algorithm, co-association based functions and finite mixture model). Nevertheless, all these methods suffer from high time complexity [7,8] and they cannot be applied directly to large images containing millions of pixels [10,13].

In this paper, two computationally efficient methods of constructing clustering ensembles based on co-association matrix and subclusters are proposed and compared. Besides, the use of co-association matrix for improving the stability of ensemble results is theoretically substantiated.

## 2. Description and theoretical foundation of the ensemble approach based on consensus co-association matrix

One of the most efficient approaches for constructing cluster ensemble is the use of consensus co-association matrix [15]. The elements of this matrix characterize the pairwise dissimilarity of objects as a number of partitions in which a pair do not belong to the same cluster. To obtain final solution, the co-association matrix is considered as a matrix of distances between objects. In this role, it is used as input for the one of the standard hierarchical clustering algorithms. This method does not require the equality of the number of clusters in all partitions. This condition is necessary for nonparametric clustering, when the number of clusters is not determined in advance.

In this work, we propose to construct an ensemble on the basis of  $L$  particular results obtained by nonparametric clustering algorithm with different values of bandwidth parameter. The method of the ensemble construction is described as follows.

Suppose that the set of classified objects consists of vectors in feature space  $R^d : X = \{x_i = (x_i^1, \dots, x_i^d) \in R^d, i = \overline{1, N}\}$ . Let a set of particular clustering results  $G = \{G^{(1)}, \dots, G^{(l)}, \dots, G^{(L)}\}$  be obtained with use of some algorithm  $\mu = \mu(\theta)$  depending on a vector of parameters  $\theta$  taken at random from certain set of parameters  $\Theta$ ; here  $G^{(l)}$  is the  $l$  th variant of partitioning on  $M^{(l)}$  clusters.

Denote by  $H(\theta_l)$  the binary matrix of size  $N \times N$  defined for the  $l$  th partition as follows:

$$H_{i,j}(\theta_l) = \begin{cases} 0, & \text{if the objects } x_i \text{ and } x_j \text{ are assigned to the same cluster,} \\ 1, & \text{otherwise,} \end{cases}$$

where  $i, j = \overline{1, \dots, N}$ ,  $i \neq j$ .

After constructing a set of particular clustering results, it is possible to determine the consensus co-association matrix

$$\mathbf{H} = \{\mathbf{H}_{i,j}\}, \quad \mathbf{H}_{i,j} = \frac{1}{L} \sum_{l=1}^L H_{i,j}(\theta_l),$$

where  $i, j = \overline{1, \dots, N}$ . The quantity  $\mathbf{H}_{i,j}$  equals the frequency of classifying  $x_i$  and  $x_j$  into separate clusters in the set of clusterings  $G$ . A value close to zero suggests that these objects have a large chance of falling into the same group. A value close to 1 indicates that the chance of being in the same cluster is negligible for the pair.

After calculating the consensus co-association matrix, hierarchical clustering algorithm UPGMA with unweighted average linkage rule is applied to find the collective clustering result [16]. This method has the advantage that it allows discovering hierarchical structure of clusters, what greatly simplifies the process of interpreting the results.

To study the properties of cluster ensemble construction method, we suggest the following probabilistic model.

Suppose that there exists a latent (directly unobservable) variable  $U$  that determines the belonging of each object to some of  $M \geq 2$  classes. Each class is characterized by certain conditional distribution  $p(x|U=r) = f_r(x)$ ,  $r = \overline{1, \dots, M}$ . Consider a model of data generation. Let an object's attributed class be determined in accordance with a priori probabilities  $P_r = P(U=r)$ ,  $r = \overline{1, \dots, M}$ , where  $\sum_{r=1}^M P_r = 1$ . Then according to distribution  $f_r(x)$  the value of  $x$  is obtained. This procedure is repeated independently for each object.

Let some clustering algorithm  $\mu$  be used to partition the dataset  $X$  into  $M$  subsets. Since the labeling of clusters do not matter, it is convenient to consider the equivalence relation, i.e. to indicate whether the algorithm assigns a pair of objects to the same class or not. For each pair of objects  $a$  and  $b$ , we define the value

$$\mathbf{H}_{a,b}(\mu) = \begin{cases} 0, & \text{if the objects are assigned to the same cluster,} \\ 1, & \text{otherwise,} \end{cases}$$

where  $a, b \in X$ ,  $a \neq b$ .

Let us choose an arbitrary pair  $a$  and  $b$  of different objects.

Let  $P_U = P(U(a) \neq U(b))$  be the probability of assigning the objects to different classes. For example, for  $M = 2$  this probability equals

$$P_U = 1 - P(U(a) = 1 | a)P(U(b) = 1 | b) - P(U(a) = 2 | a)P(U(b) = 2 | b) = 1 - \sum_{r=1}^2 \frac{f_r(a)f_r(b)P_r^2}{p(a)p(b)},$$

where  $p(\omega) = \sum_{r=1}^2 f_r(\omega)P_r$ ,  $\omega = a, b$ .

Denote by  $P_{er}(\mu)$  the probability of error for algorithm  $\mu$  in assigning  $a$  and  $b$  to different classes, where

$$P_{er}(\mu) = \begin{cases} P_U, & \text{if } \mathbf{H}_{a,b}(\mu) = 0, \\ 1 - P_U, & \text{if } \mathbf{H}_{a,b}(\mu) = 1. \end{cases}$$

One can easily notice that

$$P_{er}(\mu) = (1 - \mathbf{H}_{a,b}(\mu))P_U + \mathbf{H}_{a,b}(\mu)(1 - P_U) = P_U + (1 - 2P_U)\mathbf{H}_{a,b}(\mu).$$

Algorithm  $\mu$  depends on the random vector of parameters  $\Theta \in \Theta : \mu = \mu(\Theta)$ . To emphasize the dependence of the results on the parameter  $\Theta$ , in what follows we shall denote  $\mathbf{H}_{a,b}(\mu(\Theta)) = \mathbf{H}_{a,b}(\Theta)$ ,  $P_{er}(\mu(\Theta)) = P_{er}(\Theta)$ .

Let a collection  $H(\theta_1), \dots, H(\theta_L)$  be obtained by algorithm  $\mu$  running  $L$  times with randomly and independently selected parameters  $\Theta_1, \dots, \Theta_L$ . For definiteness, we may assume that  $L$  is odd. The function

$$\mathbf{H}(H(\theta_1), \dots, H(\theta_L)) = \begin{cases} 0, & \text{if } \frac{1}{L} \sum_{l=1}^L H(\theta_l) < \frac{1}{2}, \\ 1, & \text{otherwise} \end{cases}$$

will be called a collective (ensemble) decision by majority voting for a pair of objects. Within the framework of the model described above, the following properties of the suggested collective decision are fulfilled [15].

**Proposition 1.** Mathematical expectation and variance of the value of error probability for algorithm  $\mu(\Theta)$  are equal, respectively, to:

$$E_{\Theta} P_{er}(\Theta) = P_U + (1 - 2P_U)P_H,$$

$$\text{Var}_{\Theta} P_{er}(\Theta) = (1 - 2P_U)^2 P_H(1 - P_H),$$

where  $P_H = P(H(\Theta) = 1)$ .

Denote by  $P_{er}(\Theta_1, \dots, \Theta_L)$  the random function which for fixed arguments takes a value equal to the probability of error in classifying  $a$  and  $b$  by the ensemble algorithm. Here we denote by  $\Theta_1, \dots, \Theta_L$  independent statistical copies of random vector  $\Theta$ . Consider the behavior of error probability for collective decision.

**Proposition 2.** Mathematical expectation and variance of the value of error probability for collective decision are equal, respectively, to:

$$E_{\Theta_1, \dots, \Theta_L} P_{er}(\Theta_1, \dots, \Theta_L) = P_U + (1 - 2P_U)P_{H,L},$$

$$\text{Var}_{\Theta_1, \dots, \Theta_L} P_{er}(\Theta_1, \dots, \Theta_L) = (1 - 2P_U)^2 P_{H,L}(1 - P_{H,L}),$$

where  $P_{H,L} = P\left(\frac{1}{L} \sum_{l=1}^L H(\Theta_l) \geq \frac{1}{2}\right) = \sum_{l=[L/2]+1}^L C_L^l P_H^l (1 - P_H)^{L-l}$ ,  $[\cdot]$  denotes the integer part.

We use the following a priori information about cluster analysis algorithm. We assume that the expected probability of erroneous classification  $E_{\Theta} P_{er}(\Theta) < 1/2$ . That is, it is believed that the algorithm  $\mu$  classifies with better quality than the algorithm of random equiprobable choice. It follows from Proposition 1 that one of two variants should be fulfilled: a)  $P_H > 1/2$  and  $P_U > 1/2$ ; b)  $P_H < 1/2$  and  $P_U < 1/2$ . Let us consider, for definiteness, the first case.

**Proposition 3.** If  $E_{\Theta} P_{er}(\Theta) < 1/2$ , and at that  $P_H > 1/2$  and  $P_U > 1/2$ , then with increasing an ensemble size, the expected probability of erroneous classification decreases, tending to the limit of  $1 - P_U$ , and the variance of the value of error probability tends to zero.

The last statement allows to conclude that under the fulfillment of quite natural requirements, the usage of the ensemble-based approach will improve the quality of clustering.

### 3. Computationally efficient methods of clustering ensemble construction

Constructing ensemble solution based on consensus co-association matrix requires formation and processing of the square matrix of size  $N \times N$  ( $N$  is the number of elements). This method is not applicable for satellite image segmentation due to its quadratic complexity. This problem can be overcome by processing groups of elements instead of single elements. The way of grouping and choosing representatives may depend on algorithm specific features. Two methods of data grouping in order to construct consensus co-association matrix are proposed below.

The first method allows combining results obtained by arbitrary clustering algorithm. Having  $L$  partitions, each data element  $x_i$  can be associated with a label vector  $c_i = (c_i^1, \dots, c_i^L)$  where  $c_i^j$  is a label of cluster containing  $x_i$  in  $j$ th partition. Data elements with the same label vectors are merged into subclusters, because corresponding elements of co-association matrix (the distance between such elements) is zero. Subclusters are labeled as related data elements and will act as data elements while constructing of consensus co-association matrix. The distance between subclusters labeled as  $c_i$  and  $c_j$  is defined as

$$H_{i,j} = \frac{1}{L} \sum_{l=1}^L I(c_i^l \neq c_j^l), \text{ where } I(a) = \begin{cases} 1, & \text{if } a \text{ is true;} \\ 0, & \text{otherwise.} \end{cases}$$

In this case, the size of co-association matrix is equal to the number of subclusters, which can vary from the maximum number of clusters in partitions to their product.

The second method allows to construct an ensemble solution for nonparametric mode-seeking clustering algorithms. In this case, each cluster contains one or more local density maxima (modes). Data elements are first divided into subclusters corresponding to single modes. Modes are used as representatives of subclusters and will act as data elements while constructing consensus co-association matrix. All elements of the subcluster are labeled as corresponding mode. Consensus co-association matrix is formed on the set of representatives from the basic partition (the most detailed partition in the ensemble). The correspondence between different partitions is established by determining clusters containing the representatives from the basic partition (as points in the feature space). Therefore, the size of co-association matrix is much less than  $N$  and equal to the number of modes in base partition.

Fig. 1 illustrates the second method and shows two different partitions ( $L = 2$ ) of the same data into clusters (highlighted by colors) that correspond to some density modes. First partition is considered basic and it contains three representatives (A, B, C). When establishing the correspondence between partitions, representatives A and B are assigned to one cluster in the second partition and representative C – to another. Thus, the distance between representatives A and B equals  $1/2$  (since they lie in different clusters in the basic partition) while the distances between pairs A, B and B, C are equal to  $2/2$ .

The proposed methods were used to construct an ensemble of nonparametric clustering algorithms: MeanSC (based on Parzen density estimation), CCA and HCA (based on the histogram density estimation). Ensemble algorithms EMeanSC [17] and HECA [18] were designed using the first and the second ensemble construction method respectively. Basing on clustering algorithm CCA two ensemble algorithms were developed: CCAE (using first method) and ECCA [19] (using second method). The implementation of both proposed methods based on the same clustering algorithm allows to compare these methods.

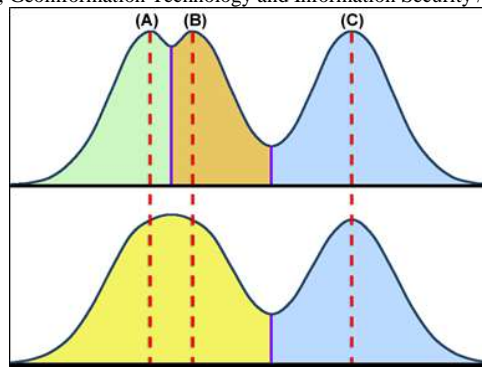


Fig. 1. Illustration of the second method of clustering ensemble construction using representatives.

#### 4. Experimental results

Experimental results show that proposed clustering ensemble construction methods improve the quality and the stability of obtained results. Moreover, they significantly simplify algorithm parameters selection. Computational efficiency of the proposed methods allows processing large satellite images.

Experimental results on both synthetic and real datasets were obtained on Intel Core i7 3.2 GHz quad-core CPU.

**Experiment 1.** Two different two-dimensional synthetic datasets [20] were clustered. First dataset consists of 3 classes containing 1000 points each with uniformly distributed “bridge” (200 points) and uniformly distributed noise (1000 points). Fig. 2 shows initial data (Fig. 2a) and the result of EMeanSC ensemble clustering algorithm (Fig. 2b). Correct data decomposition with MeanSC clustering algorithm requires precise parameter setting while ensemble algorithm successfully detects 4 clusters and noise (black points in Fig. 2) even with bad interim results (Fig. 2e). Second synthetic dataset (“bananas”) was built in PRTools toolbox [21] with parameter  $\sigma = 0.7$ . It consists of 400 points and represents 2 linearly inseparable classes (Fig. 2c). Ensemble clustering result of EMeanSC algorithm is shown in Fig. 2d.

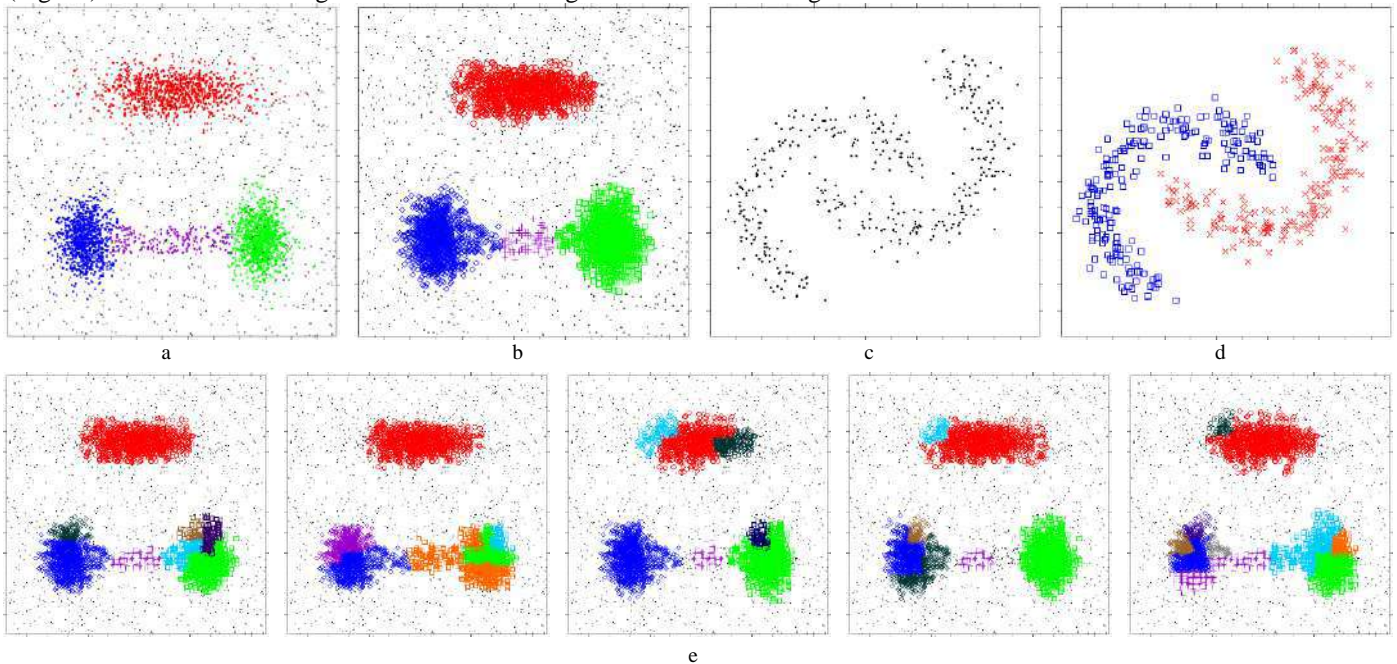


Fig. 2. Experiment 1: synthetic datasets (a,c) and clustering results obtained by ensemble EMeanSC algorithm (b,d) and MeanSC algorithm with different bandwidth parameter values (e).

**Experiment 2.** Two-dimensional synthetic dataset containing 5 classes [20] was clustered. Three classes represent normal distribution with mathematical expectation vectors  $\mu_1 = (188,100)$ ,  $\mu_2 = (75,100)$ ,  $\mu_3 = (75,150)$  and covariance matrices  $\Sigma_1 = \begin{pmatrix} 4^2 & 0 \\ 0 & 4^2 \end{pmatrix}$ ,  $\Sigma_2 = \begin{pmatrix} 12^2 & 0 \\ 0 & 12^2 \end{pmatrix}$ ,  $\Sigma_3 = \begin{pmatrix} 21^2 & 0 \\ 0 & 8^2 \end{pmatrix}$  respectively. Fourth class represents uniformly distribution along the ring with center at  $(188,100)$  and radiuses  $R_{\min} = 20$  and  $R_{\max} = 25$ . Fifth class represents uniform distribution along the circle with center at  $(188,100)$  and radius  $R = 45$ . Elements of the fifth class were further radially displaced by a random value having normal distribution with standard deviation  $\sigma = 4$ . Clusters containe 220, 600, 600, 400 and 500 points respectively (Fig. 3a). There we have a generated reference partition (Fig. 3a), so the accuracy of clustering is determined as the percentage of correctly classified elements. Each class from the reference partition is associated with a cluster (one or none) containing the largest number of elements from this class.



Fig. 3b represents results obtained by ECCA clustering algorithm ( $L=8$ ). The accuracy of clustering is 99.48%. Clustering algorithms from open source package ELKI [22] could not correctly separate all 5 classes. The best results (Fig. 3c and 3d) were obtained by density-based hierarchical algorithm OPTICS (79.7% for parameters  $\epsilon=12$ ,  $minpts=8$  and hierarchy cutoff level 4.9) and hierarchical nearest neighbor algorithm SLINK (72.72% for parameter  $threshold=5$ ).

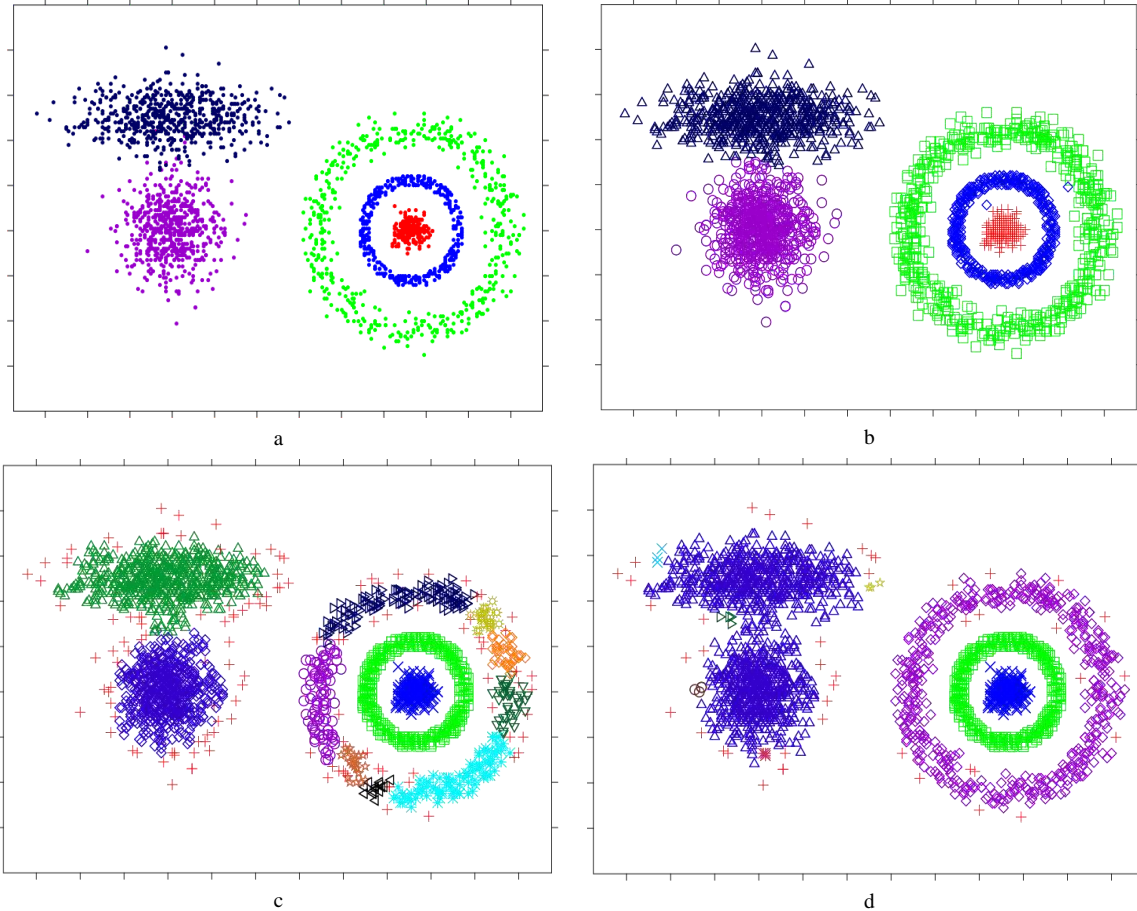


Fig. 3. Experiment 2: synthetic dataset (a) and results obtained by ECCA (b), OPTICS (c) and SLINK (d) clustering algorithms.

**Experiment 3.** The purpose of this experiment is to demonstrate an increasing stability of the ensemble clustering results with ensemble size ( $L$ ) growth. The accuracy of clustering “bananas” dataset (see Fig. 2c) with respect to grid parameter  $m$  (bandwidth) for  $CCA(m, T)$  algorithm and corresponding ensemble algorithms  $ECCA(m, L, T)$  and  $CCA(m, L, T)$  with  $L=5$  and  $L=10$  is shown in Fig. 4. Parameter  $T$  was fixed at value 0.3. As it can be seen from the graph, the stability of the results increases with ensemble size growth.

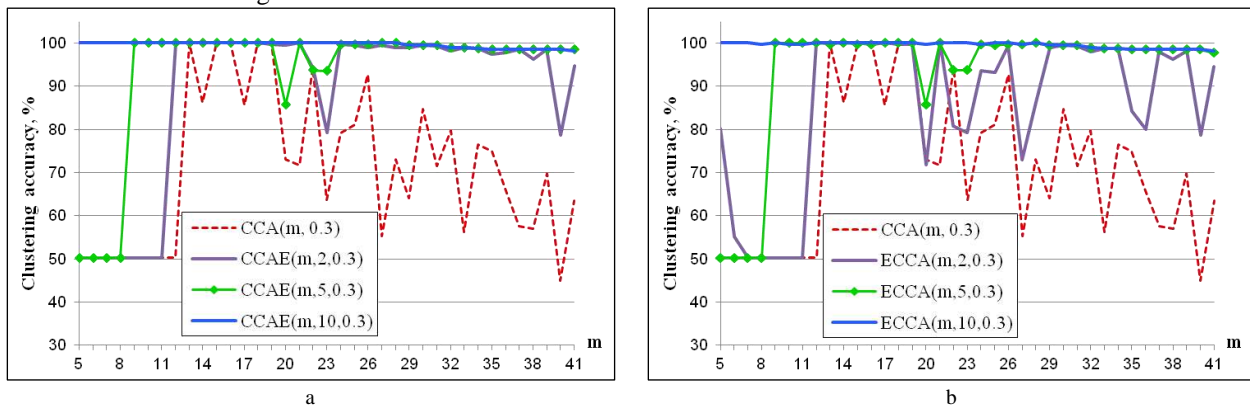


Fig. 4. Graph of the clustering accuracy with respect to grid parameter for CCA, CCAE (a) and ECCA (b) algorithms.

**Experiment 4.** Fig. 5 shows the result of WorldView-2 satellite image clustering (Burmistrovo, Novosibirsk region). Image size is  $2048 \times 2048$  pixels; four spectral bands (1, 3, 5 and 8) were used. Image was processed by ECCA algorithm (with  $L=8$ ) in 0.5 seconds. This experiment confirms the ability of the developed ensemble algorithm to effectively process multispectral satellite images.

**Experiment 5.** The purpose of this experiment is to demonstrate computational efficiency of the proposed ensemble construction methods and to compare them. Six images were used (Fig. 5a and 6) [23], including 3 satellite ones (WorldView-2 and Landsat-8, 4 bands selected). Table 1 shows the time of constructing clustering ensemble containing 8 elements with CCAE and ECCA algorithms. The time for 8 CCA runs (in parallel) performed before the ensemble construction is shown separately.

The results show that second ensemble constructing method based on the use of density-modes (ECCA algorithm) leads to much smaller size of consensus co-association matrix that substantially reduces processing time.



Fig. 5. RGB-composite (bands 5, 3, 2) of the WorldView-2 satellite image (a) and segmentation result obtained by ECCA clustering algorithm (b).

Table 1. Comparison of the proposed ensemble construction methods (the time is in seconds).

Image size (megapixels)	Number of bands	Time for 8 CCA runs	Subclusters number, CCAE	CCAEnsemble construction time	Subclusters number, ECCA	ECCA ensemble construction time
1	3	0.08	1697	0.3	155	0.005
5	3	0.3	2199	2.8	161	0.02
14	3	0.6	3520	4.9	79	0.03
4	4	0.5	9975	5.6	1278	0.06
12	4	0.9	4640	2.1	1414	0.1
50	4	2.5	9625	32	1518	0.2



Fig. 6. Color pictures and images from WorldView-2 and Landsat-8 satellites.

## 5. Conclusion

In this paper, two methods of constructing clustering ensemble based on co-association matrix are proposed and theoretically substantiated. Developed ensemble clustering algorithms are based on nonparametric density estimates and they don't make hard assumptions on the density distribution function. Experimental results on both synthetic and real datasets confirm high quality of the obtained solutions and their stability. Unlike common ensemble construction methods, the proposed approaches can be directly applied to large satellite images (containing millions of pixels) and demonstrate high performance.

## Acknowledgements

This work was partially supported by Presidium of the RAS (project 0316-2015-0006).

## References

- [1] Gonzalez RC, Woods RE. Digital image processing. Moscow: Tekhnosfera, 2006; 812 p. (in Russian)
- [2] Dey V, Zhang Y, Zhong M. A review on image segmentation techniques with remote sensing perspective. Proceedings of ISPRS TC VII Symposium – 100 Years ISPRS. Vienna: ISPRS, 2010; XXXVIII(7A): 31–42.
- [3] Pestunov IA, Sinyavsky YuN. Clustering algorithms in problems of segmentation of satellite images. Bulletin KemSU 2012; 52(4/2): 110–125. (in Russian).
- [4] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters 2010; 31(8): 651–666.
- [5] Krstinic D, Skelin AK, Slapnicar I. Fast two-step histogram-based image segmentation. Image Processing, IET 2011; 5(1): 63–72.

- [6] Ghaemi R, Sulaiman M, Ibrahim H, Mustapha N. A survey: clustering ensembles techniques. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 2009; 3(2): 365–374.
- [7] Kashef R, Kamel M. Cooperative clustering. *Pattern Recognition* 2010; 43(6): 2315–2329.
- [8] Hore P, Hall LO, Goldof DB. A scalable framework for cluster ensembles. *Pattern Recognition* 2009; 42(5): 676–688.
- [9] Strehl A, Ghosh J. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research* 2003; 3: 583–617.
- [10] Jia J, Liu B, Jiao L. Soft spectral clustering ensemble applied to image segmentation. *Frontiers of Computer Science in China* 2011; 5(1): 66–78.
- [11] Franek L, Jiang X. Ensemble clustering by means of clustering embedding in vector spaces. *Pattern Recognition* 2014; 47(2): 833–842.
- [12] Filkov V, Skiena S. Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools* 2004; 13(4): 863–880.
- [13] Tasdemir K, Moazzen Y, Yildirim I. An approximate spectral clustering ensemble for high spatial resolution remote-sensing images. *IEEE Journal, Selected Topics in Applied Earth Observations and Remote Sensing* 2015; 8(5): 1996–2004.
- [14] Berikov V, Pestunov I. Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties. *Pattern Recognition* 2017; 63: 427–436.
- [15] Berikov VB. Construction of an ensemble of decision trees in the cluster analysis. *Computational Technologies* 2010; 15(1): 40–52. (in Russian)
- [16] Gronau I, Moran S. Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters* 2007; 104(6): 205–210.
- [17] Pestunov IA, Berikov VB, Sinyavskiy YuN. Segmentation of multispectral images based on an ensemble of nonparametric clustering algorithms. *Vestnik SibGAU* 2010; 5(31): 56–64. (in Russian)
- [18] Pestunov IA, Rylov SA, Berikov VB. Hierarchical clustering algorithms for segmentation of multispectral images. *Optoelectronics, Instrumentation and Data Processing* 2015; 51(4): 329–338.
- [19] Pestunov IA, Kulikova EA, Rylov SA, Berikov VB. Ensemble of clustering algorithms for large datasets. *Optoelectronics, Instrumentation and Data Processing* 2011; 47(3): 245–252.
- [20] Rylov SA. Model datasets. URL: <https://drive.google.com/open?id=0ByK9GtU5ExExRnZwdFNmRHRWdFk> (30.05.2017).
- [21] A Matlab Toolbox for Pattern Recognition. URL: <http://www.prtools.org> (30.05.2017).
- [22] Schubert E, Koos A, Emrich T, Züfle A, Schmid KA, Zimek A. A framework for clustering uncertain data. *Proceedings of the VLDB Endowment* 2015; 8(12): 1976–1979.
- [23] Image datasets for clustering. URL: <https://drive.google.com/open?id=0ByK9GtU5ExExWXpGRjU5WVfHcDg> (30.05.2017).



# Edge Detection in Remote Sensing Images Based on Fuzzy Image Representation

E.V. Pugin<sup>1</sup>, A.L. Zhiznyakov<sup>1</sup>

<sup>1</sup>Vladimir State University named after Alexander and Nikolay Stoletovs, Gorky Street 87, Vladimir, Russia

---

## Abstract

Edge detection is an important task in image processing. There are a lot of approaches in this area: Sobel, Canny operators and others. One of the perspective techniques in image processing is the use of fuzzy logic and fuzzy sets theory. They allow us to increase processing quality by representing information in its fuzzy form. Most of the existing fuzzy image processing methods switch to fuzzy sets on very late stages, so this leads to some useful information loss. In this paper a novel method of edge detection based on fuzzy image representation and fuzzy pixels is proposed. With this approach we convert the image to fuzzy form on the first step. Different approaches to this conversion are described. Several membership functions for fuzzy pixel description and requirements for their form and view are given. A novel approach to edge detection based on Sobel operator and fuzzy image representation is proposed. Experimental testing of developed method was performed on remote sensing images. Comparison of result with Sobel, Prewitt, Roberts and Canny operators is presented. Developed method selected more details (edges) rather than Sobel, Prewitt and Roberts operators, but less than Canny operator. This is because the selected convolution kernel (Sobel) has size 3x3. There are also used only simple functions of estimating the real intensities of pixels. Later, to increase quality it is necessary to use more complex masks of size 5x5 and 7x7 or median filters. Developed approach showed its workability in solving image processing problems. The proposed fuzzy model in the future can be extended to use higher level fuzzy sets (Type-2 FS and others).

*Keywords:* edge detection; fuzzy features; fuzzy image representation; fuzzy sets

---

## 1. Introduction

To extract information from remote sensing image different methods of image processing are used. Edge detection is one of these methods. It can be used to further extraction of interesting objects. There are a lot of algorithms developed in this area. The most popular are Canny, Sobel, Prewitt, Roberts operators and some others [1, 2, 3]. One of the perspective techniques in image processing is the use of fuzzy logic and fuzzy sets theory [4, 5, 6]. Different algorithms and methods use different approaches to fuzzy processing. Linguistic variables are useful when the results are well distinguishable [7, 8, 9], but this is not very common situation. On the other hand it is better to process the results using analytical methods. Most papers suggest switching to fuzzy sets on very late stages of processing when the source image is enhanced, converted or somehow preprocessed. Papers that extract fuzzy properties based on existing crisp features can be taken to this category [10, 11]. Early defuzzification of the results has also negative influence on fuzzy processing [12]. All these things lower flexibility of fuzzy approach.

Other drawback is the use of fuzzy sets of first type (Type-1 Fuzzy Sets or T1FS), that have very little amount of uncertainty. To solve this issue type-2 fuzzy sets (T2FS) and other more high-type fuzzy sets were introduced [13, 14]. Also, other types of sets based on fuzzy sets are evolving: rough sets, soft sets, soft rough sets, blurry sets and others [15, 16].

In this paper we propose a novel approach to image processing based on fuzzy sets, that suggest a transition to fuzzy image representation with fuzzy pixels on the earliest stages of processing. With this approach all extracted features are fuzzy by definition and defuzzification process at best should be done only when retrieving information from the computer system.

## 2. Fuzzy image representation

Continuous image can be described as two-dimensional signal  $f(x, y)$ , where  $x$  and  $y$  – coordinates. During formation of digital image transition to discrete coordinates and values of intensities are performed:

$$F(x, y) = D[f(x, y)],$$

where  $D[\cdot]$  – transformation operator from continuous signal to discrete, that is implemented on hardware,  $F(x, y)$  – the discrete image. Obviously that with insufficient level of quantization, signal levels are rounded to integers. Then real intensity level of point  $R(x, y)$  could be computed as

$$R(x, y) = F(x, y) + d(x, y),$$

where  $d(x, y)$  – round error. More noticeable distortion is brought by different noises  $v(x, y)$ , that could be greater than 1. We will use the simplest noise model – additive noise. More complex models of noise could be investigated accordingly. Then the real intensity level could be calculated as

$$R(x, y) = F(x, y) + d(x, y) + v(x, y).$$

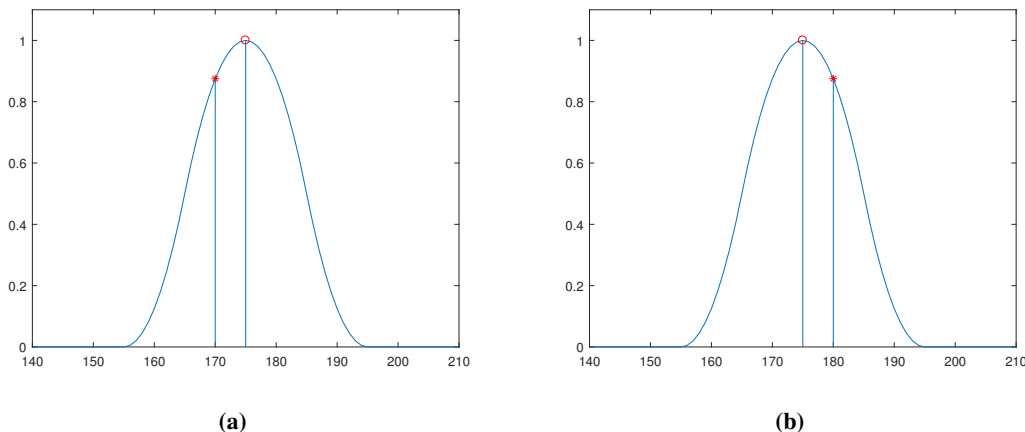


Fig. 1. Membership function of image fuzzy pixel. Real intensity level at  $\mu(F(x, y)) = 1$ : a) greater, b) less than current.

Computations that does not take into account these details, soon could accumulate big error related to source continuous image. Methods of fuzzy sets theory allow us to save the uncertainty till the latest stages of image processing and analysis. To do it we should switch to fuzzy image representation.  $U(F, x, y)$

$$U(F, x, y) = \mu(F(x, y)),$$

where  $\mu(F(x, y))$  – membership function of a pixel with coordinates  $(x, y)$  to intensity level  $F(x, y)$ . Graphically this can be represented as shown on Fig. 1.

There are a lot of membership functions known. One must select those which satisfy the following conditions

$$\lim_{l \rightarrow \infty} \mu(F(x, y)) = 0,$$

$$\int_0^{L-1} \mu(F(x, y)) dl > 0, \quad l \in [0; L - 1].$$

These include triangular, trapezoidal, bell, Gauss-like,  $\pi$  functions and others. In the simplest case we will be using  $\pi$ -function which is based on  $s$ -function:

$$\pi(l) = \begin{cases} s(l, c - b, c - \frac{b}{2}, c), & l \leq c, \\ 1 - s(l, c, c + \frac{b}{2}, c + b), & l \geq c, \end{cases}$$

$$s(l) = \begin{cases} 0, & l \leq a, \\ 2(\frac{l-a}{c-a})^2, & a \leq l \leq b, \\ 1 - 2(\frac{l-c}{c-a})^2, & b \leq l \leq c, \\ 1, & l \geq c, \end{cases}$$

where  $b = \frac{a+c}{2}$ ,  $l$  – intensity level. Form of  $\pi$ -function is shown on Fig. 1.

To pick  $\mu(F(x, y)) = \pi(F(x, y))$  function, that is, pick  $b$  and  $c$  parameters, it is necessary to extract some additional information from the image. Firstly, let us simplify this task reducing the selection to single parameter  $I_c$  — center of membership function  $\pi(I_c) = 1$ . In this case  $\pi$ -function lies symmetrically on this point. To set slope inclination, we must choose necessary width of  $w$  section, where  $\pi(x) > 0$ . Then let us use the following equation to find parameters of function  $\pi$ :

$$b = \frac{w}{2}, \quad c = I_c.$$

Value of  $w$  is chosen empirically, e.g.  $w = 60$ . With small value of  $w$  slope will be very big, and small intensity deviations will make the pixel insignificant ( $\pi < 0.5$ ).

Parameter  $I_c$  can be selected differently. In our case  $\pi(I_c) = 1$  means real intensity of the pixel, and  $\pi(F(x, y)) \neq 1$  shows interference, noise, errors in quantization and similar errors. Value of  $I_c$  can be computed using neighbor of point  $(x, y)$ , that is shown on Fig. 2. Let us consider some possible approaches:

1. average between horizontal pixels

$$I_c = (P_4 + P_8)/2,$$

2. average between vertical pixels

$$I_c = (P_2 + P_6)/2,$$

3. average in 4-neighbour  $D_4$

$$I_c = (P_2 + P_4 + P_6 + P_8)/4,$$

$P_1$	$P_2$	$P_3$
$P_8$	$P_0$	$P_4$
$P_7$	$P_6$	$P_5$

Fig. 2. Neighbour of pixel  $P_0$  with coordinates  $(x, y)$ .

-1	-2	-1
0	0	0
1	2	1

(a)

-1	0	1
-2	0	2
-1	0	1

(b)

Fig. 3. Sobel kernels: a) horizontal, b) vertical.

4. average in 8-neighbour  $D_8$

$$I_c = \frac{1}{8} \sum_{i=1}^8 P_i,$$

5. average in d-neighbour  $D_d$

$$I_c = (P_1 + P_3 + P_5 + P_7)/4,$$

In more complex case to get  $I_c$  value one could use one of the existing smoothing methods like median filters, approximations etc.

### 3. Edge detection

Most of the edge detection operators uses gradient operator that has modulo  $|\nabla G|$  and direction  $\theta$

$$|\nabla G| = \sqrt{G_x^2 + G_y^2},$$

$$\theta = \arctan \frac{G_y}{G_x},$$

where  $G_x = M_x * G$ ,  $G_y = M_y * G$  – the result of convolution operator with horizontal and vertical matrices.

Let us consider possibility of the use of fuzzy pixels in Sobel and Prewitt edge detection operator (Fig. 3, 4). To do this we must process the membership function values  $\pi(F(x, y))$  in addition to normal intensity levels. Calculations of  $G_{x\mu}$  and  $G_{y\mu}$  can be done analogously. The difference here is in gradient computations of  $\pi$ -function.

$$|\nabla \pi| = 1 - \sqrt{G_{x\mu}^2 + G_{y\mu}^2}.$$

We take complementary value because after squaring, sum and square root operations from values on interval  $[0; 1]$ , the result is near to 0. Final value of gradient will be

$$|\nabla G_\pi| = |\nabla G| |\nabla \pi|.$$

Threshold value of gradient can be selected manually or with different binarization techniques (e.g. with Otsu method [17]).

-1	-1	-1
0	0	0
1	1	1

(a)

-1	0	1
-1	0	1
-1	0	1

(b)

Fig. 4. Prewitt kernels: a) horizontal, b) vertical.

#### 4. Testing

Let us consider transition to fuzzy pixels. Test remote sensing image has size  $160 \times 160$ . Simple averaging procedures described above were applied to it. Result images are shown on Fig. 5, according membership functions are shown on Fig. 6 and the values are put in Table 1. Comparison of different methods is shown on Fig. 7.

Proposed method extracted more details (edges) than Sobel, Prewitt and Roberts operators, but less than Canny operator. This is because the selected convolution kernel (Sobel) has size  $3 \times 3$ . Also we used only simple functions during real pixel intensity estimation. To increase quality of the results more complex masks ( $5 \times 5$ ,  $7 \times 7$ ) could be used and median filters also.

On processed images we can see that proposed method found a lot of "islands", which in the original image are green plantings. Rivers were marked well. Presence of the big number of details after the use of Canny operator in some cases could bring some issues during further steps, so additional filtering may be applied. In proposed method number of details less than after Canny operator and this could be useful. Later, selected edges could be used in segmentation and object detection algorithms [18, 19].

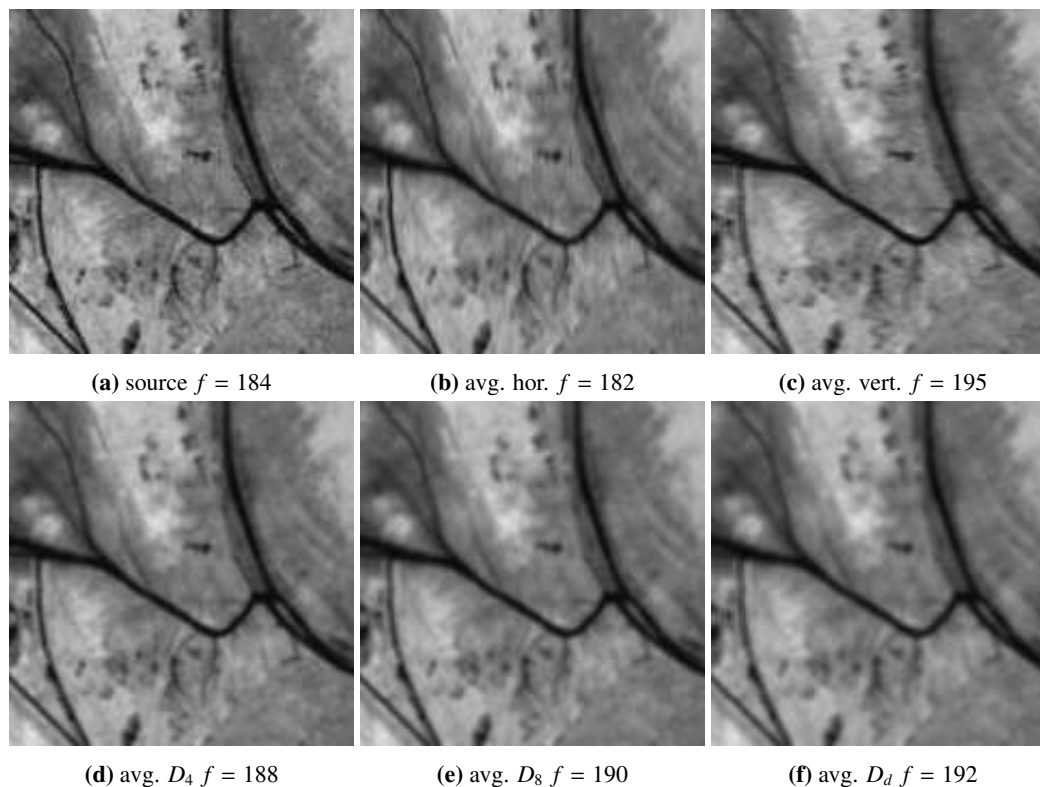


Fig. 5. Remote sensing image of a river. Intensity level of the pixel  $f = F(70, 55)$  during different approaches to  $I_c$  calculation.

Table 1. Intensities of pixels and according values of membership functions

Image	$F(70, 55)$	$\pi(F(70, 55))$
Source	184	1
Avg. hor.	182	0.9911
Avg. vert.	195	0.7311
Avg. $D_4$	188	0.9644
Avg. $D_8$	190	0.92
Avg. $D_d$	192	0.8578

#### 5. Conclusion

Proposed algorithm showed their applicability in edge detection task during image processing. Main feature is the use of fuzzy image representation based on fuzzy pixels. This approach is very perspective because it saves the uncertainty much better rather other existing algorithms. In the future this model could be extended to type-2 and higher fuzzy sets and also to other kinds of fuzzy sets.

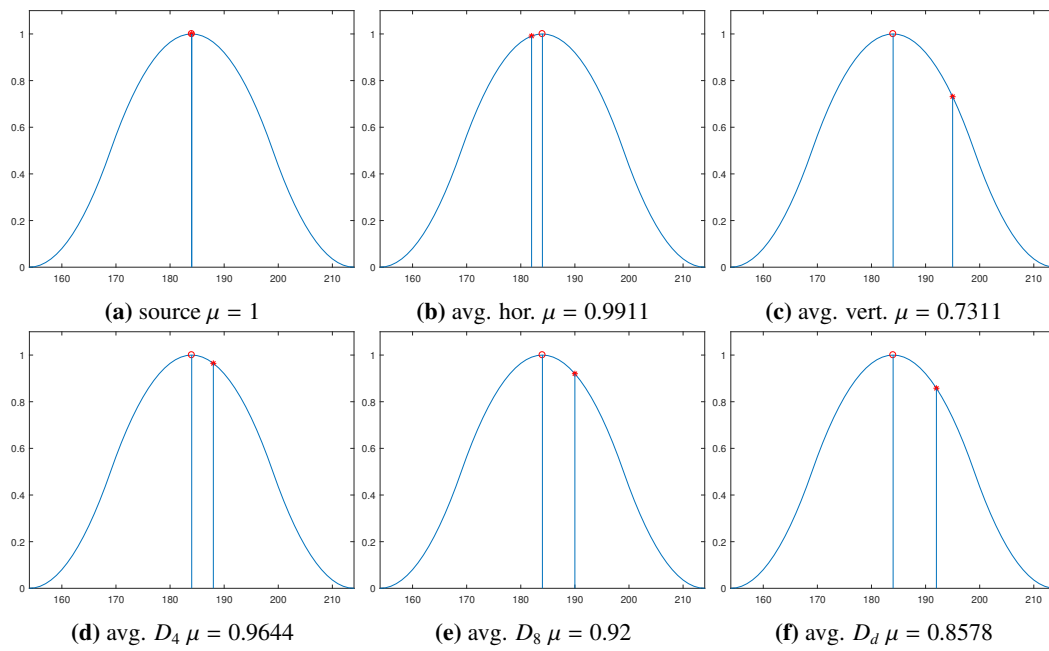


Fig. 6. Value of membership function  $\mu = \mu(F(70, 55))$  during different approaches to  $I_c$  calculation.

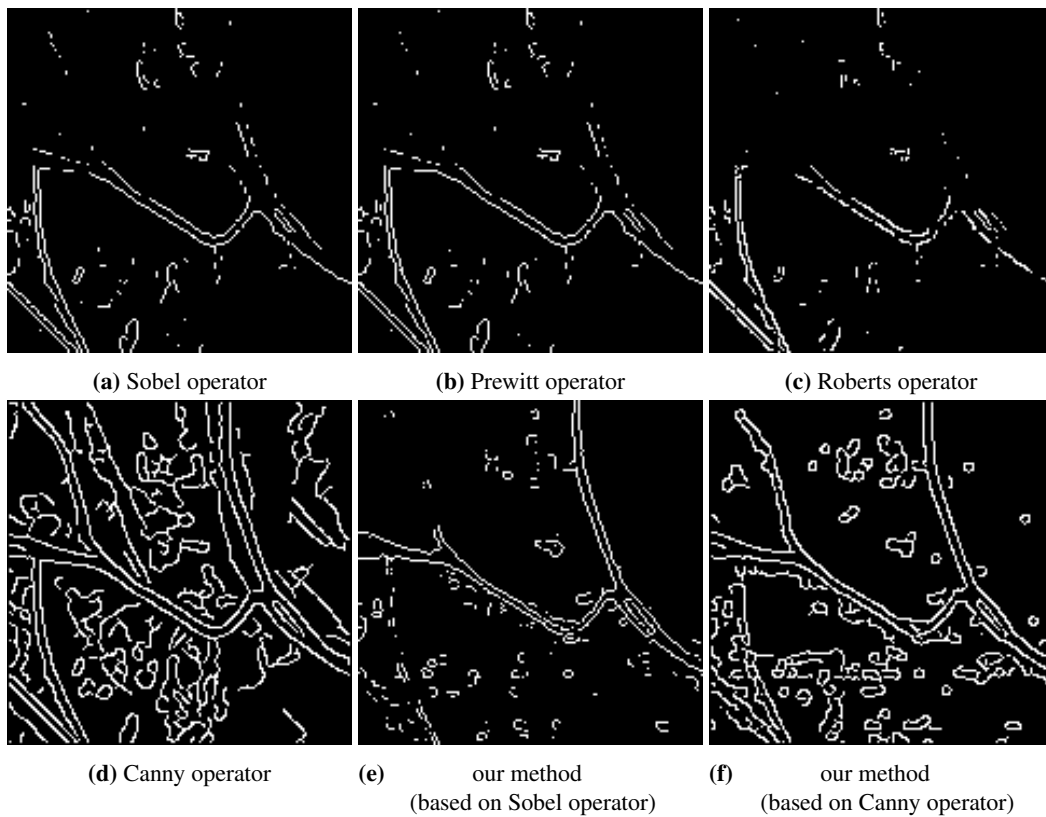


Fig. 7. Application of different edge detection operators.

## References

- [1] Canny, J. A computational approach to edge detection / John Canny // *Pattern Analysis and Machine Intelligence*, IEEE Transactions on. — 1986. —11. —Vol. PAMI-8, no. 6. —P. 679–698.
- [2] Kanopoulos, N. Design of an image edge detection filter using the sobel operator / N. Kanopoulos, N. Vasanthavada, R.L. Baker // *IEEE Journal of Solid-State Circuits*. — 1988. —apr. —Vol. 23, no. 2. —P. 358–367.
- [3] Haralick, R. M. Digital step edges from zero crossing of second directional derivatives / Robert M. Haralick // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. — 1984. —jan. —Vol. PAMI-6, no. 1. —P. 58–68.
- [4] Kuo, Y.-H. A new fuzzy edge detection method for image enhancement / Yau-Hwang Kuo, Chang-Shing Lee, Chao-Chin Liu // *Proceedings of 6th International Fuzzy Systems Conference*. —[S. l.] : Institute of Electrical and Electronics Engineers (IEEE), 1997.
- [5] El-Khamy, S. A modified fuzzy sobel edge detector / S.E. El-Khamy, M. Lotfy, N. El-Yamany // *Proceedings of the Seventeenth National Radio Science Conference*. 17th NRSC 2000 (IEEE Cat. No.00EX396). —[S. l.] : Institute of Electrical and Electronics Engineers (IEEE), 2002.
- [6] Melin, P. An improved method for edge detection based on interval type-2 fuzzy logic / Patricia Melin, Olivia Mendoza, Oscar Castillo // *Expert Systems with Applications*. —2010. —dec. —Vol. 37, no. 12. —P. 8527–8535.
- [7] Becerikli, Y. A new fuzzy approach for edge detection / Yasar Becerikli, Tayfun M. Karan // *Computational Intelligence and Bioinspired Systems*. — [S. l.] : Springer Nature, 2005. —P. 943–951.
- [8] A novel fuzzy ant system for edge detection / Om Prakash Verma, Madasu Hanmandlu, Ashish Kumar Sultania, Dhruv // *2010 IEEE/ACIS 9th International Conference on Computer and Information Science*. —[S. l.] : Institute of Electrical and Electronics Engineers (IEEE), 2010. —aug.
- [9] Hu, L. A high performance edge detector based on fuzzy inference rules / Liming Hu, H.D. Cheng, Ming Zhang // *Information Sciences*. —2007. —nov. —Vol. 177, no. 21. —P. 4768–4784.
- [10] Russo, F. Edge detection in noisy images using fuzzy reasoning / F. Russo // *IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference. Where Instrumentation is Going (Cat. No.98CH36222)*. — [S. l.] : Institute of Electrical and Electronics Engineers (IEEE), 1998.
- [11] Interval type-2 fuzzy sets constructed from several membership functions: Application to the fuzzy thresholding algorithm / Miguel Pagola, Carlos Lopez-Molina, Javier Fernandez [et al.] // *IEEE Transactions on Fuzzy Systems*. —2013. —apr. —Vol. 21, no. 2. —P. 230–244.
- [12] *Fuzzy Techniques in Image Processing* / Ed. by Etienne E. Kerre, Mike Nachtgaeel. —[S. l.] : Physica-Verlag HD, 2000.
- [13] Edge-detection method for image processing based on generalized type-2 fuzzy logic / Patricia Melin, Claudia I. Gonzalez, Juan R. Castro [et al.] // *IEEE Transactions on Fuzzy Systems*. —2014. —dec. —Vol. 22, no. 6. —P. 1515–1525.
- [14] An improved sobel edge detection method based on generalized type-2 fuzzy logic / Claudia I. Gonzalez, Patricia Melin, Juan R. Castro [et al.] // *Soft Computing*. —2014. —dec. —Vol. 20, no. 2. —P. 773–784.
- [15] Molodtsov, D. Soft set theory first results / D. Molodtsov // *Computers & Mathematics with Applications*. — 1999. — Vol. 37, no. 4. —P. 19 – 31. — URL: <http://www.sciencedirect.com/science/article/pii/S0898122199000565>.
- [16] Pawlak, Z. Rough sets / Zdzislaw Pawlak // *International Journal of Computer and Information Sciences*. — 1982. —oct. — Vol. 11, no. 5. — P. 341–356. —URL: <http://dx.doi.org/10.1007/BF01001956>.
- [17] Otsu, N. A threshold selection method from gray-level histograms / N Otsu // *IEEE Trans. Sys., Man., Cyber*. — 1979. —Vol. 9. —P. 62–66.
- [18] Privezentsev, D. G. Use of characteristic image segments in tasks of digital image processing / Denis G. Privezentsev, Arkady L. Zhiznyakov // *2015 International Conference "Stability and Control Processes" in Memory of V.I. Zubov (SCP)*. —[S. l.] : Institute of Electrical and Electronics Engineers (IEEE), 2015. —oct.
- [19] Zhiznyakov, A. L. Using fractal features of digital images for the detection of surface defects / A. L. Zhiznyakov, D. G. Privezentsev, A. A. Zakharov // *Pattern Recognition and Image Analysis*. —2015. —jan. —Vol. 25, no. 1. —P. 122–131.

# Compressing deep convolutional neural networks in visual emotion recognition

A.G. Rassadin<sup>1</sup>, A.V. Savchenko<sup>1</sup>

<sup>1</sup>National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, 25/12 Bolshaya Pecherskaya Street, 603155, Nizhny Novgorod, Russia

---

## Abstract

In this paper, we consider the problem of insufficient runtime and memory-space complexities of deep convolutional neural networks for visual emotion recognition. A survey of recent compression methods and efficient neural networks architectures is provided. We experimentally compare the computational speed and memory consumption during the training and the inference stages of such methods as the weights matrix decomposition, binarization and hashing. It is shown that the most efficient optimization can be achieved with the matrices decomposition and hashing. Finally, we explore the possibility to distill the knowledge from the large neural network, if only large unlabeled sample of facial images is available.

*Keywords:* deep learning; convolutional neural networks; deep compression; visual emotion recognition; deep compression; binarized neural networks; tensor decomposition; SqueezeNet; XNOR-Net; distilling the knowledge of neural network

---

## 1. Introduction

Emotion recognition in the wild has many potential applications in various information systems with man-machine interaction. Emotions can be automatically extracted from voice [1], text [2] and body language. However, one of the most practical way of classifying of human emotions is the usage of facial expressions in either video or still images seems to be one of the major research directions in the area of image recognition [3] – [5]. It is known that contemporary deep convolutional neural networks (CNNs) [6] – [8] cause much more accurate solutions than the traditional techniques. However, their runtime complexity becomes insufficient for application in practical tasks, especially with implementation on mobile platforms. For example, the size of the file with the neural model trained on EmotiW [9] dataset is approximately equal to 475 Mb [10]. Moreover, it is impossible to classify images with this model faster than 10 FPS even on common laptop. At the same time, the most exciting applications of emotion recognition appear in mobile hardware. Hence, the performance optimization of deep CNN is now considered as one of the most important studies in deep learning.

The most remarkable research direction in this field is the optimization of algorithms and neural network architectures. For instance, the work on the CNNs compression [11] received the Best Paper Award in very prestigious International Conference on Learning Representation (ICLR'16). To compare various methods, we will use such goals as recognition accuracy, and space (memory) complexity of the training and inference procedures. It is also important to take into account GPU total training time and average inference time.

The visual emotion recognition problem is particularly difficult because there does not exist a large database of training images. In this context, it is worth mentioning the EmotiW challenge [9], which provides one of the most famous datasets playing a key role in the growth of the field. Unfortunately, this dataset is not publicly available. Thus, in this paper we, firstly, selected the most promising and effective optimization possibilities introduced in the papers from the last year. Secondly, we examine the possibility to distill the knowledge [12] of large CNN [10] trained on the EmotiW dataset [9] by classifying the images from unlabeled facial dataset in order to train the most efficient CNN architecture.

The rest of the paper is organized as follows. In Section 2, we provide a survey of recent literature devoted to the performance improvements of deep neural networks. Section 3 contains an experimental study of performance optimization methods in visual emotion recognition within the already done model. Section 4 explains the possibility to build a powerful yet efficient model by distilling the knowledge on architecture independent basis. Finally, concluding comments are given in Section 5.

## 2. Review of CNN compression techniques

There are several types of classification of deep neural networks performance optimization methods, which can differ: by

- 1) *accuracy loss*: lossless, optimization with accuracy loss, optimization-accuracy trade-off;
- 2) *applicability level*: architectural, operational (by model / framework modification), computational (exactly while training or inference), hardware;
- 3) *limitations*: architecture-dependent, and architecture-independent;
- 4) *implementation*: runtime implementation, two-step (training -> optimization), sequential (training -> optimization -> re-training);
- 5) *optimization building block*: all blocks, convolutional layers, fully connected layers.

Perhaps the most fundamental approach and in the same time one of the most efficient and universal is the *pruning* [11], [13]. It is known that in huge amount of weights (connections) in the trained network even with the superior generalization ability the

contribution to every neuron (connection) is different. One can alternately remove connections with low (by absolute value) weight and, in turn, minimal impact to the prediction results, and fine-tune after every pruning, until achieving the allowable loss in the accuracy. It is important to note that the pruning can be applied to any neural network architecture and both before and after every another performance optimization technique being the most general approach.

*Distilling the knowledge* technique has been initially suggested by Hinton et al. in 2014 [12]. The idea of this approach is to train a cumbersome neural model or an ensemble of models with the superior generalization ability (“teacher”) and then transfer its predictive power to another, thinner but usually deeper model (“student”) by training the latter to predict the same labels as the original one. The main disadvantage of this approach is that the time cost for the optimization is on the same level with the training from scratch. Another obvious drawback is that the “student” model is not protected from the mistakes of the “teacher” model. In fact, the resulted model is even weaker, because it can tends to unexpected behavior in predictions. Moreover, it is not an absolute performance optimization but rather relative to the original teacher network. This technique have not become widely used. We can only mention the work of Romero et al. [14] in which some limitations of the initial approach were overcome.

The idea of *weights hashing (quantization)* [15] is based on that close values of the CNN weights may be considered equal (with some precision), which makes it possible to share the same memory unit, and, in turn, drastically reduce the memory costs. This approach continues to exploit the idea of lower precision computations. It is exactly the key part of the famous Deep Compression method [11], in which a very effective pipeline to optimize the performance and the size of the network is described. Unfortunately, it is hard to distill from the paper the real influence of quantization to overall compression quality, because it also includes pruning, which is the most important factor, which allowed the authors to achieve their outstanding results in compression of AlexNet [16] architecture.

The *tensor decomposition* exploits a very intuitive idea: since that deep neural network contains high order matrices (tensors) of weights in each layer, they can be decomposed to the sequence of lower order matrices and vectors. The most popular techniques nowadays are CP (CANDECOMP/PARAFAC or Canonical Polyadic Decomposition) [17], Tucker [18] and the most recent one – Tensor Train [19], [20]. Such approaches allow to explicitly choose between the amount of memory consumption and the accuracy loss by setting the rank of the decomposition.

The group of *binarization* methods is based on the observation that it is to enough for weights to be stored in FP32 and continues the trend of lower precision computations. These techniques differs from the hashing (quantization) by going deeper into performance optimization problem caring out not only about the storage and native (because of lower precision) computational efficient. Original idea is followed by the observation that only 1 bit ( $\{0, 1\}$  or  $\{-1, +1\}$ ) is enough for weights values. Thus, it is possible to store only the sign of values instead of usage of full FP32 precision. Hence, the arithmetic operations can be replaced to much faster logical operations. However, the binarization of the network right after traditional training leads to the complete loss of the predictive power of the network. It is important to apply binarization iteratively, epoch-by-epoch. The procedure of binary weights backpropagation was suggested in [21] to implement this approach. Initial idea to binarize only weights outgrew to binarizing the whole network including the input vector. Such an approach [22] leads to the complete replacement of the arithmetic operations by XNOR. It has recently been shown [23] that the applied binary mapping does not matter, hence, the sign of the variable is usually the simplest and fastest technique.

There exist other methods, which optimize a fixed architecture or even already learned model by using several *architectural tricks*. Among these methods, it is important to mention the SqueezeNet [24] and the Tiny Darknet [25], which achieve the accuracy compared to the AlexNet [16], but are much smaller and even faster. The PVANet [26] is the architecture for the object detection task with minimal computational cost obtained by adapting and combining recent technical innovations. The BranchyNet [27] introduces early exits (classifiers) along the architecture by which the researcher can explicitly balance between the inference speed and the accuracy.

To summarize our brief survey, we present in Table 1 the potential of the most important discussed methods to achieve four optimization goals, which we mentioned in introduction. As we can see here, despite of the large number of reviewed papers, there are no “silver-bullet” methods, which guarantee the training speedup or memory consumption while training. Most of these techniques dedicated on reduction the memory consumption while inference. The pruning can be very common approach, e.g. integrated in every modern framework but unfortunately, it is still not common. Next, we consider a set of experiments similarly to [28].

### 3. Experimental results

In this section, we will discuss the computational experiments dedicated on performance optimization power of already done model using the following techniques: HashedNet, BWN, XNOR-Net and CP-decomposition. We have used the author’s code that guarantees us the exact implementation and results reproducibility. These methods were evaluated on the real task of emotion recognition from facial images detected in the widely used Radboud Faces Database (RaFD) [29] which contains pictures of eight emotional expressions: anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. Each emotion was shown with three different gaze directions and all images were taken from five camera angles simultaneously. In our experiments, we omitted profile images and use only frontal faces and pictures with  $45^\circ$  rotation. The neural networks were trained from scratch using identical training samples and learning procedures. Inspired by the well-known facial expression recognition CNN [10], we choose the VGG-S architecture for the HashedNet, BWN and XNOR-Net as a baseline. All the neural



network models are freely available at (<https://mega.nz/#F!2FVz1SAT!dRdzpfc7UEwHC-jI9jEkIQ>). In fact, in [10] authors trained an ensemble of neural networks using RGB and different kind LBP (Local Binary Patterns) visual features, but we decided to use a single RGB input for the simplicity. All the experiments were done on the same machine using Tesla M2090 with 6 GB of memory under Ubuntu 14.04 with CUDA Toolkit 8.0.

Table 1. Reported results of deep neural networks performance optimization.

	Memory reduction while training	Memory reduction while inference	Inference speedup	Baseline model	Dataset
Deep Compression [11]	no	49	unknown	VGG-16	ImageNet
FitNets [14]	no	36	13.36	Maxout	CIFAR-10
HashedNets [15]	no	64	unknown	same-size	MNIST
CP-Decomposition [17]	no	12	4.5	AlexNet	ImageNet
TensorNet [20]	unknown	80	unknown	simple	CIFAR-10
BinaryNet [21]	~32 (theoretical)	~32 (theoretical)	3.4~23	Maxout	CIFAR-10
Binary-Weight-Network and XNOR-Net [22]	no	67	58 (CPU)	ResNet-18	ImageNet
SqueezeNet [24]	unknown	50	1.	AlexNet	ImageNet
Tiny Darknet [25]	unknown	60	2.9	AlexNet	ImageNet
BranchyNet [27]	no	no	1.9	ResNet-110	CIFAR-10

The HashedNets, BWN and XNOR-Net have been trained using RGB images from the same distinct and balanced training / testing subsets of the RaFD [29] dataset using the Torch framework. We used SGD with momentum equal to 0.9, learning rate fixed at 0.001 and mini-batch of 20 sample. The common baseline model [10] was trained with the same settings. This baseline CNN converged to accuracy 97.13% after 100 epochs (Fig. 1). Here and bellow, the testing error rate is in practically all cases less than the training error rate. Though such behavior seems to be not obvious, it is reasonable due to the usage of dropout regularization layer, which is activated while training phase and deactivated when evaluating on the validation set. Moreover, the training error rate is computed as the mean error rate for all mini-batches in one epoch. On the contrary, the testing error rate is computed only after each epoch with more optimal weights, which were learned during this epoch. Let us compare this result with the performance optimization techniques.

We used default compression settings, provided by the authors of the HashedNet technique [15]: compression rate is equal to 0.125 and the bias hashing was set. The latter option leads to the 81.64% reduction in the weights count. Despite this reduction, the training process (Fig. 2) is practically identical to the baseline (Fig. 1): the network converged to 96.31% accuracy after 100 epochs, which is 0.8% lower when compared to the baseline CNN (Fig. 1). However, the training procedure is 6.7 times slower when compared to the baseline. The inference procedure of the HashedNet is also 4.7 times slower. We believe that such slowdown can be drastically reduced by replacing the current third-party implementation of hashing, which does not allow us saving trained model and measure memory consumption while inference accurately.

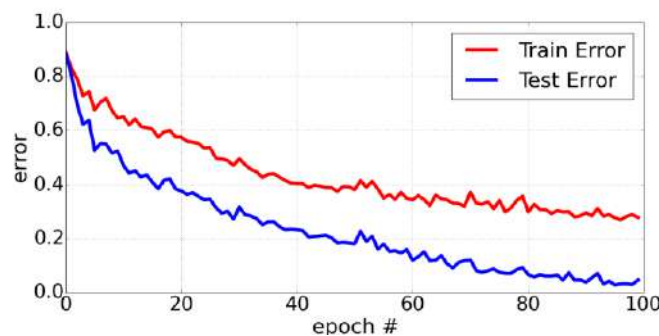


Fig. 1. The training/testing error rates for the baseline VGG-S neural network model.

The testing of the CP-decomposition [17] was performed using the SqueezeNet-1.1 [24] architecture instead of VGG-S (Fig. 3). Indeed, convolutional layers take a small portion of weights in such architectures with massive fully connected layers, as the VGG-S. Hence, the CP-decomposition is appropriate only for such convolutional architectures without fully connected layers as the SqueezeNet. The baseline model was trained with Caffe framework using stochastic gradient descent (SGD) with momentum 0.9, fixed learning rate 0.001 and 32 images in a mini-batch. To compare the neural networks computing efficiency we measured: 1) epoch time for single forward pass and subsequent gradient update on GPU for mini-batch in one random sample, averaged over 1000 runs; and 2) GPU inference time for single random sample, averaged over 1000 runs. We additionally estimated the accuracy loss and the reduction in number of weights. Original version of the SqueezeNet-1.1 architecture has relatively small number of filters in every convolutional layer. Hence, the decomposition of every layer to a lower rank, e.g., 16, tends to the complete loss in accuracy. However, when only two last convolutional layers were decomposed with the rank equal to 192, the number of parameters reduced at 23.5% with 1.65% of the accuracy loss (from 89.14% to 87.5%). Unfortunately, the inference in the resulted network became even 1.5 times slower. It seems that replacement of the

single large convolutional layer to four sequentially connected small layers causes higher computing complexity in parallel environment.

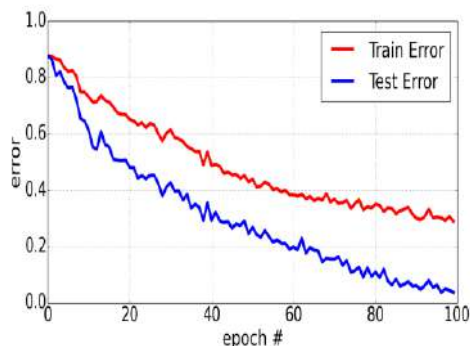


Fig.2. The training/testing error rates for the HashedNet.

In next experiments, the BWN and the XNOR-Net are implemented according to the paper [22]. Every *conv-bn-activation* block excluding the first was replaced with the *bn-activation-conv* block. The dependences of the testing and training error rates of the BWN on the epoch number are shown in Fig. 4. Here the BWN converged to the very low error rate 1.43% after forty epochs. After that time both training and testing error rate started to grow. We cannot precisely explain this behavior but probably, advanced learning rate policy can suppress this binarization shortcoming. In fact, all our experiments demonstrate that BWN model always converges 2-4 times faster, when compared to the baseline CNN, which can be explained by very strong regularization effect introduced by the BWN architecture. We have not observed the inference memory reduction or inference speedup. The number of parameters also remains unchanged.

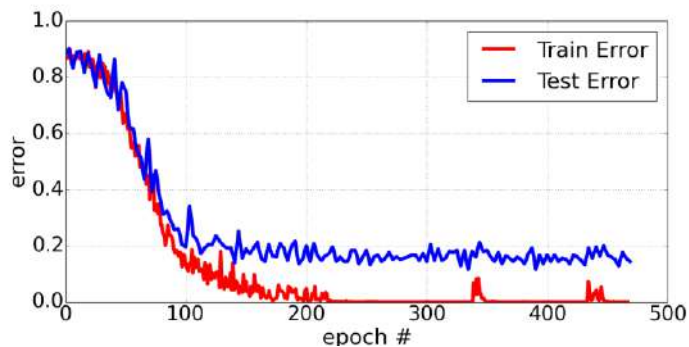


Fig.3. The training/testing error rates for the SqueezeNet.

The XNOR-Net [22] was not converged in our experiments (Fig. 5). The lowest error rate for the testing set was equal to 41.19%. The only advantage of this method is the slight (2.4%) reduction in the memory consumption while inference, which is the benefit of the modified binarized activation layer. It is interesting to note that using only binarized activation layer without weights binarization leads to the same parameters reduction and even slight epoch time speedup. What is more important, such modification is capable to converge much closer to the accuracy of the baseline model – 88.32% – within the same learning procedure (Fig. 6).

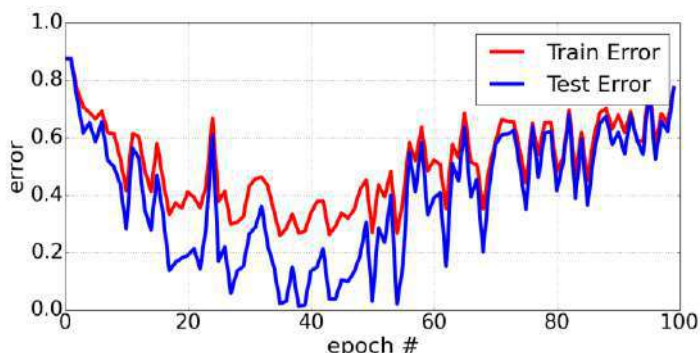


Fig. 4. The training/testing error rates for the BWN.

All results of these experiments are briefly summarized in Table 2. The best value in each column is marked by bold. Here in “Model size” column we count only the minimum amount of memory needed to store all weights of the CNN. In fact, the real

size of the file with the model can be much larger. For instance, the real size of the baseline VGG-S model is equal to 474 MB, i.e., it is approximately 100 MB larger than the model size in Table 2. The best accuracy achieves with the BWN technique, while the SqueezeNet outperforms other networks by the model size and execution times.

Table 2. Summary of evaluation results of CNN compressing methods.

	Training time per one epoch, ms	Inference time, ms	Model size, MB	Accuracy, %
VGG-S (baseline)	43.7	33.4	372.2	97.13
SqueezeNet-1.1 (baseline)	<b>22.94</b>	<b>4.94</b>	2.8	89.14
SqueezeNet-1.1, CP-Decomposition	<b>22.94</b>	7.74	<b>2.1</b>	87.5
HashedNets	294.8	158.2	68.3	96.31
Binary-Weight-Network (BWN)	83.8	33.5	11.6	<b>98.57</b>
XNOR-Net	84.3	34.2	11.6	58.81
XNOR-Net w/o weights activation	43.4	34.1	11.6	88.32

#### 4. Distilling the knowledge of neural network in unsupervised environment

Let us consider the well-known practical case of visual emotion recognition, when the large training dataset is unavailable. However, there exist several pre-trained large CNN models, which do not satisfy the requirements of space complexity and run-time efficiency. Due to lack of original or suitable dataset it is impossible to directly implement compact architecture described above. Hence, in this section we examine the potential of distilling the knowledge [12], [14] of these CNNs using one of the known face datasets, which are widely applied in face recognition tasks.

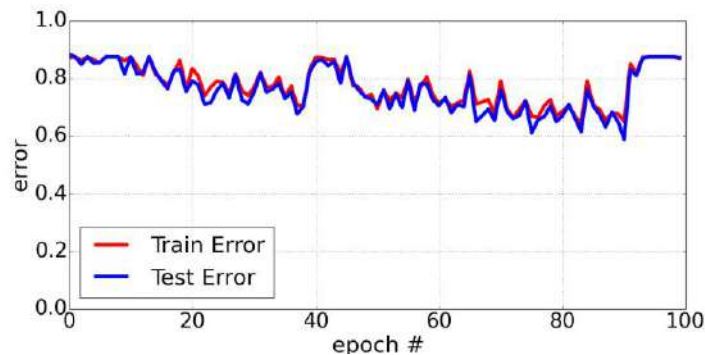


Fig. 5. The training/testing error rates for the modified XNOR-Net w/o weights activation.

The main disadvantage of the distilling the knowledge technique from paper [12] is its strong dependence on the network architecture. However, Tramèr et al. [30] have shown that probably any classifier of multimedia data can be reproduced based only on the labels, which are returned by this classifier for images from large enough dataset, even if nothing is known about its internal structure (architecture or even a kind of model). Hence, we can train an arbitrary architecture (small-size and efficient network like SqueezeNet [24]) using labels obtained by the existing (large) CNN or even an ensemble of such networks. This problem is the special case of unsupervised learning, because images from these available datasets usually do not contain the emotion labels.

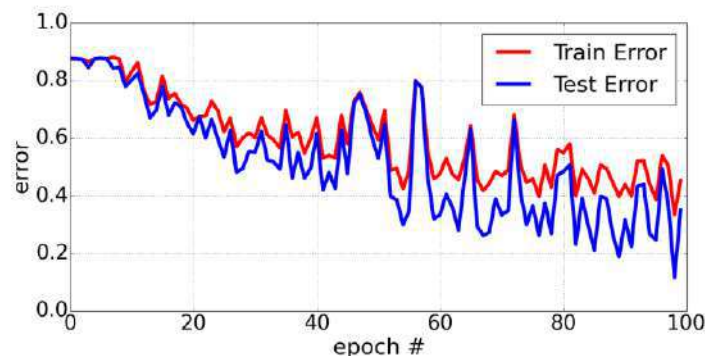


Fig. 6. The training/testing error rates for the modified XNOR-Net.

In this paper, we propose to extend this idea and train the small network using not only the labels predicted by large (“teacher”) CNN, but the vectors of posterior probabilities of all emotion classes at the output of softmax layer of this network. The loss function is defined as the Kullback-Leibler divergence (KLD) between these posterior probabilities and the output of the softmax layer of the trained small (“student”) CNN. It is expected that having also the scoring for each label can drastically improve the accuracy of the system. This architecture was implemented using Keras framework with Theano backend (<https://github.com/arassadin/cnn-compression>). The sketch of this network is shown in Fig. 7.

This architecture (hereinafter “softmax outputs”) is experimentally compared with the traditional (“label-only”) by architecture-independent distillation the knowledge. Rather large VGG-S network was used as a teacher, and trained the most promising architecture discussed in the previous section, namely, SqueezeNet-1.1 model [24]. Due to the lack of computational resources, the VGG-S knowledge was distilled on 13813 facial images from PubFig83 dataset [31]. The resulted architecture was tested with the RaFD [29] dataset. However, unlike the previous section, here we examine all images from this set with either frontal or profile orientation. We used two VGG-S teacher models, namely, the publicly available model from pre-trained [10] on EmotiW [9] dataset, and our own model trained directly on the RaFD dataset. The accuracies of these models on the whole RaFD testing set are approximately equal to 41.45% and 81%, respectively. The estimated accuracies of resulted (SqueezeNet) CNNs using either training or testing datasets are presented in Table 3.

Table 3. Experimental results of knowledge distillation.

	VGG on EmotiW		VGG on RaFD	
	Label-only	Softmax outputs	Labels-only	Softmax outputs
Training (PubFig83) accuracy, %	66.5	<b>73</b>	75.5	<b>77</b>
Testing (RaFD) accuracy, %	12.3	<b>23.8</b>	40.9	<b>46.9</b>

This experiment shows the strong domination of the learning on posterior probabilities at the softmax layer (Fig. 7) over the traditional (labels-only) approach. However, we cannot consider the experiment with VGG (EmotiW) model very representative due to very low accuracy rate (23.8% for the softmax outputs and 12.3% for labels only). Such behavior can be explained by the very low capabilities of the initial model (near the 40% accuracy according to the paper [10]). However, it is very revealing that labels-only accuracy is on rate of random guessing while the accuracy of the proposed architecture (Fig. 7) is almost twice higher. Another teacher network allowed labels-only training the small model achieving near the 41% of accuracy. At the same time the model trained on the softmax outputs was able to achieve near the 47% of accuracy rate. Such two simple experiments show the potential of the knowledge distillation via the training on both labels and softmax of the large (“teacher”) architecture.

## 5. Conclusion

In this paper, we have reviewed several modern approaches to reduce the space requirements and run-time complexity of deep CNNs in the problem of visual emotion recognition based on facial expressions. We emphasized the obvious trends in this field, namely, efficient tensor (or CP) decomposition techniques, lower precision calculations and more accurate network binarization. It was experimentally shown, that the most promising CNN performance optimization methods include the usage of special architectures, e.g., SqueezeNet [24], and binarization techniques [22], [23]. Additional set of experiments was intended to demonstrate the potential of the knowledge distillation methods using the pre-trained large CNN as a teacher network, which allows training a small CNN even with limited computational resources and the absence of the massive specialized datasets.

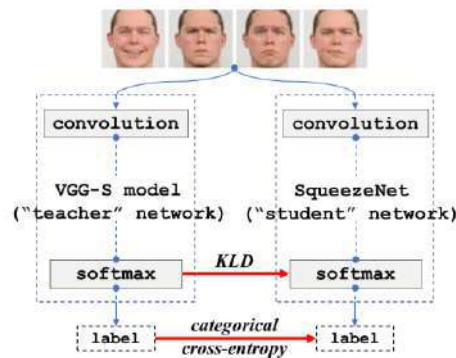


Fig. 7. The sketch of the proposed architecture: distillation the knowledge from large CNN with matching of posterior probabilities at the softmax outputs along with the labels correspondence.

The main direction for further research will be concentrated on combining of the most successful reviewed techniques. It is important to test these methods with other datasets, e.g., in the group-level emotion recognition in the EmotiW 2017 challenge. Another research direction is the implementation of the complete pipeline to video-based emotion recognition [9]. Finally, it is necessary to examine the possibility to implement discussed methods in image recognition on mobile platforms.

## Acknowledgments

The work was conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

## References

[1] Fayek HM, Lech M, Cavedon L. Towards real-time Speech Emotion Recognition using deep neural networks. Proceedings of the International Conference on Signal Processing and Communication Systems (ICSPCS) 2015: 1–5.

- [2] Socher R, Perelygin A, Wu J, Chuang J, Manning C, Ng A, Potts C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013; p. 1642.
- [3] Kahou SE, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K, Jean S, Froumenty P, Dauphin Y, Boulanger-Lewandowski N, Ferrari RC, Mirza M, Warde-Farley D, Courville A, Vincent P, Memisevic R, Pal C, Bengio Y. EmoNets: Multimodal deep learning approaches for emotion recognition in video, 2015. ArXiv preprint arXiv:1503.01800.
- [4] Ruiz-Garcia A, Elshaw M, Altahan A, Palade V. Deep Learning for Emotion Recognition in Faces. Proceedings of the International Conference on Artificial Neural Networks and Machine Learning (ICANN). Lecture Notes in Computer Science 1988; 7: 38–46.
- [5] Kahou SE, Michalski V, Konda K, Memisevic R, Pal C. Recurrent Neural Networks for Emotion Recognition in Video. Proceedings of the ACM International Conference on Multimodal Interaction 2016; 467–474.
- [6] Lin M, Chen Q, Yan S. Network In Network, 2013. ArXiv preprint arXiv:1312.4400.
- [7] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions, 2014. ArXiv preprint arXiv:1409.4842.
- [8] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, 2015. ArXiv preprint arXiv:1512.03385.
- [9] Emotion Recognition in the Wild Challenge. URL: <https://sites.google.com/site/emotiwchallenge/>.
- [10] Levi G, Hassner T. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. Proceedings of the ACM International Conference on Multimodal Interaction (ICMI) 2015; 503–510.
- [11] Han S, Mao H, Dally WJ. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, 2015. ArXiv preprint arXiv:1510.00149.
- [12] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network, 2015. ArXiv preprint arXiv:1503.02531.
- [13] Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning Convolutional Neural Networks for Resource Efficient Transfer Learning, 2016. ArXiv preprint arXiv:1611.06440.
- [14] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for Thin Deep Nets, 2014. ArXiv preprint arXiv:1412.6550.
- [15] Chen W, Wilson JT, Tyree S, Weinberger KQ, Chen Y. Compressing Neural Networks with the Hashing Trick, 2015. ArXiv preprint arXiv: 1504.04788.
- [16] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NIPS) 2012; 25: 1106–1114.
- [17] Lebedev V, Ganin Y, Rakhuba M, Oseledets I, Lempitsky V. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition, 2014. ArXiv preprint arXiv:1412.6553.
- [18] Kim Y-D, Park E, Yoo S, Choi T, Yang L, Shin D. Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications, 2015. ArXiv preprint arXiv:1511.06530.
- [19] Novikov A, Podoprikin D, Osokin A, Vetrov D. Tensorizing Neural Networks, 2015. ArXiv preprint arXiv:1509.06569.
- [20] Garipov T, Podoprikin D, Novikov A, Vetrov D. Ultimate Tensorization: Compressing Convolutional and FC Layers Alike, 2016. ArXiv preprint arXiv:1611.03214.
- [21] Courbariaux M, Hubara I, Soudry D, El-Yaniv R, Bengio Y. Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1, 2016. ArXiv preprint arXiv:1602.02830.
- [22] Rastegari M, Ordonez V, Redmon J, Farhadi A. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks, 2016. ArXiv preprint arXiv:1603.05279.
- [23] Merolla P, Appuswamy R, Arthur J, Esser SK, Modha D. Deep Neural Networks Are Robust to Weight Binarization And Other Non-linear Distortions, 2016. ArXiv preprint arXiv:1606.01981.
- [24] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level Accuracy With 50x Fewer Parameters And <0.5MB Model Size, 2016. ArXiv preprint arXiv:1602.07360.
- [25] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [26] Hong S, Roh B, Kim K-H, Cheon Y, Park M. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection, 2016. ArXiv preprint arXiv:1608.08021.
- [27] Teerapittayanon S, McDanel B, Kung HT. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. Proceedings of the IEEE International Conference on Pattern Recognition (ICPR) 2016: 2464–2469.
- [28] Rassadin AG, Savchenko AV. Deep Neural Networks Performance Optimization in Image Recognition. Proceedings of the III International Conference on Information Technologies and Nanotechnologies (ITNT) 2017: 649–654.
- [29] Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A. Presentation and Validation of the Radboud Faces Database. Cognition & Emotion 2010; 24(8): 1377–1388.
- [30] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing Machine Learning Models via Prediction APIs. 25th USENIX Security Symposium (USENIX Security 16) 2016: 601–618.
- [31] Pinto N, Stone Z, Zickler T, Cox D. Scaling Up Biologically-inspired Computer Vision: A Case Study In Unconstrained Face Recognition On Facebook. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)2011: 35–42.

# Real-time tracking of multiple objects with locally adaptive correlation filters

A.N. Ruchay<sup>1</sup>, V.I. Kober<sup>1</sup>, I.E. Chernoskulov<sup>1</sup>

<sup>1</sup>Chelyabinsk State University, 129 Bratiev Kashirinykh st., Chelyabinsk 454001, Russia

---

## Abstract

A tracking algorithm using locally adaptive correlation filtering is proposed. The algorithm is designed to track multiple objects with invariance to pose, occlusion, clutter, and illumination variations. The algorithm employs a prediction scheme and composite correlation filters. The filters are synthesized with the help of an iterative algorithm, which optimizes discrimination capability for each target. The filters are adapted online to targets changes using information of current and past scene frames. Results obtained with the proposed algorithm using real-life scenes, are presented and compared with those obtained with state-of-the-art tracking methods in terms of detection efficiency, tracking accuracy, and speed of processing.

*Keywords:* tracking; locally adaptive filters; correlation filters; matching.

---

## 1. Introduction

Nowadays, object tracking is a widely investigated topic in engineering and computer vision [1, 2]. Video surveillance, vehicle navigation, human-computer interaction, and robotics are examples of tracking applications [3, 4, 5, 6, 7, 8, 9, 10, 11]. In tracking, objects are localized in a current frame automatically by applying a detection engine [12, 13, 14, 15]. A main difficulty in object tracking is that the observed scene is commonly degraded by additive noise, the presence of a cluttered background, geometric modifications such as pose changing and scaling, gesticulations, and nonuniform illumination. Additionally, eventual occlusions and real-time requirements are challenges that a modern tracking algorithm must solve.

Object tracking based on correlation-based methods are widely utilized as an attractive alternative to existing tracking algorithms [16, 17, 18]. Correlation filters have a good formal basis, and they can be easily implemented for real-time applications [19, 20]. Recognition methods involving template matching are not useful in some cases, for instance, when articulation changes global features like the object outline. So, conventional correlation filters without training may yield a poor performance to recognize objects possessing incomplete information [21, 22, 23]. Adaptive approach to the filter design helps us to synthesize adaptive filters for object tracking [24, 25].

In this work, we propose an algorithm for object tracking based on locally adaptive correlation filtering. The algorithm is able to carry out object tracking with a high accuracy in an video without offline training. The objects are selected at the beginning of the algorithm. Afterwards, a composite correlation filter optimized for distortion tolerant pattern recognition is designed to recognize the target in the next frame. The impulse responses of optimum correlation filters are used to synthesize composite filters for distortion invariant object tracking. Two techniques are used to improve the detection performance: adaptive procedure that achieves a prespecified performance for a typical scene background, and multiple composite filters (bank of composite filters) when numerous views are available for training. The filter is dynamically adapted to each frame using information of current and past scene observations.

The paper is organized as follows. Section 2 recalls the optimum composite filter design. Section 3 describes the suggested algorithm for object tracking by locally adaptive correlation filtering. Computer simulation results obtained with the proposed algorithm are presented and compared with common algorithms in terms of detection efficiency and location accuracy in section 4. Finally, section 5 presents our conclusions.

## 2. Composite filter design using optimum correlation filters

We are interested in the design of a correlation filter which is able to recognize an object embedded into a disjoint background in the scene corrupted with additive noise. The designed filter should be also able to recognize geometrically distorted versions of the target. Let  $T = \{t_i(x, y); i = 1, \dots, N\}$  be an image set containing geometrically distorted versions of the target to be recognized. The input scene is assumed to be composed by the target  $t(x, y)$  embedded into a disjoint background  $b(x, y)$  at unknown coordinates  $(\tau_x, \tau_y)$ , and the whole scene is corrupted with additive noise  $n(x, y)$ , as follows:

$$f(x, y) = t(x - \tau_x, y - \tau_y) + b(x, y)\bar{w}(x - \tau_x, y - \tau_y) + n(x, y), \quad (1)$$

where  $\bar{w}(x, y)$  is a binary function defined as zero inside the target area, and unity elsewhere. The optimum filter for detecting the target, in terms of the signal to noise ratio (SNR) and the minimum variance of measurements of location errors (LE), is the generalized matched filter (GMF) [26], whose frequency response is given by

$$H^*(u, v) = \frac{T(u, v) + \mu_b \bar{W}(u, v)}{P_b(u, v) \otimes |\bar{W}(u, v)|^2 + P_n(u, v)}. \quad (2)$$



In (2),  $T(u, v)$  and  $\overline{W}(u, v)$  are the Fourier transforms of  $t(x, y)$  and  $\overline{w}(x, y)$ , respectively;  $\mu_b$  is the mean value of the background  $b(x, y)$ ;  $P_b(u, v)$  and  $P_n(u, v)$  denote power spectral densities of  $b_0(x, y) = b(x, y) - \mu_b$  and  $n(x, y)$ , respectively. The symbol  $\otimes$  denotes convolution.

Let  $h_i(x, y)$  be the impulse response of a GMF constructed for the  $i$ th available view of the target  $t_i(x, y)$  in  $T$ . Let  $H = \{h_i(x, y); i = 1, \dots, N\}$  be the set of all GMF impulse responses constructed for all training images  $t_i(x, y)$ . Additionally, let  $S = \{s_i(x, y); i = 1, \dots, M\}$  be an image set containing  $M$  unwanted patterns to be rejected. We want to synthesize a filter capable to recognize all target views in  $T$  and to reject the false patterns in  $S$ , by combining the optimum filter templates contained in  $H$ , and by using only a single correlation operation. The required filter  $p(x, y)$ , can be constructed as follows [26]:

$$p(x, y) = \sum_{i=1}^N \alpha_i h_i(x, y) + \sum_{i=N+1}^{N+M} \alpha_i s_i(x, y), \quad (3)$$

where the coefficients  $\{\alpha_i; i = 1, \dots, N + M\}$  are chosen to satisfy prespecified output values for each pattern in  $U = T \cup S$ . Using vectormatrix notation, we denote by  $\mathbf{R}$  a matrix with  $N + M$  columns, where each column is the vector version of each element of  $U$ . Let  $\mathbf{a} = [\alpha_i; i = 1, \dots, N + M]^T$  be a vector of coefficients. Thus, (3) can be rewritten as

$$\mathbf{p} = \mathbf{R}\mathbf{a}. \quad (4)$$

Let us denote by

$$\mathbf{u} = \left[ \underbrace{1, \dots, 1}_{N \text{ones}}, \underbrace{0, \dots, 0}_{M \text{zeros}} \right]^T,$$

the desired responses to the training patterns, and denote by  $\mathbf{Q}$  the matrix whose columns are the elements of  $U$ . The response constraints can be expressed as

$$\mathbf{u} = \mathbf{Q}^+ \mathbf{p}, \quad (5)$$

where superscript  $+$  denotes complex conjugate. Substituting (4) into (5), we obtain

$$\mathbf{u} = \mathbf{Q}^+ \mathbf{R}\mathbf{a}.$$

Thus, the solution for  $\mathbf{a}$ , is

$$\mathbf{a} = [\mathbf{Q}^+ \mathbf{R}]^{-1} \mathbf{u}. \quad (6)$$

Finally, substituting (8) into (4), the solution for the composite filter is given by

$$\mathbf{p} = \mathbf{R}[\mathbf{Q}^+ \mathbf{R}]^{-1} \mathbf{u}. \quad (7)$$

Note that the value of the correlation peak when using the filter given in Eq. 7, is expected to be close to unity for true-class objects, and close to zero for false-class objects.

The discrimination capability (DC) is a measure of the ability of the filter to distinguish a target from unwanted objects; it is defined by [26]

$$DC = 1 - \frac{|c^b|^2}{|c^t|^2},$$

where  $c^b$  is the value of the maximum correlation sidelobe in background area and  $c^t$  is the value of the correlation peak generated by the target. A  $DC$  value close to unity indicates that the filter has a good capability to distinguish between the target and any false object. Negative values of the  $DC$  indicate that the filter is unable to detect the target. Also, if the obtained  $DC$  is greater than a prespecified threshold ( $DC > DC_{th}$ ), then the target is considered as detected and, otherwise, the target is rejected.

### 3. Object tracking with locally adaptive correlation filtering

In this section we describe the proposed algorithm for object tracking based on composite correlation filtering. The proposed algorithm is robust to pose changes and appearance modifications of objects, as well as to the presence of scene noise, illumination changes, and target occlusions.

The algorithm starts with an initialization step where the objects are selected. Next, an optimum correlation filter for reliable detection and location estimation of the target is designed. Afterwards, a composite locally adaptive correlation filter is synthesized. The proposed algorithm incorporates an automatic re-initialization mechanism that reestablishes the tracking if it fails. The block diagram of the proposed algorithm is depicted in Fig. 1. The detailed operation steps are explained below.

Step 1: For each object select a small target  $t_i(x, y)$  from a captured scene frame  $f_i(x, y)$  containing the object to be tracked.

Step 2: Synthesize an optimum correlation filter  $h_i(x, y)$  with (2) for reliable detection and location estimation of the target  $t_i(x, y)$  in the observed local frame  $l_i(x, y)$ .

Step 3: Synthesize a composite locally adaptive correlation filter  $p_i(x, y)$  as follows. First, detect and locate the target by  $h_i(x, y)$  filter from the observed local frame  $l_i(x, y)$ . If the obtained  $DC$  is greater than a prespecified threshold ( $DC > DC_{rec}$ ), then the target is considered as successfully detected,  $t_i(x, y)$  added into the set  $T$  and recursion should be stopped. Otherwise, the target  $s_i(x, y)$  corresponding to a false peak added into the set  $S$ . Second, synthesize a composite filter  $p_i(x, y)$  with the help of (7). Third, detect and locate the target by  $p_i(x, y)$  filter from the observed local frame  $l_i(x, y)$  recursively until the condition  $DC > DC_{rec}$  is satisfied.

Step 4: Detect and locate the target in the observed local frame  $l_{i+1}(x, y)$  from a new scene frame  $f_{i+1}(x, y)$  by  $p_i(x, y)$  filter. The coordinates of the observed local frame  $l_{i+1}(x, y)$  are provided by a prediction process that analyzes the motion kinematics of the target. If the obtained  $DC$  is greater than a prespecified threshold ( $DC > DC_{th}$ ), then the target is considered as successfully detected and  $p_i(x, y)$  filter added to the bank  $B$  of composite correlation filters. Otherwise, the target is lost in the observed local frame  $l_{i+1}(x, y)$  and we recursively used the filters from bank  $B$  until condition  $DC > DC_{con}$  is satisfied. The filter from bank  $B$  with condition  $DC > DC_{con}$  is used to a new scene frame. If the target is lost in the observed local frame  $l_{i+1}(x, y)$  with help the filters from bank  $B$ , then the coordinates of the target is set coordinates of the past scene frame  $f_i(x, y)$  and we proceed to a new scene frame  $f_{i+2}(x, y)$ .

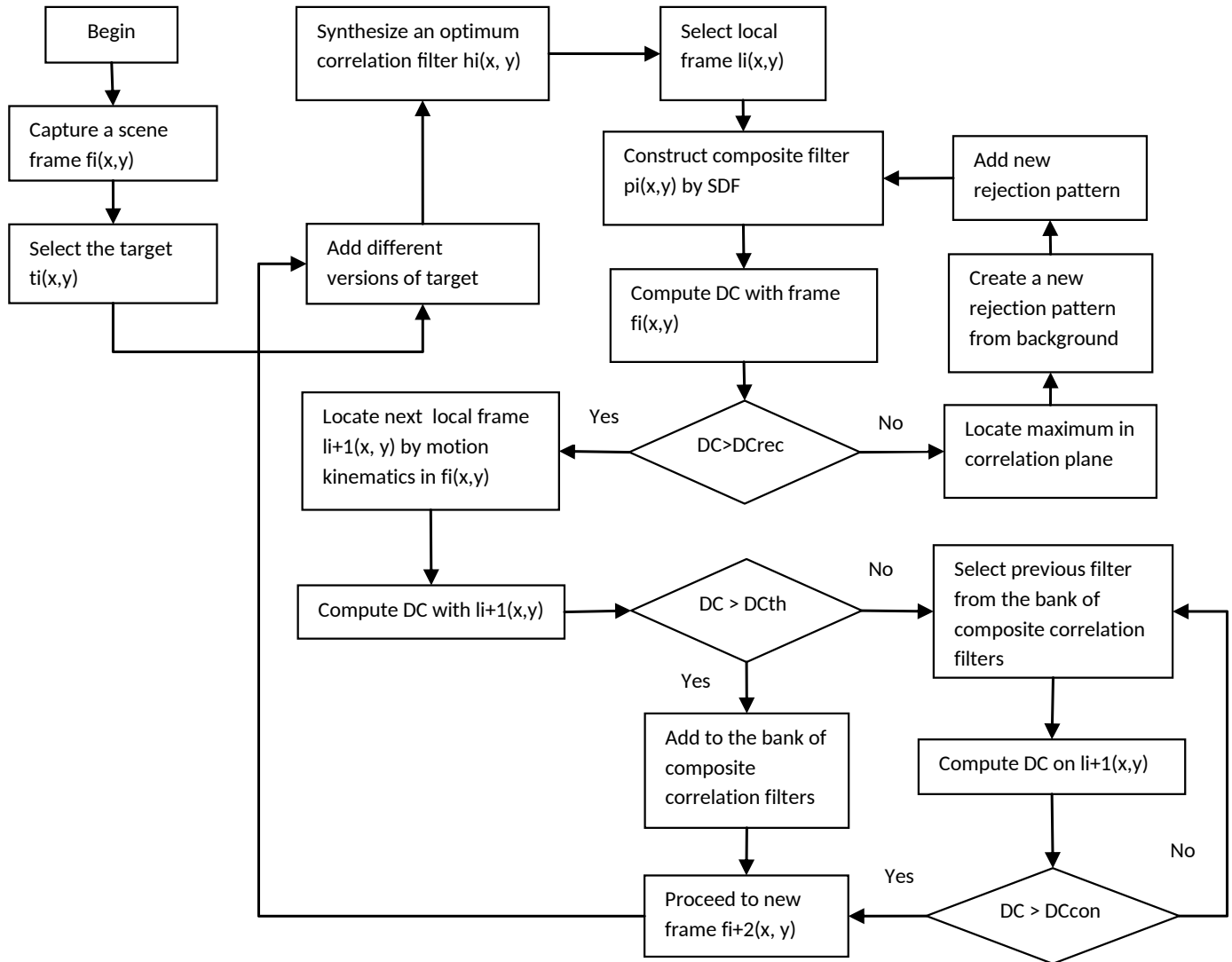


Fig. 1. Block diagram of the proposed tracking algorithm based on locally adaptive correlation filtering.

#### 4. Computer simulation

In this section, computer simulation results obtained with the proposed algorithm for object tracking are presented and compared with common algorithms in terms of detection efficiency, tracking accuracy, and speed of processing.

In order to evaluate the performance of our tracker, we conduct experiments on 100 challenging image sequences from Object Tracking Benchmark (TB-100 database) [27]. These sequences cover most challenging situations in object tracking: Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC), Low Resolution (LR).



For comparison, we run 3 state-of-the-art algorithms with the same initial position of the target. The first tracking algorithm (SURF) [28] is based on matching of local features and descriptors. The second tracking algorithm (STRUCK) predicts the target location change between frames on the basis of structured learning [29]. The third collaborative tracking algorithm (SCM) is combined a sparsity-based discriminative classifier and a sparsity-based generative model [30]. The work [27] performed large-scale experiments to evaluate the performance of recent 33 object-tracking algorithms. Tracking algorithms STRUCK and SCM perform much better than the others.

For evaluating of detection efficiency we use an evaluation metric of the overlap score. Given a tracked bounding box  $r_t$  and the ground-truth bounding extent  $r_0$  of a target object, the overlap score is defined as

$$S = \frac{\|r_t \cap r_0\|}{\|r_t \cup r_0\|}, \quad (8)$$

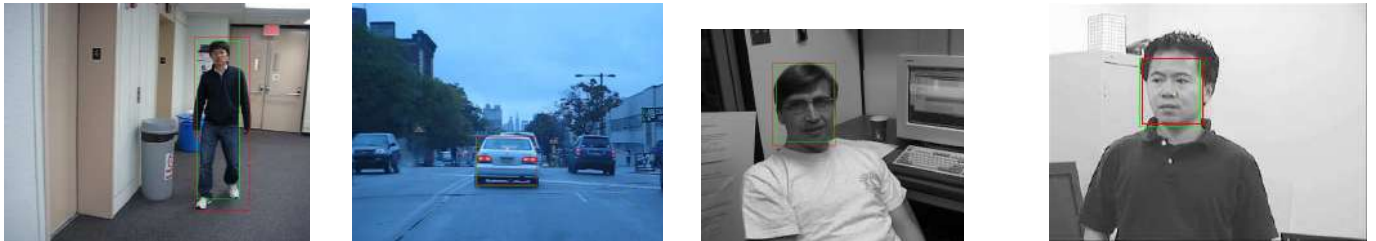
where  $\cap$  and  $\cup$  represent the intersection and union operators, respectively, and  $\|\cdot\|$  denotes the number of pixels in a region. This average overlap score (AOS) can be used as the performance measure. In addition, the overlap scores can be used for determining whether an algorithm successfully tracks a target in a frame, by testing whether  $S$  is larger than a threshold of 0.5. Also we evaluate the tracking algorithms using the average center location error (ACLE) for all image sequences from database.

Table 1 shows the average overlap score (AOS), the average center location errors (ACLE) and the Average Processing Time (APT) on a scena for all the tracking algorithms with the overlap threshold of 0.5. The evaluation results show that our proposed algorithm is faster than the others and more accurate in terms of the average center location errors.

**Table 1.** Evaluation results of the state-of-the-art STRUCK, SCM, SURF and proposed algorithms by the average overlap score (AOS), the average center location errors (ACLE), and the Average Processing Time (APT)

Tracker	All	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	APT	ACLE
Proposed	53.3	50.7	51.1	60.0	56.4	43.5	56.7	55.7	44.6	50.5	41.7	51.4	0.2005	68.8
STRUCK	57.5	59.3	52.4	55.6	57.0	59.0	59.1	59.9	55.9	57.3	58.9	57.8	0.2894	61.5
SCM	54.4	61.3	51.5	42.8	51.8	61.1	61.7	45.2	56.8	57.0	56.4	55.8	0.3122	64.8
SURF	35.2	37.4	25.8	41.6	39.7	37.3	23.0	45.4	36.0	34.8	46.7	33.0	0.1668	276.6

When an object moves fastly on the FM subset, the proposed algorithm performs much better than the others. However, the proposed algorithm does not perform well in the subset (IV, OCC, OV) due to illumination variation, and partial occlusion of the target. On the other subsets, the Struck, SCM, and the proposed algorithms outperform other the state-of-the-art algorithms. Fig. 2 shows sample tracking results of the proposed algorithms where the target objects are marked with red rectangles and the actually tracked objects by the proposed algorithm are marked with green rectangles.



**Fig. 2.** Results of tracking by proposed algorithm.

## 5. Conclusion

A tracking algorithm using locally adaptive correlation filtering is proposed. The algorithm is designed to track multiple objects with invariance to pose, partial occlusion, clutter, and illumination variations. The algorithm employs a prediction scheme and composite correlation filters. The filters are synthesized with the help of an iterative algorithm, which optimizes discrimination capability for each target. The filters are adapted online to targets changes using information of current and past scene frames. The evaluation results show that our proposed algorithm is faster than the others and more accurate in terms of the average center location errors. On the majority test sets the proposed algorithm performs much better than the state-of-the-art algorithms.

## Acknowledgments

This work was supported by the Russian Science Foundation, grant no. 15-19-10010.

## References

- [1] Karasulu, B. Performance Evaluation Software: Moving Object Detection and Tracking in Videos [Text] / B. Karasulu, S. Korukoglu. — New York : Springer, 2013.
- [2] Talmale, S. Object tracking in images and videos [Text] / S.K. Talmale, N.J. Janwe // International Journal Of Engineering And Computer Science. —2016. —Vol. 5(1). —P. 15482–15486.
- [3] Accurate three-dimensional pose recognition from monocular images using template matched filtering [Text] / Kenia Picos, Victor H. Diaz-Ramirez, Vitaly Kober [et al.] // Optical Engineering. —2016. —Vol. 55, no. 6. —P. 063102.
- [4] Echeagaray-Patron, B. A. Conformal parameterization and curvature analysis for 3d facial recognition [Text] / B. A. Echeagaray-Patron, D. Miramontes-Jaramillo, V. Kober // 2015 International Conference on Computational Science and Computational Intelligence (CSCI). — [S. l. : s. n.], 2015. —P. 843–844.
- [5] Echeagaray-Patron, B. A. 3d face recognition based on matching of facial surfaces [Text] / Beatriz A. Echeagaray-Patron, Vitaly Kober. — Vol. 9598. — [S. l. : s. n.], 2015. —P. 95980V–95980V–8.
- [6] Diaz-Escobar, J. A robust hog-based descriptor for pattern recognition [Text] / Julia Diaz-Escobar, Vitaly Kober. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 99712A–99712A–7.
- [7] Diaz-Escobar, J. Text Detection in Digital Images Captured with Low Resolution Under Nonuniform Illumination Conditions [Text] / Julia Diaz-Escobar, Vitaly Kober // Pattern Recognition: 8th Mexican Conference, MCPR 2016, Guanajuato, Mexico, June 22-25, 2016. Proceedings / Ed. by José Francisco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, Victor Ayala Ramirez [et al.]. — Cham : Springer International Publishing, 2016. —P. 3–12.
- [8] An efficient algorithm for matching of slam video sequences [Text] / Jose A. Gonzalez-Fraga, Victor H. Diaz-Ramirez, Vitaly Kober [et al.]. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 99712Z–99712Z–10.
- [9] Effective indexing for face recognition [Text] / I. Sochenkov, A. Sochenkova, A. Vokhmintsev [et al.]. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 997124–997124–9.
- [10] Face recognition based on a matching algorithm with recursive calculation of oriented gradient histograms [Text] / A. V. Vokhmintsev, I. V. Sochenkov, V. V. Kuznetsov, D. V. Tikhonkikh // Doklady Mathematics. —2016. —Vol. 93, no. 1. —P. 37–41.
- [11] Tihonkih, D. A modified iterative closest point algorithm for shape registration [Text] / Dmitrii Tihonkih, Artyom Makovetskii, Vladislav Kuznetsov. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 99712D–99712D–8.
- [12] Miramontes-Jaramillo, D. A Robust Tracking Algorithm Based on HOGs Descriptor [Text] / Daniel Miramontes-Jaramillo, Vitaly Kober, Víctor Hugo Díaz-Ram // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 19th Iberoamerican Congress, CIARP 2014, Puerto Vallarta, Mexico, November 2-5, 2014. Proceedings / Ed. by Eduardo Bayro-Corrochano, Edwin Hancock. — Cham : Springer International Publishing, 2014. —P. 54–61.
- [13] Miramontes-Jaramillo, D. Multiple objects tracking with hogs matching in circular windows [Text] / Daniel Miramontes-Jaramillo, Vitaly Kober, Victor H. Diaz-Ramirez. — Vol. 9217. — [S. l. : s. n.], 2014. —P. 92171N–92171N–8.
- [14] Miramontes-Jaramillo, D. Robust illumination-invariant tracking algorithm based on hogs [Text] / Daniel Miramontes-Jaramillo, Vitaly Kober, Víctor Hugo Díaz-Ramírez. — Vol. 9599. — [S. l. : s. n.], 2015. —P. 95991Q–95991Q–8.
- [15] Miramontes-Jaramillo, D. Real-time tracking based on rotation-invariant descriptors [Text] / Daniel Miramontes-Jaramillo, Vitaly Kober // 2015 International Conference on Computational Science and Computational Intelligence (CSCI). —2015. —Vol. 00. —P. 543–546.
- [16] Ontiveros-Gallardo, S. E. Objects tracking with adaptive correlation filters and kalman filtering [Text] / Sergio E. Ontiveros-Gallardo, Vitaly Kober. — Vol. 9598. — [S. l. : s. n.], 2015. —P. 95980X–95980X–8.
- [17] Ontiveros-Gallardo, S. E. Correlation-based tracking using tunable training and kalman prediction [Text] / Sergio E. Ontiveros-Gallardo, Vitaly Kober. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 997129–997129–9.
- [18] Ruchay, A. A correlation-based algorithm for recognition and tracking of partially occluded objects [Text] / Alexey Ruchay, Vitaly Kober. — Vol. 9971. — [S. l. : s. n.], 2016. —P. 99712R–99712R–9.
- [19] Facial recognition using composite correlation filters designed with multiobjective combinatorial optimization [Text] / Andres Cuevas, Victor H. Diaz-Ramirez, Vitaly Kober, Leonardo Trujillo. — Vol. 9217. — [S. l. : s. n.], 2014. —P. 921710–921710–8.
- [20] Aguilar-González, P. M. Adaptive composite filters for pattern recognition in nonoverlapping scenes using noisy training images [Text] / Pablo Mario Aguilar-González, Vitaly Kober, Víctor Hugo Díaz-Ramírez // Pattern Recogn. Lett. —2014. —Vol. 41. —P. 83–92.
- [21] Díaz-Ramírez, V. H. Object Tracking in Nonuniform Illumination Using Space-Variant Correlation Filters [Text] / Víctor Hugo Díaz-Ramírez, Kenia Picos, Vitaly Kober // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part II / Ed. by José Ruiz-Shulcloper, Gabriella Sanniti di Baja. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2013. —P. 455–462.
- [22] Real-time tracking of multiple objects using adaptive correlation filters with complex constraints [Text] / Victor H. Diaz-Ramirez, Viridiana Contreras, Vitaly Kober, Kenia Picos // Optics Communications. —2013. —Vol. 309. —P. 265–278.
- [23] Diaz-Ramirez, V. H. Target tracking in nonuniform illumination conditions using locally adaptive correlation filters [Text] / Victor H. Diaz-Ramirez, Kenia Picos, Vitaly Kober // Optics Communications. —2014. —Vol. 323. —P. 32–43.
- [24] Robust Face Tracking with Locally-Adaptive Correlation Filtering [Text] / Leopoldo N. Gaxiola, Víctor Hugo Díaz-Ramírez, Juan J. Tapia [et al.] // Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 19th Iberoamerican Congress, CIARP 2014, Puerto Vallarta, Mexico, November 2-5, 2014. Proceedings / Ed. by Eduardo Bayro-Corrochano, Edwin Hancock. — Cham : Springer International Publishing, 2014. —P. 925–932.
- [25] Target tracking with dynamically adaptive correlation [Text] / Leopoldo N. Gaxiola, Victor H. Diaz-Ramirez, Juan J. Tapia, Pascuala Garcia-Martinez // Optics Communications. —2016. —Vol. 365. —P. 140–149.
- [26] Ramos-Michel, E. M. Adaptive composite filters for pattern recognition in linearly degraded and noisy scenes [Text] / Erika M. Ramos-Michel, Vitaly Kober // Optical Engineering. —2008. —Vol. 47, no. 4. —P. 047204–047204–7.
- [27] Wu, Y. Object tracking benchmark [Text] / Y. Wu, J. Lim, M. H. Yang // IEEE Transactions on Pattern Analysis and Machine Intelligence. —2015. —Vol. 37, no. 9. —P. 1834–1848.
- [28] Al-asadi, T. Object detection and recognition by using enhanced speeded up robust feature [Text] / T.A. Al-asadi, A.J. Obaid // International Journal of Computer Science and Network Security. —2016. —Vol. 16(4). —P. 66–71.
- [29] Torr, P. H. S. Structured output tracking with kernels [Text] / Philip H. S. Torr, Sam Hare, Amir Saffari // 2011 IEEE International Conference on Computer Vision (ICCV 2011). —2011. —Vol. 00. —P. 263–270.
- [30] Zhong, W. Robust object tracking via sparsity-based collaborative model [Text] / Wei Zhong // Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). —CVPR '12. —Washington, DC, USA : IEEE Computer Society, 2012. —P. 1838–1845.

# Methods for automated vectorization of point objects on cartographic images

S. Rychazhkov<sup>1</sup>, V. Fedoseev<sup>1,2</sup>, R. Yuzkiv<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

In this paper, we propose methods designed to automate vectorization of point objects at cartographic image digitizing. These methods are based on the analysis of the skeleton-contour image representation and allow to detect point objects, estimate their parameters, and restore a regular grid of points. The experiments show high accuracy of the developed methods, which results in the possibility of using them in order to accelerate the process of digitizing cartographic images.

*Keywords:* vectorization; digitizing; cartographic image; point object; skeletonization; geoinformation system; GIS; skeleton-contour representation

---

## 1. Introduction

At present, the problem of digitizing important information from paper carriers is of great urgency [1]. These include old books, drawings, engravings important for the preservation of cultural heritage. Also, we can note plans, charts, and maps printed 20 or more years ago, which digital sources did not exist or were not preserved. Such documents in addition to historical value can have great practical importance. Thus, topographic maps and plans are an important data source for geoinformation systems (GIS) and can be used to solve a lot of practical problems of territory analysis and development: digital terrain model creation, maintenance of forest and water registers, identification of flooding areas and others [2]. So the subject of this paper is digitizing such cartographic images.

The simplest way to digitize paper maps is to scan them in bitmaps with subsequent coordinate binding in GIS. In such case, these data can be used as an underlying layer supplementing the information on a particular territory. However, they can not be used for solving any analytical tasks in GIS. For that, raster maps must be vectorized and become a part of the GIS database. In the past 20 years, a couple of methods and software tools were developed to simplify the process of digitizing bitmaps. For instance, we can mention such tools as ArcScan (in ArcGIS software), PowerTRACE (in CorelDRAW), EasyTrace, LineTracer, Spotlight, etc. However, cartographic images often contain various figures (topographic signs), and for their fast digitizing it is often necessary to develop new specialized methods and software tools.

The current paper is devoted to the development of methods for solving one of the particular problems in the described area: vectorization of point objects. In general, all the objects plotted on the maps can be divided into two main groups. The first one includes linear objects, defined by their axial line, drawing method, and thickness. At their digitizing, the most important task is to restore the axial line with high precision. For that, there are used methods based on either raster skeletonization [3, 4] or vector skeleton-contour representation [5]. For the first group, the contour line is less significant than the axial line. The second group consists of all other objects, which are topographic signs, letters, and objects of complex shape. For them, unlike the first group, the contour line is more important than the axial line. Point objects, formally related to the second group, are a special case, and for their exact digitizing, as will be shown later, both the contour and the axial line are of high importance.

Point objects are quite common on paper maps. They are used, for example, for designation of plantations, bushes, elevations, sand, orchards, etc. (see Figure 1) [6]. When vectorizing in a GIS, each such object is associated with a point object, defined by its coordinates  $(x, y)$  and radius  $R$ . In this case, depending on the type of data source, the radius can be either arbitrary (for example, if the point objects depict sands) or constant (most other objects). Often the exact point radius is unknown or is not of great significance. In this case, point objects of the same type should be justified in size after the vectorization. Another common case is the strictly ordered placement of point objects on maps. This is typical, for example, for the designation of orchards or forest plantations (see Figure 1). In this case, vector objects in GIS should be placed on the terrain according to the same rule. Thus, we see some special cases in the problem of digitizing point objects. This fact necessitates the development of specialized algorithms for automated vectorization of point objects that is the goal of this paper.

The paper is organized as follows. The second section describes the general technology of cartographic image digitizing that we use and also denotes its stages requiring the development of specialized methods for point objects. Section 3 contains a description of the methods developed to detect such objects, restore their radius, and localize them on a continuous or discrete grid. The results of experimental studies of the methods developed are presented in Section 4. The paper ends with the conclusion and acknowledgments.

## 2. General technology of cartographic image digitizing

In our work, we used a typical technology of cartographic image digitizing that includes the following stages:

1. Preprocessing a bitmap.
2. Transformation of the image into the skeleton-contour representation.
3. Classification of objects to be vectorized.
4. Type-adapted object vectorization.
5. Correction of vector results based on analysis of object groups.

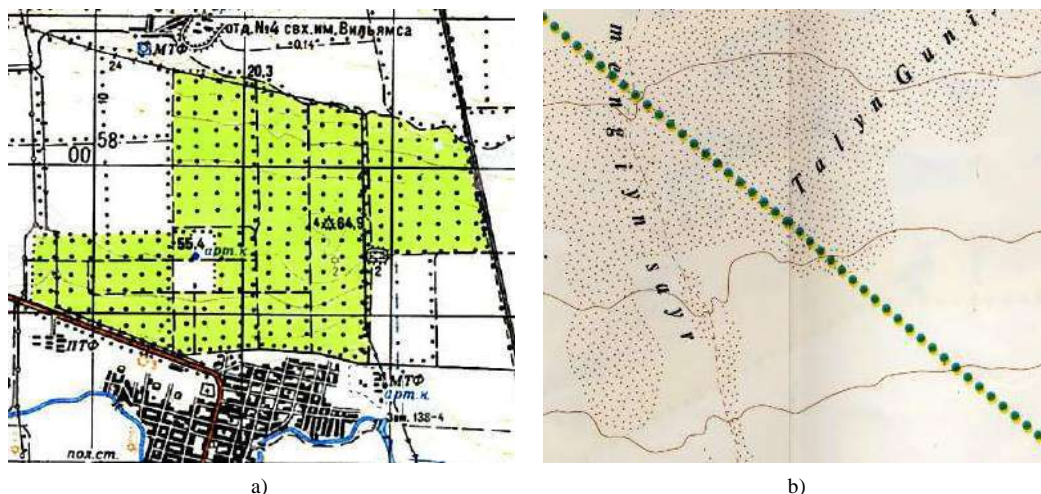


Fig. 1. Examples of cartographic images with point objects.

At the first stage, the original color scanned image is converted into one or more binary images prepared for subsequent vectorization. Thus, the mandatory procedures used at this stage are color segmentation and binarization. The need for color segmentation is due to the fact that most maps contain several thematic layers displayed by predetermined colors (for example, blue is used for hydrography, red or brown for relief, green for vegetation, etc.) [7, 8]. Thus, the separation of the original image into thematic layers by the color feature will avoid the overlay of dissimilar objects and improve the digitizing quality. Binarization is necessary because of the need for subsequent skeletonization, which usually requires binary input data [5]. In addition to the two mentioned procedures, at the first stage, additional processing operations can be carried out to improve image quality: linear or nonlinear filtering for noise reduction, morphological processing for gluing lines etc.

The binary image obtained at the first stage can be considered as a set of figures. At the second stage, we estimate contours and skeletons for each of these figures. According to [5], the contour of a figure is an ordered set of vertices  $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$  defining a polygon approximating the boundary representation of the figure. Such a polygon can not have intersections with other contours and self-intersections. For a formal definition of a figure skeleton, we use a definition based on the concept of a maximal empty circle [5]. An empty circle of a figure  $A$  is a closed set of points  $\mathcal{S}_r(p) = \{q: q \in R^2, d(p, q) \leq r\}$  such that  $\mathcal{S}_r(p) \cap A = \emptyset$ . This set is a circle of the radius  $r \geq 0$  with the center at a point  $p \in R^2$ . The maximum empty circle is an empty circle, which is not contained in any other empty circle. The skeleton of a figure is the set of centers of all its maximal empty circles.

The skeleton can be described by a flat graph [5], whose vertices are the centers of maximal empty circles having either one common point with the boundary of the figure or three and more points. In this case, the edges of a graph are lines that consist of the centers of those empty circles that touch the boundary exactly at two points. As shown in [5], the construction of a skeleton-contour representation for a raster image is associated with a number of difficulties. Therefore, in this paper, we used algorithms [9] developed by Leonid Mestetskiy and widely used in the papers of his school.

Since different methods of processing the skeleton-contour data are required for different object types (linear, point, area), at the third stage it is necessary to classify each object (figure) for correct processing performed at the fourth stage. The result of the fourth stage is the vector data, which can be further corrected in the fifth stage based on a priori known constraints on the geometry of similar objects. For example, many conventional topographic signs should have a constant size, and some groups of symbols must also be located at the same distance from each other [6]. The methods used at the stages 3-5 for vectorizing point objects are described in the next section.

### 3. Methods for point object vectorizing

#### 3.1. Estimation of a circle radius

As studies have shown, to classify point objects (i.e. to check whether an object is point-like), it is useful to estimate the object radius. The initial approximation of the circle radius  $R_0$  can be calculated based on the skeleton edge lengths  $l_j, j = 1..N$  and radiuses  $r_{j,1}$  and  $r_{j,2}$  of maximal empty inscribed circles for corresponding vertices. Let  $L$  be the length of the entire figure skeleton:

$$L = \sum_{j=1}^N l_j.$$

Then

$$R_0 = \sum_{i=1}^N \frac{r_{i,1} + r_{i,2}}{2} \frac{l_i}{L}.$$

Further, two different methods can be used to refine the radius of the circle: the method based on contour length, and the one based on figure area.

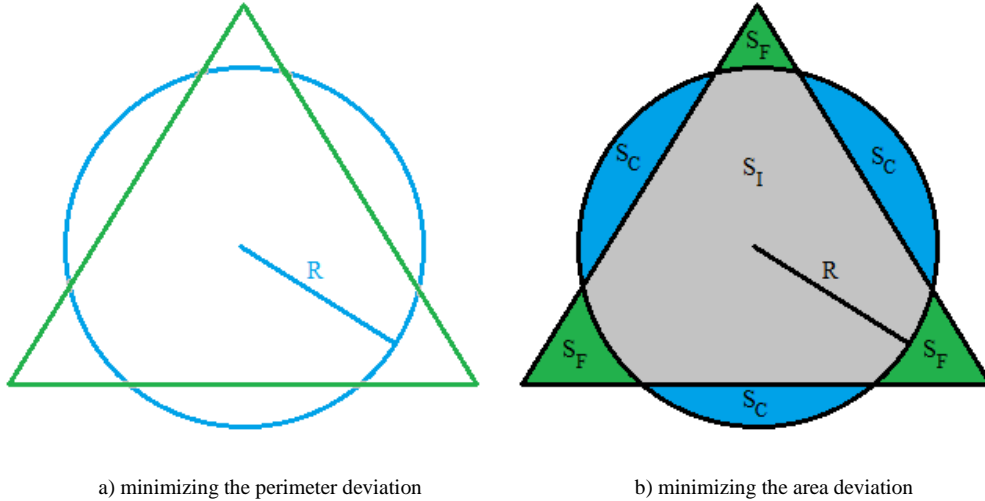


Fig. 2. An illustration of two methods for the radius refinement. The figure to be analyzed is a triangle.

Method 1 is based on minimizing the deviation of the figure perimeter from the theoretical circumference based on the value of the radius  $r$  :

$$R = \arg \min_{\rho} \left( \left| \frac{D}{2\pi\rho} - 1 \right| \right), \tag{1}$$

where  $D$  is the length of the figure contour.

The problem (1) can be solved numerically (for example, using the golden section search) with the initial approximation  $r = R_0$ . Method 1 is illustrated in Fig. 2a, where the figure contour (shown as a triangle) and the theoretical circumference are highlighted in color.

Method 2 is based on minimizing the deviation of the figure area from the theoretical value of circle area, determined by the radius  $r$ . Let  $S$  be the figure area,  $S_I$  – the area of intersection of a circle with a figure,  $S_F$  – the figure area that is outside of the circle,  $S_C$  – the circle area that is outside of the figure (see Fig. 2b). We used the following accuracy characteristics:

$$a = \frac{S_I}{S_I + S_F} = \frac{S_I}{S}, \quad b = \frac{S_I}{S_I + S_C} = \frac{S_I}{\pi r^2}.$$

Let  $g$  be the harmonic mean of  $a$  and  $b$  :

$$g = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2S_I}{S + \pi r^2}.$$

Then the radius  $R$  can be calculated by minimizing the deviation of  $g$  from unity:

$$R = \arg \min_{\rho} \left( \left| \frac{2S_I}{S + \pi\rho^2} - 1 \right| \right). \tag{2}$$

The problem (2) can be solved by the same method as the problem (1).

### 3.2. Detection of point objects

To detect point objects on a map, we need to calculate some features and classify whether the object is from the point class. Our studies have shown that a set of only two features is sufficient and the linear separating function can be applied for

$$f_1^1 = \left| \frac{D}{2pR} - 1 \right|,$$

$$f_1^2 = \left| \frac{2S_H}{S + pr^2} - 1 \right|.$$

To reduce the computational complexity, it is preferable to use the feature  $f_1^1$  for the first method of a circle radius refinement and the feature  $f_1^2$  for the second one.

The second feature is defined as the ratio of the skeleton length to the radius estimation:

$$f_2 = \frac{L}{R}.$$

For point objects, such ratio should be small, as shown in Fig. 3.

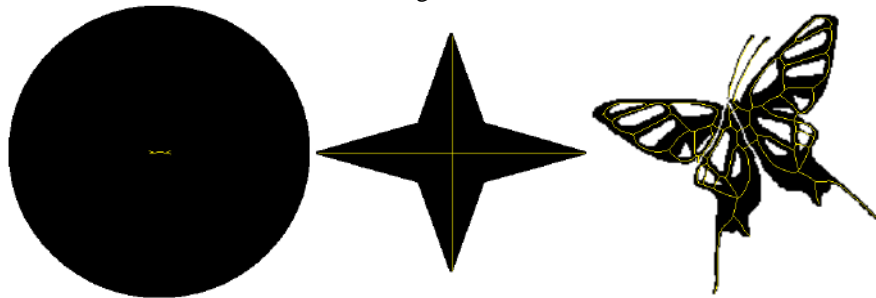


Fig. 3. Skeletons of figures with various shapes (figure shows that the skeleton length of a point object should be much less than the radius).

### 3.3. Reconstruction of a regular grid of objects

In Section 1 it was noted that some point symbols (for example, orchards, wooded areas) on maps follow the same distance from each other. Thus, the resulting vector objects must also keep a regular layout. However, this requirement is not guaranteed when objects are processed individually. To restore a regular object grid, the following method is proposed.

Let one of the grid nodes is located in the origin of coordinates. Then a two-dimensional grid can be defined by two vectors  $\vec{V}_1$  and  $\vec{V}_2$  (which are often orthogonal to each other) (see Fig. 4). Then the coordinates  $(x^{(i)}, y^{(i)})$  of a particular grid node can be found as follows:

$$(x^{(i)}, y^{(i)}) = k_1 \vec{V}_1 + k_2 \vec{V}_2, \tag{3}$$

where  $k_1, k_2 \in \mathbb{Z}$  are the indices of the grid node.

The initial approximation of the grid can be based on three adjacent support objects, one of which is central, and the other two specify the direction of the vectors  $\vec{V}_1$  and  $\vec{V}_2$  (see Fig. 4). In practice, these support objects can be selected by the user. When reconstructing a regular grid of points, it is necessary to solve next three problems:

- refine the vectors  $\vec{V}_1$  and  $\vec{V}_2$ ;
- detect all point objects lying on the grid

$$S = \left\{ (x^{(i)}, y^{(i)}) : (x^{(i)}, y^{(i)}) = k_1 \vec{V}_1 + k_2 \vec{V}_2 \quad \forall k_1, k_2 \in \mathbb{Z} \right\}_r.$$

- correct the center coordinates of the found objects taking into account the refined vectors  $\vec{V}_1$  and  $\vec{V}_2$
- Here is the proposed algorithm:

**Step 0.** Set the initial approximations of vectors  $\vec{V}_1$  and  $\vec{V}_2$ ; add 3 reference objects to the set  $S$ .

**Step 1.** Find the theoretical coordinates of all grid nodes not yet added to the set  $S$  and having at least one neighbor in the set  $S$  by the rule of four neighbors.

**Step 2.** Among all the point objects of the map, find objects that are located at a distance no more than the threshold value  $d$  from the theoretical coordinates found in Step 1. If at least one object is found, go to Step 3, otherwise – to Step 5.

*A loop for all found objects (grid nodes). After the loop go to Step 1.*

**Step 3.** Refine vectors  $\vec{V}_1$  and  $\vec{V}_2$ :

$$\vec{V}_1 = \frac{N_1 \vec{V}_1 + \vec{A}_1}{N_1 + 1}, \quad \vec{V}_2 = \frac{N_2 \vec{V}_2 + \vec{A}_2}{N_2 + 1},$$



where  $N_1, N_2 \in \mathbb{Z}$  are the counts of vectors used to estimate the vectors;  $\vec{A}_1, \vec{A}_2$  are the estimations of vectors  $\vec{V}_1$  and  $\vec{V}_2$  for current grid node by Eq. (3) and known indices  $k_1, k_2$ . If one of these indices is zero, then it is obvious that only one of the vectors  $\vec{V}_1$  and  $\vec{V}_2$  is recalculated.

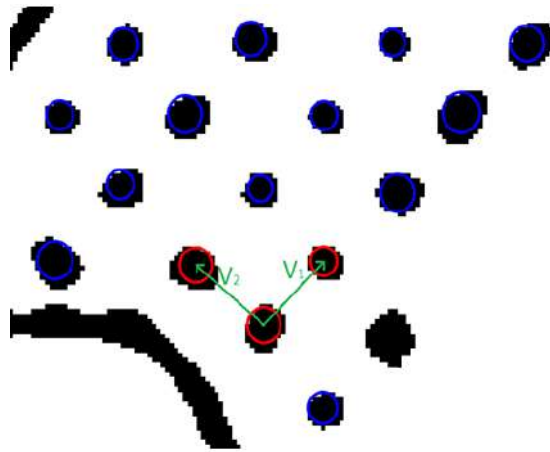


Fig. 4. An initial approximation of a grid of point objects.

**Step 4.** Add the current object to the set  $S$ .

**Step 5.** Recalculate the coordinates of the initial reference object (which corresponds to the indices  $k_1 = k_2 = 0$ ). Go to Step 1. Finish the algorithm if no objects have been added at step 4.

#### 4. Experimental research

For the experiments, we primarily used a scanned image containing many regularly located point objects with some linear ones (see Fig. 5a). For this image, a manual vectorization was performed, as well as automatic vectorization using the proposed methods. The results of manual vectorization were considered as reference ones. Also, during the experiments, a synthesized image was used (see Fig. 5b), for which the exact number of point objects and their locations are known.

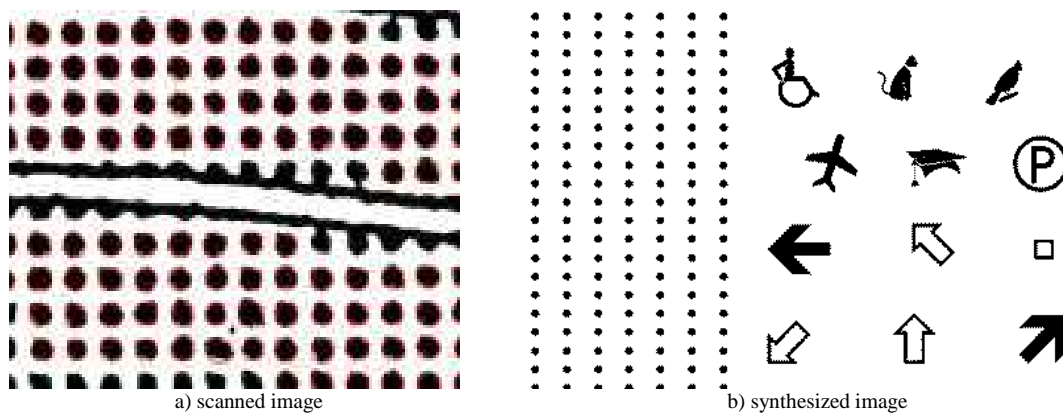


Fig. 5. Fragments of the test images.

##### 4.1. Detection and localization of point objects

Since the average point size influence on the parameters of a point classifier and we had only two test images of different origin, we decided not to train a particular classifier (which would have to be trained on a fragment of the same image). Instead, several fixed threshold values  $\Delta_1^1$  (or  $\Delta_1^2$ ) and  $\Delta_2$  for features  $f_1^1$  (or  $f_1^2$ ) and  $f_2$  were considered, and the classification was performed according to the rule

$$f_1^1 < \Delta_1^1 \left( f_1^2 < \Delta_1^2 \right)$$

The value of F-measure was used as a quality measure.

During the experiment on the real image, for the refinement method 1, the largest value of F-measure was achieved at  $\Delta_1^1 = 0.35$ ,  $\Delta_2 = 3.5$  and was equal to 0.9943. For the refinement method 2, the largest value was achieved at  $\Delta_2^1 = 0.2$ ,  $\Delta_2 = 4.5$  and reached 0.9957. Fig. 6 shows how  $\Delta_1^1$  (for the fixed)  $\Delta_2 = 3.5$  and  $\Delta_1^2$  (for fixed)  $\Delta_2 = 4.5$  influence on F-measure. As can be seen from the graphs, for a wide range of values, the F-measure has very high values, which confirms the possibility of separation of point objects in this image.

In addition, Table 1 specifies some other statistics. As can be seen, the second method showed a somewhat higher accuracy both in detecting the points and in estimating their location.

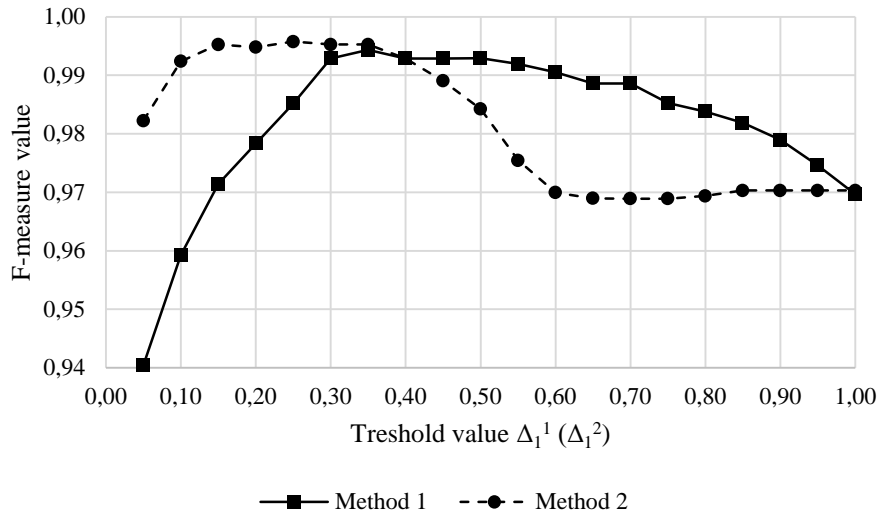


Fig. 6. Dependencies of F-measure on features  $f_1^1$  ( $f_1^2$ ).

Table 1. Comparison of the best results for two methods of radius refinement.

Characteristic	Method 1	Method 2
Number of detected objects	1056	1059
Number of false positives	6	6
Number of missed objects	6	3
F-measure for hits of the center of a point inside the true contour of a point	0.9953	0.9957
Median deviation of the centers of circles, pixels	1.065	0.7908
Median radius deviation, pixels	2.3146	1.1011

For the synthesized image (Fig. 5b), in a sufficiently large range of threshold values, it is possible to separate all 2243 point objects from all 326 figures having other shapes by both methods.

#### 4.2. Reconstruction of a regular grid

Testing of the grid reconstruction method was carried out using a fragment of the image from Fig. 5a, containing 147 circles, with  $d = 10$  pixels. As the measure of the method effectiveness, the number of objects matched with the grid nodes and the total deviation of their centers from the grid nodes were used.

During the experiment, at the first iteration (steps 0-4, until the coordinates of the support objects were refined), all circles were found which should lie on the grid, and the total deviation was 11.02 pixels (for 147 objects). At the second iteration (after changing the coordinates of the reference object), the deviation was 6.47 pixels, and the total deviation was reduced to 6.41 pixels.

Figure 7 shows the results of processing after the first iteration (Fig. 7a) and the final result (Fig. 7b). For comparison, both figures show the initial contours of points before the procedure starts. The result in Fig. 7b has a smaller deviation than initial approximation on Fig. 7a.

### 5. Conclusion

In this paper we have proposed a number of algorithms which use the skeleton-contour image representation:

1. the algorithm for estimating the radius of a point object;
2. the algorithm for point objects classification;
3. The algorithm for reconstructing a regular grid of objects.

The experimental study has shown, that the proposed methods have high accuracy and allow significantly accelerate the process of raster map digitizing.

### Acknowledgements

Research has been supported by the RFBR grants (projects No. 15-07-05576 and 16-41-630676). During the research, BSTransLib library by Leonid Mestetskiy was used [9]. The authors express their gratitude to the colleague Daria Terentyeva, who performed manual vectorization of the test images, which allowed to estimate the accuracy of the developed methods.



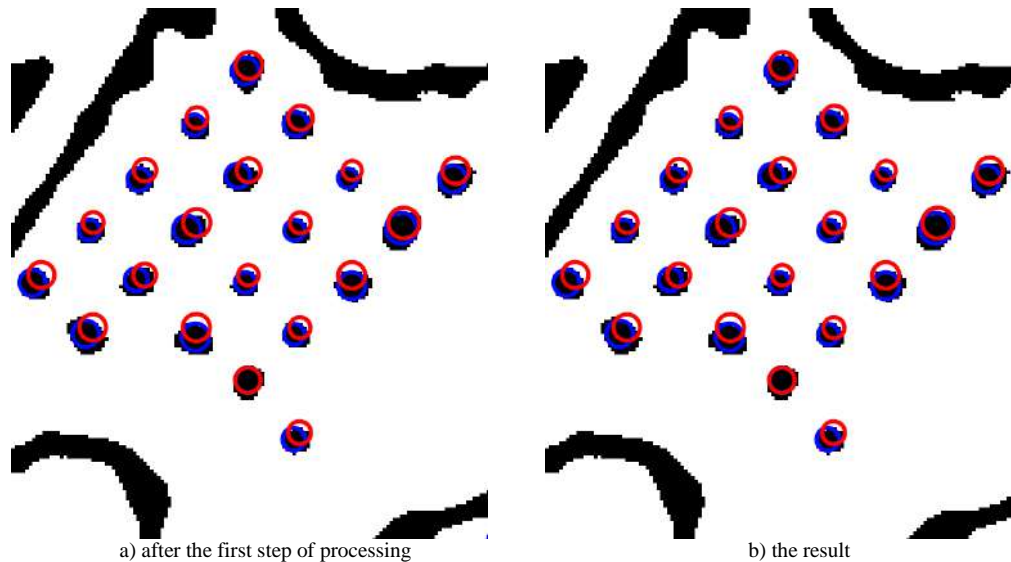


Fig. 7. Reconstruction of a regular grid.

## References

- [1] Conway P. Overview: Rationale for digitization and preservation. Handbook for digital projects. Andover, Massachusetts: NEDCC, 2000: 5–20.
- [2] Awange JL, Kyalo Kieam JB. Environmental geoinformatics. Springer-Verlag Berlin Heidelberg, 2013; 541 p. DOI: 10.1007/978-3-642-34085-7.
- [3] Zhang TY, Suen CY. A fast parallel algorithm for thinning digital patterns. Communications of ACM 1984; 27(3): 236–239. DOI: 10.1145/357994. 358023.
- [4] Oka S, Garg A, Varghese K. Vectorization of contour lines from scanned topographic maps. Automation in Construction 2012; 22: 192–202. DOI: 10.1016/j.autcon.2011.06.017.
- [5] Mestetskiy LM. Continuous morphology of the binary images: the figures, skeletons, circulars. Moscow: Fizmatlit, 2009; 288 p.
- [6] Topographic map symbols for the scales 1:5000, 1:2000, 1:1000, 1:500. Moscow: KartGeoCenter, 2005; 287 p. (in Russian)
- [7] Chiang Y-Y, Leyk S, Knoblock CA. A survey of digital map processing techniques. ACM Computing Surveys 2014; 47(1): 1–44. DOI: 10.1145/2557423.
- [8] Liu T, Miao Q, Tian K, Song J, Yang Y, Qi Y. SCTMS: Superpixel based color topographic map segmentation method. Journal of Visual Communication and Image Representation 2016; 35: 78–90. DOI: 10.1016/j.jvcir.2015.12.004.
- [9] Continuous morphological models and algorithms (a lecture course by Mestetskiy LM). URL: <http://www.machinelearning.ru/wiki/index.php?title=Morphmodels> (16.05.2017).

# EEG Beta Wave Trains Are Not the Second Harmonic of Mu Wave Trains in Parkinson's Disease Patients

Olga S. Sushkova<sup>1</sup>, Alexei A. Morozov<sup>1,2</sup>, Alexandra V. Gabova<sup>3</sup>

<sup>1</sup>*Kotel'nikov Institute of Radio Engineering and Electronics of RAS, Mokhovaya 11-7, Moscow, 125009, Russia*

<sup>2</sup>*Moscow State University of Psychology & Education, Sretenka 29, Moscow, 107045, Russia*

<sup>3</sup>*Institute of Higher Nervous Activity and Neurophysiology of RAS, Butlerova 5A, Moscow, 117485, Russia*

---

## Abstract

The goal of this study is development of a novel signal processing and analysis method for detailed investigation of the time-frequency dynamics of brain cortex electrical activity. The idea of our method of electroencephalograms (EEG) analyzing is in that we consider EEG signal as a composition of so-called wave trains. The wave train term is used to denote a signal localized in time, frequency, and space. We consider the wave train as a typical component of EEG, but not as a special kind of EEG signals.

In contrast to papers devoted to detecting wave trains of one or two specific types, such as alpha spindles and sleep spindles, we analyze any kind of wave trains in a wide frequency band. Using this method, we have found three interesting frequency areas where differences were detected between a group of Parkinson's disease (PD) patients and a control group of healthy volunteers. The goal of this work is to check whether the regularities in the mu and beta frequency bands are independent ones, that is, the beta wave trains observed in the analysis were not the second harmonics of the mu wave trains.

We have developed a special algorithm that eliminates from the analysis all beta wave trains in EEG signal that were observed simultaneously with the mu wave trains. Analysis of a real experimental data set processed by this algorithm has confirmed that the beta frequency band regularity is separate from the mu frequency band regularity. Moreover, a new significant difference between the left hand tremor and right hand tremor Parkinson's disease patients was discovered.

*Keywords:* Wave train, Wave packet, Burst, Electroencephalogram, EEG, Beta, Mu, Wavelet, Visualizing EEG data, Decrease of quantity of wave trains, Parkinson's disease

---

## 1. Introduction

The goal of our research is development of a novel signal processing and analysis method for detailed investigation of time-frequency dynamics of brain cortex electrical activity. The idea of our method of analyzing EEG is in that we consider EEG signal as a composition of so-called wave trains. In contrast to papers devoted to detecting wave trains of one or two specific types, such as alpha spindles [1] and sleep spindles [2, 3, 4, 5, 6, 7], we analyze any kind of wave trains in a wide frequency area. The developed method differs from analogous method for detailed analysis of time-frequency dynamics of EEG [8] in that the statistical analysis of samples of wave trains and a new method for visualizing the results of the analysis are proposed. The algorithm used for detecting wave trains also is different. In particular, a new kind of diagrams based on ROC curves was developed to visualize the neurophysiological data (see an example of the diagram in Figure 6, Section 2).

In physics, a wave train (or a wave packet) is a short "burst" or "envelope" of localized wave action that travels as a unit. In this paper, the wave train term is used to denote a signal localized in time, frequency, and space. We consider the wave train as a typical pattern in EEG signals. Recently we have demonstrated that the number of wave trains in the EEG beta frequency range (12–25 Hz) is significantly decreased in early stage Parkinson's disease patients [9, 10]. In previous paper [11], a method of visualization of EEG analysis results based on ROC curves was described. Using this method, we have revealed three interesting frequency ranges where differences between a Parkinson's disease patient group and a control group are detected. The first range is 7.5–9.5 Hz (approximately the mu frequency band),

the second is 10.5 – 13.5 Hz (also the mu frequency band), and the third range is 18 – 24 Hz (approximately the beta-2 frequency band). The presence of the first and second frequency bands gives an evidence for a shift of EEG mu rhythm to the lower frequency areas in Parkinson's disease patients. The third frequency band is a confirmation of the regularity reported in [9]. Note that in [9] another frequency band was investigated (12 – 25 Hz). As a consequence, the numerical characteristics of EEG given in this paper differ slightly from the numerical characteristics considered in the paper [9]. In addition, the frequency ranges of beta-1 and beta-2 were not separated in the paper [9], but in the course of further research we came to the conclusion that the characteristics of the wave trains in these beta sub-bands should be investigated separately.

The goal of this work is to check whether the regularities in the mu and beta frequency bands are independent ones, that is, the beta wave trains observed in the analysis were not the second harmonics of the mu wave trains. This check is important for understanding the development of neurodegenerative processes and the formation of compensatory neurophysiological mechanisms in Parkinson's disease, since the frequency ranges of mu and beta correspond to the work of different functional systems in the brain [12, 13, 14, 15]). We have developed a special algorithm that eliminates from the analysis all beta wave trains in EEG signal that were observed simultaneously with the mu wave trains. Analysis of a real experimental data set processed by this algorithm has confirmed that the beta frequency band regularity is separate from the mu frequency band regularity. Furthermore, a new significant difference between the left hand tremor and right hand tremor Parkinson's disease patients was discovered.

## 2. A Problem Statement

Let  $M$  be a local maximum in a wavelet spectrogram (see Figure 1). We estimate the full width at half maximum (FWHM) of  $M$  in the time plane  $FWHM_{TIME}$  and in the frequency plane  $FWHM_{FREQUENCY}$ . Then we check whether

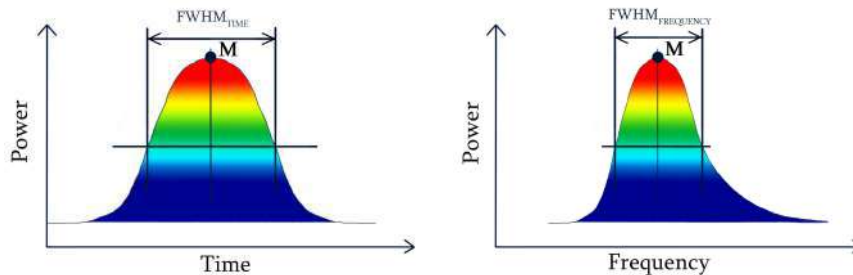


Figure 1: An example of a spectrogram of a wave train in a time-frequency domain. The diagram at the left shows the spectrogram of the signal in the time plane. The abscissa indicates a time and the ordinate indicates a power. The diagram at the right shows the spectrogram of the signal in the frequency plane. The abscissa indicates a frequency and the ordinate indicates a power.

there are no values in the rectangle area

$$FWHM_{TIME} \times FWHM_{FREQUENCY}$$

that are bigger than the  $M$  value (see Figure 2). We consider  $M$  as a case of a wave train if  $FWHM_{TIME}$  of  $M$  is greater or equal to the  $T_D$  threshold (see Figure 1). The  $T_D$  threshold is a function of the frequency  $f$  of the maximum  $M$ :

$$FWHM_{TIME} \geq T_D = N_P/f,$$

where  $N_P$  is a constant given by an expert. In this paper, we apply the value:  $N_P = 2$ .

An example of wave trains in the time-frequency domain in a PD patient is shown in Figure 3. The figure demonstrates the wave trains in the background EEG in the right hand tremor patient, the C3 cortex area. Each circle indicates a wave train in the wavelet spectrogram. It is obvious that most wave trains are located in the mu frequency band (8 – 12 Hz). In addition, there are wave trains in the delta (1 – 4 Hz), theta (4 – 8 Hz), and beta (12 – 30 Hz) frequency bands.

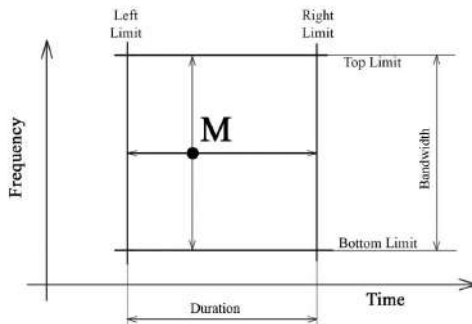


Figure 2: Time and frequency bounds of the  $M$  wave train in the wavelet spectrogram. The abscissa indicates the time and the ordinate indicates the frequency.

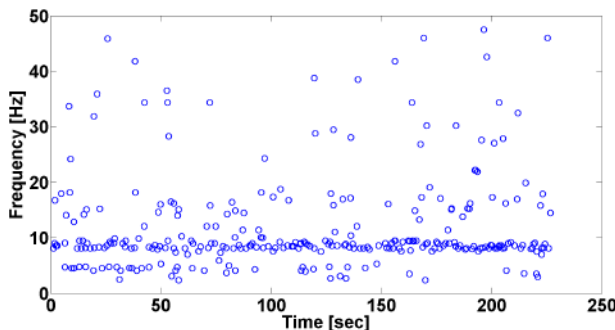


Figure 3: Wave trains of the PD patient in the time-frequency domain. A background EEG in the right hand tremor patient is presented, the C3 cortex area, the delta (1 – 4 Hz), theta (4 – 8 Hz), mu (8 – 12 Hz), and beta (12 – 30 Hz) frequency bands. Each circle indicates a wave train in the wavelet spectrogram. The abscissa is the time and the ordinate is the frequency.

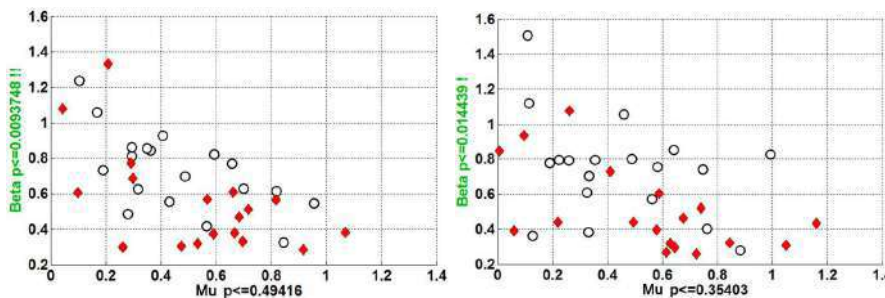


Figure 4: The scattering of the quantity of beta wave trains in the patients and the healthy volunteers. The C3 cortex area is shown at the left, the C4 cortex area is shown at the right. The abscissa is the quantity of wave trains per second in the mu frequency band. The ordinate is the quantity of wave trains per second in the beta frequency band. The patients are indicated by diamonds and the healthy volunteers are indicated by circles.

The method of EEG wave train analysis revealed a new effect in the Parkinson’s disease [9, 10]. The number of beta (12 – 25 Hz) wave trains in the C3 and C4 cortex areas is significantly decreased (Mann-Whitney,  $p < 0.02$ ), see Figure 4. The patients are indicated by red diamonds, and the healthy volunteers are indicated by white circles. Each diamond and each circle correspond to a particular person. The number of wave trains is standardized by the length of EEG record in seconds, because durations of EEG records were slightly different in the subjects. Note that in this test the quantity of the wave trains is considered only, but not the amplitude of the wave trains.

In previous paper [11], a method of visualization of results of EEG analysis based on ROC curves was described. Let  $MinFreq$ ,  $MaxFreq$  are frequency bounds of a four-dimensional area  $S$  in the space of the wave trains. Let

$MinPower$ ,  $MaxPower$  are power bounds of the area  $S$ ;  $MinDurat$ ,  $MaxDurat$  are duration bounds of the area  $S$ ; and  $MinBandwidth$ ,  $MaxBandwidth$  are bandwidth bounds of the area  $S$ . It was calculated a number of wave trains per second located in the area  $S$  in every individual patient and healthy volunteer and created histograms of the quantity of the wave trains per second (see an example in Figure 5). A statistical difference between the diagrams may indicate that the area  $S$  contains wave trains that are typical for Parkinson's disease patients, but not for the control group, or vice versa. Another interesting issue is whether it can be specified a threshold (a limit of the number of the wave trains in the area  $S$ ) that separates adequately the histograms, because the presence of such threshold means that the quantity of the wave trains in the area  $S$  may be used for the clinical diagnosis of Parkinson's disease. For instance, there is a strong statistical difference between the histograms in the Figure 5 (the Mann-Whitney test,  $p < 0.009$ ). The diagram demonstrates that a typical number of the wave trains in the control group is about 0.13 per second in the given frequency band. At the same time, a typical number of the wave trains in the patients is about 0.06.

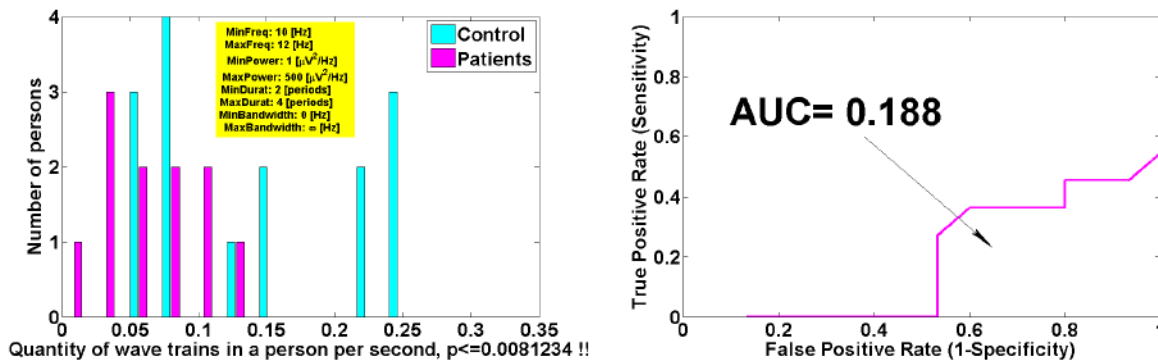


Figure 5: At the left: histograms of the quantity of the wave trains per second in the patients and the control group (the left hand tremor patients, the C3 cortex area). The wave trains are considered in the  $S$  space bounded by the following limits: a frequency range is 10 – 12 Hz, a power range is 1 – 500  $\mu V^2/Hz$ , the duration range is 2 – 4 periods, the bandwidth range has no limits. The patients' histogram is indicated by the dark magenta color; and the control group histogram is indicated by the light cyan color. At the right: a ROC curve based on the histograms. The abscissa indicates the False Positive Rate. The ordinate indicates the True Positive Rate. The area under the ROC curve ( $AUC$ ) indicates whether the area  $S$  is applicable for separation of the patients and the control group.  $AUC < 0.5$  indicates that the wave trains quantity is greater in the control group than in the patients.

Thus, in mathematical terms, the goal of our investigation was searching such areas in the multidimensional space of the wave trains, where  $AUC$  differs sufficiently from 0.5 and is approached to 1 or to 0.  $AUC > 0.5$  indicates that the wave trains quantity is greater in the patients than in the healthy volunteers. Similarly,  $AUC < 0.5$  indicates that the wave trains quantity is greater in the control group. An exhaustive search of the values  $MinFreq$ ,  $MaxFreq$ ,  $MinPower$ ,  $MaxPower$ ,  $MinDurat$ ,  $MaxDurat$ ,  $MinBandwidth$ , and  $MaxBandwidth$  can be implemented to investigate the multidimensional space, but we consider different slices of the space using various 2D and 3D diagrams not to miss any interesting regularities in the space of the wave trains. An example of this analysis is presented in the Figure 6.

Let us compute  $AUC$  values for various frequency ranges. In Figure 6, the functional dependence of  $AUC$  is shown, where the arguments of the function are the  $MinFreq$  and  $MaxFreq$  bounds. The frequency values varied from 2 to 25 Hz (with the 0.5 Hz step); the  $MinPower$ ,  $MaxPower$ ,  $MinDurat$ ,  $MaxDurat$ ,  $MinBandwidth$ , and  $MaxBandwidth$  were constant:  $MinPower = 1$ ,  $MaxPower = \infty$ ,  $MinDurat = 0$ ,  $MaxDurat = \infty$ ,  $MinBandwidth = 0$ ,  $MaxBandwidth = \infty$ . The upper left triangle of the diagram indicates the values of  $AUC$  corresponding to the  $MinFreq$ – $MaxFreq$  frequency range. The lower right triangle of the diagram indicates the  $AUC$  values corresponding to the total frequency band 2 – 25 Hz except the  $MaxFreq$  –  $MinFreq$  band. Note that in the lower right triangle of the diagram the  $MinFreq$  indicates the upper limit (but not the lower limit) of the excepted frequency band and the  $MaxFreq$  indicates the lower bound of the excepted values.

Using this method we have revealed three interesting frequency ranges where differences between a Parkinson's disease patient group and a control group were detected [11]. The first range is 7.5 – 9.5 Hz (approximately the mu frequency band), the second is 10.5 – 13.5 Hz (also the mu frequency band), and the third range is 18 – 24 Hz (approximately the beta-2 frequency band). The task of this paper is to check whether the regularities in EEG in the

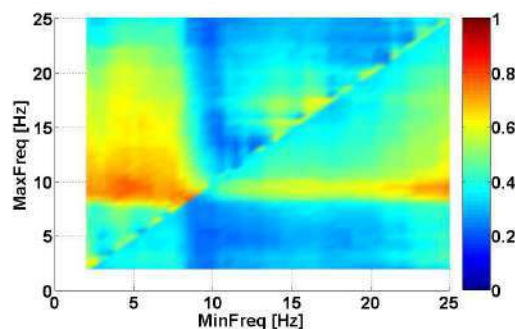


Figure 6: A diagram of *AUC* values calculated for various frequency bands. In the upper left triangle of the diagram: the abscissa is the lower bound of the frequency band and the ordinate is the upper bound of the frequency band. In the lower right triangle of the diagram: the abscissa is the upper bound of the *excluded* frequency band and the ordinate is the lower bound of the *excluded* frequency band. The frequency varied from 2 to 25 Hz with the 0.5 Hz step. The background EEG was analyzed, the right hand tremor patients, the C3 cortex area.

central cortex area in the frequency bands 7.5 – 13.5 Hz (mu) and 18 – 24 Hz (beta-2) are independent. In other words, we have to demonstrate that the beta wave trains observed in the analysis are not the second harmonics of the mu wave trains. In this paper we demonstrate that these regularities in the mu and beta frequency bands are independent ones.

### 3. Description of the Algorithm for Beta Wave Trains Elimination

We have developed a special algorithm that eliminates from the analysis all beta wave trains in EEG signal that were observed simultaneously with the mu wave trains. There are four hypothetical cases when the beta wave trains can be observed simultaneously with the mu wave trains (superpositions of the mu wave train and the beta wave train):

1. The time interval of the beta wave train is located inside the time interval of the mu wave train (Figure 7). The duration of the beta wave train is smaller than the duration of the mu wave train.
2. The time interval of the mu wave train is located inside the time interval of the beta wave train (Figure 8). The duration of the beta wave train is bigger than the duration of the mu wave train.
3. The left end of the beta wave train is intersected with the right end of the mu wave train (Figure 9).
4. The right end of the beta wave train is intersected with the left end of the mu wave train (Figure 10).

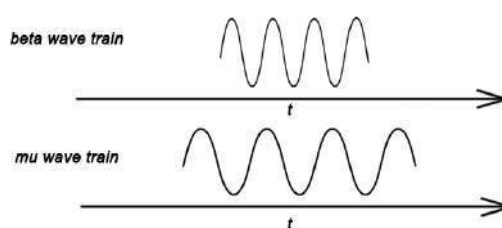


Figure 7: Superposition of a mu wave train and a beta wave train. The time interval of the beta wave train is located inside the time interval of the mu wave train.

The algorithm eliminates all beta wave trains that are intersected in time with any wave trains in the 2 – 14 Hz frequency range. Thus, all four cases are covered by this algorithm. In addition, the algorithm removes extra wave trains to exclude even the hypothetical possibility that the second or third harmonics of theta and mu EEG signals (4 – 12 Hz) are observed in the beta frequency range. This is done because EEG signals are not sine-shaped and the EEG wave train corresponds to a certain frequency band. Therefore, it is fundamentally impossible to predict exactly the frequency bands of the second and third harmonics of the wave trains under consideration.

Below we compare the results of EEG analysis with using and without using this algorithm of the wave train elimination.



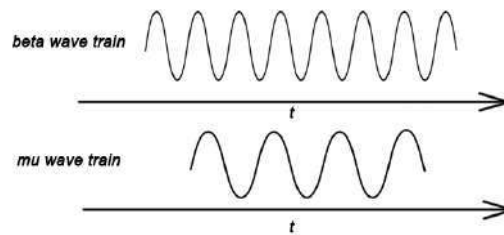


Figure 8: Superposition of a mu wave train and a beta wave train. The time interval of the mu wave train is located inside the time interval of the beta wave train.

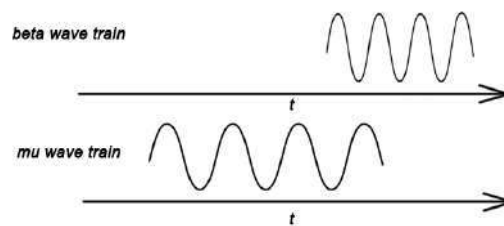


Figure 9: Superposition of a mu wave train and a beta wave train. The left end of the beta wave train is intersected with the right end of the mu wave train.

#### 4. The Experimental Setting

We considered a set of EEG wave trains detected in a group of de novo Parkinson's disease patients and a healthy volunteer group. The group of patients included 17 patients with right hand tremor and 11 patients with left hand tremor in the first stage of Parkinson's disease without Parkinson's disease treatment. The group of healthy volunteers included 15 people.

The ages of the patients were from 38 to 71 years old; the mean age was 60 years old. The ages of the healthy volunteers were from 48 to 81 years old; the mean age was 58 years old. No statistically significant differences between the patients' ages and the volunteers' ages were detected. The amount of the male patients was 11; the amount of the female patients was 17. The amount of the male healthy volunteers was 5; the amount of the female healthy volunteers was 10. The size of the groups is typical for a neurophysiological examination.

The patients were clinically diagnosed according to the standard Hoehn and Yahr scale. All patients and volunteers were right-handed. A standard 10x20 EEG acquisition schema was used for the data collection. A background EEG was recorded in standard conditions. Examined person sat in an armchair relaxing with arms disposing on the armrests and fingers dangling freely from the ends of armrests. The eyes were closed during the recordings. A 41-channel digital EEG system Neuron-Spectrum-5 (Neurosoft Ltd.) was used. The sampling rate was 500 Hz. The 0.5 Hz high-pass filter, the 35 Hz low-pass filter, and the 50 Hz notch filter were used. The duration of every record was about 3 minutes. The record was analyzed as is, without selection of areas in the signal.

In this paper, the C3 and C4 cortex areas are considered only, because these areas approximately correspond to the motor cortex areas and are situated in the scalp area that produces a minimal number of muscle artifacts.

#### 5. Results of the Analysis and Discussion

Before applying the algorithm of the elimination of beta wave trains intersected with the mu wave trains, no statistically significant differences between quantities of wave trains in the right hand tremor patients and the healthy volunteers in both the C3 and C4 cortex areas were observed (see Figure 11). Also there were no significant differences between the left hand tremor patients and healthy volunteers (see Figure 12). Note that in [9] the Mann-Whitney test revealed significant differences between the data samples. This inconsistency of the results can be explained by the fact that another frequency range is considered in comparison with the previous work [9] (the 18 – 30 Hz frequency

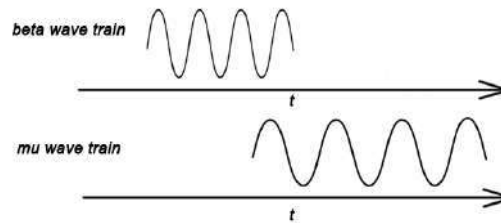


Figure 10: Superposition of a mu wave train and a beta wave train. The right end of the beta wave train is intersected with the left end of the mu wave train.

range is considered instead of the 12–25 Hz range). In addition, the right hand tremor patients and the left hand tremor patients are separated in present work. The algorithm of wave train detection also has been modified in comparison with the paper [9].

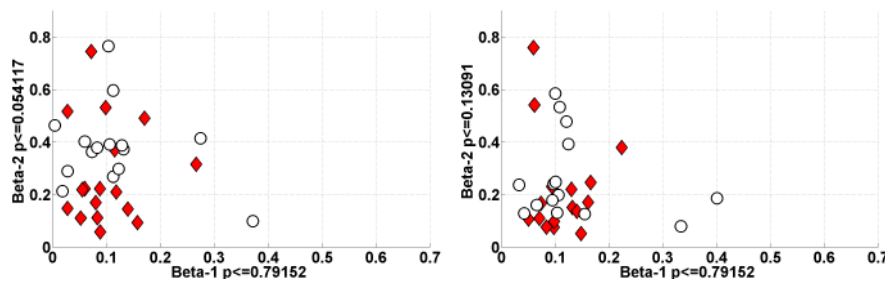


Figure 11: The scattering of the quantity of the beta wave trains in the right hand tremor patients and the healthy volunteers in the C3 cortex area (at the left) and in the C4 cortex area (at the right) *before* applying the algorithm of the elimination of beta wave trains intersected with the mu wave trains. There are no significant differences *before* the groups. The abscissa is the quantity of wave trains per second in the beta-1 frequency band. The ordinate is the quantity of wave trains in the beta-2 frequency band. The patients are indicated by diamonds and the healthy volunteers are indicated by circles.

After the processing of the experimental data by the wave trains elimination algorithm, the analysis of the data has confirmed that the beta band regularity does exist in its own. The statistically significant decrease of the beta wave train quantity was observed between the right hand tremor patients and the healthy volunteers in the C3 cortex area (18–30 Hz, Mann-Whitney,  $p < 0.02$ , see Figure 13, left). Note that there is no significant difference in the C4 cortex area (see Figure 13, right), however, there was a statistical trend ( $p < 0.19$ ). Thus, we have refined substantially the space and frequency localization of the investigated neurophysiological regularity in comparison with the paper [9].

The wave trains elimination algorithm decreases the quantity of beta-2 wave trains in all groups (the right hand tremor patients, the left hand tremor patients, and the control group) by approximately 30%. The fact that the statistically significant difference in the quantities of wave trains between the patients and the control group was detected by the Mann-Whitney test after the elimination of coincident mu and beta wave trains indicates that the eliminated wave trains are not a cause of the regularity. On the contrary, these wave trains were a noise that complicates the detection of the regularity.

It is interesting that a new neurophysiological regularity was discovered in the beta-2 frequency band after applying the algorithm. The statistically significant differences of quantities of the beta-2 wave trains were observed between the right hand tremor patients and the left hand tremor patients in both the C3 cortex area (Mann-Whitney,  $p < 0.03$ , see Figure 14, left) and the C4 cortex area (Mann-Whitney,  $p < 0.02$ , see Figure 14, right). The results of the test demonstrate that EEG of the right hand tremor patients and the left hand tremor patients are differed in both hemispheres.



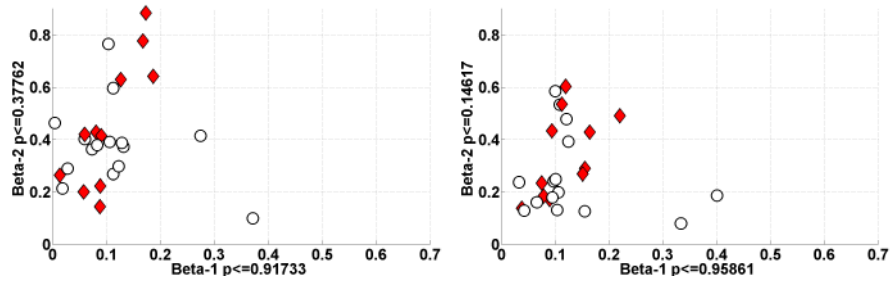


Figure 12: The scattering of the quantity of the beta wave trains in the left hand tremor patients and the healthy volunteers in the C3 cortex area (at the left) and in the C4 cortex area (at the right) *before* applying the algorithm of the elimination of beta wave trains intersected with the mu wave trains. There are no significant differences between the groups. The abscissa is the quantity of wave trains per second in the beta-1 frequency band. The ordinate is the quantity of wave trains in the beta-2 frequency band. The patients are indicated by diamonds and the healthy volunteers are indicated by circles.

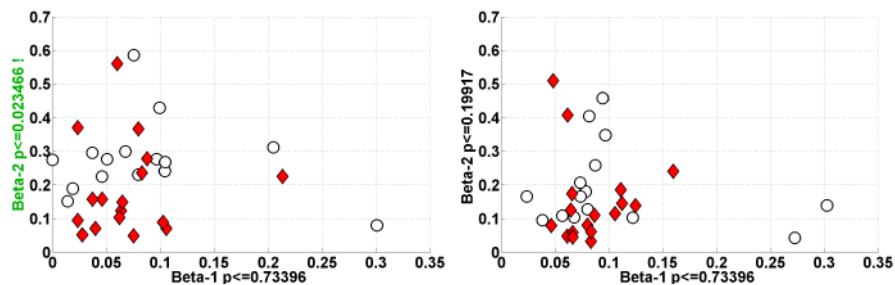


Figure 13: The scattering of the quantity of the beta wave trains in the right hand tremor patients and the healthy volunteers in the C3 cortex area (at the left) and in the C4 cortex area (at the right) *after* the processing the experimental data by the wave trains elimination algorithm. There is a significant difference between the groups at the left, but there is no significant difference between the groups at the right. The abscissa is the quantity of wave trains per second in the beta-1 frequency band. The ordinate is the quantity of wave trains per second in the beta-2 frequency band. The patients are indicated by diamonds and the healthy volunteers are indicated by circles.

## 6. Conclusions

A new method of signal processing and analysis for detailed investigation of time-frequency dynamics of the cortex electrical activity is developed. A distinctive feature of the method is a possibility to separate and analyze a specified set of wave trains in EEG signals. The results of the research give evidence that EEG analysis method based on the wave trains is prospective for:

- Looking for group statistical regularities in the early stages of Parkinson's disease that gives a basic knowledge about the disease and compensatory mechanisms in the brain cortex.
- Searching EEG features that are prospective for the early stages of Parkinson's disease diagnostics.

Using a special algorithm, we have demonstrated that a neurophysiological regularity discovered in the beta frequency band of EEG signals [9] is not an artifact caused by another regularity observed in the mu frequency band. In the course of the investigation, also, another neurophysiological regularity is discovered in the beta frequency band. The results of the analysis indicate that the right hand tremor patients and the left hand tremor patients differed in both hemispheres. This may be evidence that right hand tremor and left hand tremor Parkinson's disease patients must be diagnosed in different ways. Probably different approaches are necessary for treatment of the left hand tremor and right hand tremor patients as well.

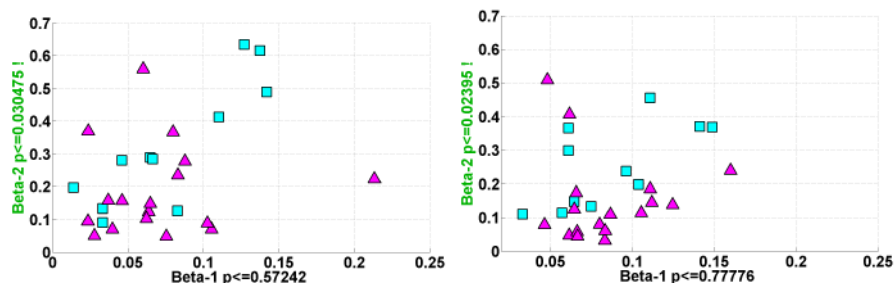


Figure 14: The scattering of the quantity of the beta wave trains in the right hand tremor patients and the left hand tremor patients in the C3 cortex area (at the left) and in the C4 cortex area (at the right) after the processing the experimental data by the wave trains elimination algorithm. There is a significant difference between the groups. The abscissa is the quantity of the wave trains per second in the beta-1 frequency band. The ordinate is the quantity of the wave trains in the beta-2 frequency band. The right hand tremor patients are indicated by triangles and the left hand tremor patients are indicated by squares.

## 7. Acknowledgment

Authors are grateful to Alexei V. Karabanov for selection and medical examination of the patients, to Galina D. Kuznetsova and Alexander F. Polupanov for co-operation and a help in the research, and Yuriy V. Obukhov and Mikhail N. Ustinin for a help in the statement of the problem. We acknowledge a partial financial support from the Russian Foundation for Basic Research, grant 15-07-07846.

## References

- [1] Lawhern V., Kerick S., Robbins K. A. Detecting alpha spindle events in EEG time series using adaptive autoregressive models // *BMC Neuroscience*. — 2013. — Vol. 14:101. <http://www.biomedcentral.com/1471-2202/14/101>.
- [2] Determination of dominant simulated spindle frequency with different methods / E. Huupponen, W. D. Clercq, G. Gómez-Herrero et al. // *Journal of Neuroscience Methods*. — 2006. — Vol. 156. — Pp. 275–283.
- [3] Sleep spindle detection through amplitude-frequency normal modelling / A. Nonclercq, C. Urbain, D. Verheulpen et al. // *Journal of Neuroscience Methods*. — 2013. — Vol. 214. — Pp. 192–203.
- [4] Improved spindle detection through intuitive pre-processing of electroencephalogram / A. Jaleel, B. Ahmed, R. Tafreshi et al. // *Journal of Neuroscience Methods*. — 2014. — Vol. 233. — Pp. 1–12.
- [5] Camilleri T. A., Camilleri K. P., Fabri S. G. Automatic detection of spindles and K-complexes in sleep EEG using switching multiple models // *Biomedical Signal Processing and Control*. — 2014. — Vol. 10. — Pp. 117–127.
- [6] Sleep spindle detection using time-frequency sparsity / A. Parekh, I. Selesnick, D. Rapoport, I. Ayappa // *IEEE Signal Processing in Medicine and Biology Symposium*. — Philadelphia, PA: IEEE, 2014. — Pp. 1–6.
- [7] O'Reilly C., Nielsen T. Automatic sleep spindle detection: benchmarking with fine temporal resolution using open science tools // *Frontiers in Human Neuroscience*. — 2015. — Vol. 9:353. <http://doi.org/10.3389/fnhum.2015.00353>.
- [8] Obukhov Y., Korolyov M., Gabova A. et al. Patent No 2484766 Rossiiskaya Federacia. Sposob rannei electroencephalographicheskoi diagnostiki boleznii Parkinsona. 20.06.2013. — 2013. — In Russian.
- [9] Sushkova O., Morozov A., Gabova A. A method of analysis of EEG wave trains in early stages of Parkinson's disease // *International Conference on Bioinformatics and Systems Biology (BSB-2016) / IEEE*. — 2016. — Pp. 1–4.
- [10] Sushkova O. S., Morozov A. A., Gabova A. V. Development of a method of analysis of EEG wave packets in early stages of Parkinson's disease // *Proceedings of the International conference Information Technology and Nanotechnology (ITNT 2016, Samara, Russia, May 17–19, 2016)*. — Samara: CEUR, 2016. — Pp. 681–690. <http://ceur-ws.org/Vol-1638/Paper82.pdf>.
- [11] Data mining in EEG wave trains in early stages of Parkinson's disease / O. S. Sushkova, A. A. Morozov, A. V. Gabova, A. V. Karabanov // *Proceedings of the 12th Russian-German Conference on Biomedical Engineering (RGC XII, Suzdal, July 4-7 2016)*. — Suzdal: Vladimir State University, 2016. — Pp. 80–84.
- [12] Neural mass models describing possible origin of the excessive beta oscillations correlated with Parkinsonian state / C. Liu, Y. Zhu, F. Liu et al. // *Neural Networks*. — 2017. — Vol. 88. — Pp. 65–73.
- [13] Pavlides A., Hogan S. J., Bogacz R. Computational models describing possible mechanisms for generation of excessive beta oscillations in Parkinson's disease // *PLOS Computational Biology*. — 2015. — no. 12. — Pp. 1–29.
- [14] Kapitsa I., Nerobkova L., Voronina T. EEG correlates of an early stage of a Parkinson illness in experiment on mice of the strain C57BL/6 // *Biomeditsina*. — 2014. — no. 1. — Pp. 54–60. — In Russian.
- [15] Resting state oscillatory brain dynamics in Parkinson's disease: An MEG study / J. Bosboom, D. Stoffers, C. Stam et al. // *Clinical Neurophysiology*. — 2006. — Vol. 117. — Pp. 2521–2531.

# Effectiveness of correlation and information measures for synthesis of recurrent algorithms for estimating spatial deformations of video sequences

A.G. Tashlinskiy<sup>1</sup>, A.V. Zhukova<sup>1</sup>

<sup>1</sup>Ulyanovsk State Technical University, Severnii Venets, 32, 432027, Ulyanovsk, Russia

---

## Abstract

A comparative analysis of the efficiency of correlation (cross-correlation, Tanimoto coefficient and Kendall's rank correlation coefficient) and information (mutual information of Tsallis and Shannon, F-information measure and entropy of the joint probability distribution) measures of image similarity for the synthesis of recursive estimation algorithms is presented for the problem of estimating parameters of spatial deformations of a sequence of images. Unbiased additive Gaussian noise was used as an interfering factor in the experimental studies. It is shown that the potentially high convergence rate of the estimated parameters and the smaller variance of the estimation error from the investigated correlation measures are ensured by the Tanimoto coefficient, and from the I-information of the F-information among the information measures. According to these criteria, the Kendall's rank correlation coefficient and the M-measure of F-information are inferior, respectively.

*Keywords:* image; recurrent estimation; similarity measures; cross-correlation; Tanimoto coefficient; Kendall's rank correlation coefficient; Tsallis mutual information; Shannon mutual information; F-information measures

---

## 1. Introduction

The detection and evaluation of spatial changes (deformations) in a sequence of images  $\mathbf{Z}^{(n)}$ ,  $n = \overline{1, N}$  is one of the key tasks of processing video sequences. Various approaches to the solution of this problem are implemented in the frequency [1] and in the spatial [2] domains.

In this case, the solution is usually based on the search for an extremum of some similarity measure (SM) between each pair of adjacent images  $\mathbf{Z}^{(n)} = \{z_{\bar{j}}^{(n)}\}$  and  $\mathbf{Z}^{(n+1)} = \{z_{\bar{j}}^{(n+1)}\}$ , where  $\bar{j}$  – coordinates of the nodes of the grid of samples on which the images are defined. With the assumed model of interframe spatial deformations of images, estimates are sought for the strain parameters  $\bar{\alpha}$  of one of the images at which the extremum of the SM is reached.

When recursively searching for the SM extremum at each iteration  $t$  of the estimation the current estimates  $\bar{\alpha}$  of the spatial deformation parameters are corrected by a certain amount in the direction  $\bar{d}(\hat{\alpha}, Z_t)$  [3]:

$$\hat{\alpha}_{t+1} = \hat{\alpha}_t - \Lambda_t \bar{d}(\hat{\alpha}, Z_t), \quad (1)$$

where  $\Lambda_t$  — a positively defined matrix (usually diagonal);  $t = \overline{1, T}$  — iteration number.

The direction  $\bar{d}(\cdot)$  is defined by the SM gradient estimated using a small subsample [4]  $Z_t = \{z_{\bar{j}_t}^{(n)}, z_{\bar{j}_t}^{(n+1)}\}$  of (usually random) points  $z_{\bar{j}_t}^{(n)} \in \tilde{\mathbf{Z}}^{(n)}$  and  $z_{\bar{j}_t}^{(n+1)} \in \mathbf{Z}^{(n+1)}$  drawn on  $t$ -th iteration, where  $\tilde{\mathbf{Z}}^{(n)}$  - interpolated image obtained using current deformation parameters' estimates  $\hat{\alpha}_t$  of image  $\mathbf{Z}^{(n)}$  [5]. In order to improve the accuracy of parameters' estimates  $\hat{\alpha}$  we need to increase a number of points in the subsamples but it will lead to increase in computational time.

There are many SM for images [6]. The choice of a particular SM is determined by the conditions and requirements of the applied problem, the nature of the possible spatial deformations of the video sequence, the properties of the images and the interfering factors.

## 2. Problem formulation

The chosen SM largely determines the potential effectiveness of the recursive algorithms for estimating spatial deformations in a sequence of images synthesized on its basis. However, criteria that allow a priori evaluation of the potential efficiency of algorithms by SM are poorly investigated. A variant of the solution of this problem was considered in [7,8]. In this work several correlation correlations have been selected for the study: cross-correlation (the coefficient of interframe correlation) [9], the Tanimoto coefficient [10] the Kendall's rank correlation coefficient [11], and a number of informational SMs: Tsallis [12] and Shannon mutual information, F-information measures, and the entropy of the joint probability distribution [13]. Unbiased additive Gaussian noise was used as an interfering factor in the studies. The relation between SM characteristics and the probabilistic properties of estimates of the parameters of spatial deformations formed by recurrent algorithms synthesized on the basis of these measures was also investigated.

A method for calculating the probabilistic properties of estimates of the parameters of inter-frame spatial deformations of images for a finite number of iterations of recurrent estimation [14] was proposed in [16]. It is based on finding the probabilities of demolition at each iteration (the probabilities of changing the estimates of the investigated parameters  $\hat{\alpha}$  towards optimal

values). With allowance for (1) the probability of demolishing of the estimates at the next iteration can be interpreted as the probability that the projection of the gradient of the SM gradient on the axis of this parameter will be negative. In [15] this characteristic was used to find the error in estimates of the parameters of inter-frame spatial deformations of images formed by relay procedures of the form (1). However, this approach can be used for recurrent procedures of other classes.

Since procedure (1) is based on the use of SM gradient estimates its capabilities are largely determined by the nature of the slope  $K$  of the SM used in the synthesis of the procedure. Fig. 1 shows examples of the dependence of the slope of the SMs chosen for the study on the magnitude of the parallel shift  $h$  of the images, where  $h = 0 - 50$  is the shift in the grid steps of the image counts. Curves of the curvature module of the SM for correlation measures are shown on the left graph, where curve 1 corresponds to the Kendall rank correlation coefficient, curve 2 - to the Tanimoto coefficient, curve 3 - to the interframe correlation coefficient (CC). The right graph shows the dependencies for information SMs, here curve 1 corresponds to Shannon MI, curve 2 - to Tsallis MI, curve 3 - to I-measure of F-information, curve 4 - to M-measure of F-information, curve 5 - to excluding F-information, curve 6 - the entropy of the joint probability distribution (JPD). The legend is the same for all other figures in this work.

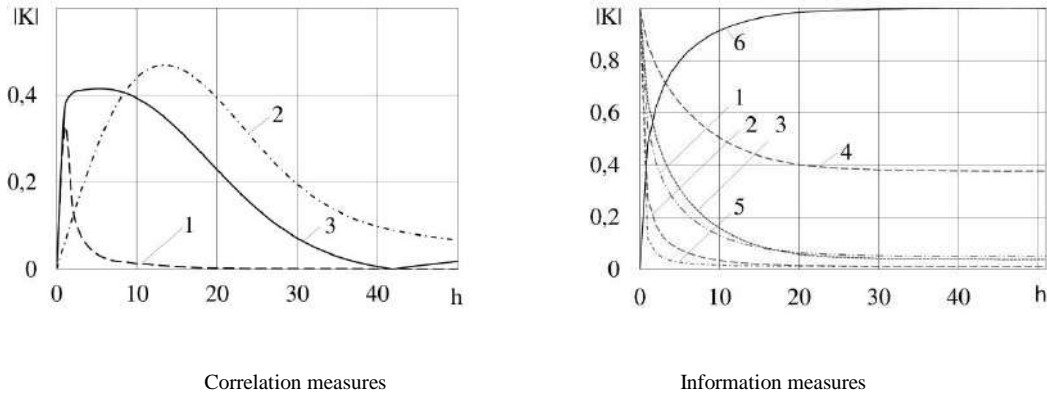


Fig. 1. Slope of the studied SM.

From Fig. 1 it can be seen that the nature of the slope of different SMs differs significantly, which must affect the potential characteristics of the recurrent evaluation procedures synthesized on their basis. A number of efficiency criteria based on an analysis of the nature of the SM slope was proposed in [8]. We will consider these criteria with reference to the correlation and information SMs chosen for the study.

### 3. Studied image similarity measures

#### 3.1. Probability-based measures

CC (cross-correlation) [9] is one of the most popular similarity measures. CC can be defined as

$$r = \frac{1}{\mu \hat{\sigma}_{z_n} \hat{\sigma}_{z_{n+1}}} \sum_{\tilde{j}_t \in \Omega_t} (\tilde{z}_{\tilde{j}_t}^{(n)} - M[\tilde{\mathbf{Z}}^{(n)}]) (z_{\tilde{j}_t}^{(n+1)} - M[\mathbf{Z}^{(n+1)}]),$$

where  $\mu$  — the number of points in a subsample;  $M[\mathbf{Z}] = \sum_{\tilde{j}_t \in \Omega_t} z_{\tilde{j}_t} / \mu$  — image mean estimation  $\mathbf{Z}$ ;

$\hat{\sigma}_{z_n}^2 = \sum_{\tilde{j}_t \in \Omega_t} (z_{\tilde{j}_t} - M[\mathbf{Z}])^2 / \mu$  — estimation of image  $\mathbf{Z}$  variance. Correlation coefficient  $r$  changes between  $-1$  and  $+1$ . The

value  $r = +1$  means the full linear relation,  $r = -1$  — negative linear relation. If  $r$  is different to  $\pm 1$  then the relation of

corresponding pixels  $\tilde{z}_{\tilde{j}_t}^{(n)}$  and  $z_{\tilde{j}_t}^{(n+1)}$  can be described as:  $z_{\tilde{j}_t}^{(n+1)} = \frac{\hat{\sigma}_{z_n}}{\hat{\sigma}_{z_{n+1}}} (\tilde{z}_{\tilde{j}_t}^{(n)} - M[\tilde{\mathbf{Z}}^{(n)}]) + M[\mathbf{Z}^{(n+1)}]$  the measure of relation

linearity between  $\mathbf{Z}^{(n)}$  and  $\mathbf{Z}^{(n+1)}$  which allows us to efficiently use correlation coefficient in case of additive noise or linear intensity distortions. In terms of computational complexity, CC is very effective, as it requires a small number of additions and multiplications for each element in the subsample. It is in the order of  $\mu$ .

Tanimoto coefficient [10] between two images  $\tilde{\mathbf{Z}}^{(n)}$  and  $\mathbf{Z}^{(n+1)}$  is defined as:

$$S_T = \frac{\sum_{\tilde{j}_t \in \Omega_t} \tilde{z}_{\tilde{j}_t}^{(n)} z_{\tilde{j}_t}^{(n+1)}}{\sum_{\tilde{j}_t \in \Omega_t} \tilde{z}_{\tilde{j}_t}^{(n)} z_{\tilde{j}_t}^{(n+1)} + \sum_{\tilde{j}_t \in \Omega_t} (\tilde{z}_{\tilde{j}_t}^{(n)} - z_{\tilde{j}_t}^{(n+1)})^2}. \quad (2)$$

In comparison with CC in Tanimoto coefficient normalization of intensity multiplication with respect to their standard deviations is replaced by the normalization with respect to the sum of squared differences between corresponding sample counts, which effects in the same way. Using the inner product of intensities in the denominator of Tanimoto coefficient gives

the same effect as the normalization with respect to mean values of images. Computational complexity of Tanimoto coefficient is on the same order as for correlation coefficient.

*Kendall rank correlation coefficient* [11]. If  $\tilde{z}_i^{(n)}$  and  $z_j^{(n+1)}$  are intensities of corresponding image pixels then for their differences  $(\tilde{z}_i^{(n)} - \tilde{z}_j^{(n)})$  and  $(z_i^{(n+1)} - z_j^{(n+1)})$  when  $i \neq j$  there are two possible situations: concordance - when  $\text{sign}(\tilde{z}_i^{(n)} - \tilde{z}_j^{(n)}) = \text{sign}(z_i^{(n+1)} - z_j^{(n+1)})$  and mismatch - when  $\text{sign}(\tilde{z}_i^{(n)} - \tilde{z}_j^{(n)}) = -\text{sign}(z_i^{(n+1)} - z_j^{(n+1)})$ . If we take a large sample from images  $\tilde{\mathbf{Z}}^{(n)}$  and  $\mathbf{Z}^{(n+1)}$  and the number of concordances is greater than the number of mismatches then we can conclude that image intensities are bounded. Let assume that from  $\mu/2$  pixel pairs are concordances and  $N_d$  are mismatches then Kendall correlation coefficient can be defined as follows [9]:

$$\tau = \frac{2(N_c - N_d)}{\mu(\mu - 1)}. \quad (3)$$

Kendal correlation coefficient is one of the most complex measures in terms of computational complexity. It requires concordance and mismatch computations for  $\mu(\mu - 1)/2$  corresponding pixel pairs. Therefore, its computational complexity of  $\tau$  is on the order of  $\mu^2$  operations.

### 3.2. Information measures

*Shannon mutual information (MI)* is one of the most widely used similarity measure in image registration [12,13] as it provides an extremely high accuracy when images have linear and non-linear intensity distortions, occlusions and also in case of additive noise and multimodal images. Generalized Shannon MI can be defined in terms of entropy as:

$$J(\hat{\alpha}, \tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}) = H(\tilde{\mathbf{Z}}^{(n)}) + H(\mathbf{Z}^{(n+1)}) - H(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}), \quad (4)$$

where  $H(\mathbf{Z}) = -\sum_i p_z(z_i) \log p_z(z_i)$  - image  $\mathbf{Z}$  entropy estimation;  $p_z$  - marginal PDF estimation of the image sample;  $H(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}) = -\sum_i \sum_k p_{z_n, z_{n+1}}(z_i, z_k) \log p_{z_n, z_{n+1}}(z_i, z_k)$  - joint entropy estimation;  $p_{z_n, z_{n+1}}$  - joint PDF estimation of intensities.

*Tsallis MI* is defined as [12]:

$$R_g = S_g(\tilde{\mathbf{Z}}^{(n)}) + S_g(\mathbf{Z}^{(n+1)}) + (1 - q) S_g(\tilde{\mathbf{Z}}^{(n)}) S_g(\mathbf{Z}^{(n+1)}) - S_q(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}), \quad (5)$$

where  $S_g(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}) = (g - 1)^{-1} \left( 1 - \sum_{i=0}^N \sum_{j=0}^N p_{i,j}^g \right)$  - Tsallis entropy of the order  $g$ ,  $g$  - a real number. When  $g=1$  Tsallis entropy approaches Shannon entropy.

*F-information measures* are based on divergence or distance between joint probability distributions and multiplication of marginal distribution of image pair. A class of divergence measures that uses MI is the F-information measures. *F-information measures* are [13]:

$$\begin{aligned} I\text{-measure:} \quad I_\alpha &= \frac{1}{\alpha(\alpha - 1)} \left( \sum_{i=0}^N \sum_{j=0}^N \frac{p_{i,j}^\alpha}{(p_i p_j)^{\alpha-1}} - 1 \right), \\ M\text{-measure:} \quad M_\alpha &= \sum_{i=0}^N \sum_{j=0}^N \left| p_{i,j}^\alpha - (p_i p_j)^\alpha \right|^{\frac{1}{\alpha}}, \\ \chi\text{-measure:} \quad \chi_\alpha &= \sum_{i=0}^N \sum_{j=0}^N \frac{|p_{i,j} - p_i p_j|^\alpha}{(p_i p_j)^{\alpha-1}}. \end{aligned} \quad (6)$$

$I$ -measure is defined for  $\alpha \neq 0$ ,  $\alpha \neq 1$ , and when  $\alpha = 1$  it approaches Shannon MI.  $M$ -measure is defined for  $0 \leq \alpha \leq 1$ , and  $\chi$ -measure - for  $\alpha > 1$ .

*Exclusive F-information* is related to entropy of JPD and Shannon MI:

$$D_f(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}) = 2H(\tilde{\mathbf{Z}}^{(n)}, \mathbf{Z}^{(n+1)}) - H(\tilde{\mathbf{Z}}^{(n)}) - H(\mathbf{Z}^{(n+1)}). \quad (7)$$

*Entropy of JPD* is defined as:

$$E = \sum_{i=0}^N \sum_{j=0}^N p_{i,j}^2, \quad (8)$$

where  $p_{i,j}$  - an element of JPD which can be estimated, for example using histograms. The stronger the relationship between two variables the smaller the entropy of JPD.

## 4. Effectiveness analysis of similarity measures

### 4.1. Effectiveness criteria

As already noted, a number of criteria for the effectiveness of SM's use in the synthesis of recurrent procedures for estimating spatial deformations of images were proposed in [8]. They are based on an analysis of the slope of SM depending on the Euclidean mismatch distance [16]. Among the characteristics that determine efficiency, the maximum slope  $K_{max}$  in the area of interest of the parameters being evaluated, the effective range of the parameters  $P$ , and the region of curvature growth  $S$  can be attributed.

It is shown that the maximum slope of SM corresponds to the maximum probability of improving the estimates of the parameters of interframe spatial deformations toward optimal values, and this extremum determines the potential rate of convergence of the estimation vector [7].

The effective range in this work refers to as subdomain of the domain of definition of registration parameters in which the required accuracy values are attained under given constraints, e.g. computational complexity, number of iteration etc. A condition under which a point of the space of estimated parameters falls into the effective range is that the slope of the SM at this point should exceed a certain critical value which does not guarantee the required convergence rate of the estimate vector  $\bar{a}$  any longer. Note that the actual effective range also depends on many other factors, in particular, the parameters of the estimation algorithm, the type of spatial deformation, so the critical value of the slope here is considered, in fact, as a criterion for determining the range of values of the discrepancies of the estimated parameters, for which the probability of demolition of estimates exceeds the specified threshold value.

### 4.2. Experimental results

For SM efficiency analysis it is reasonable to use simulated images whose intensity PDF and correlation function can be priori defined during their synthesis. In conducted experiments simulated images based on wave model [17] with intensity PDF and correlation function close to Gaussian were used. In addition, an unbiased white noise was used as a disturbing factor.

For example, Fig. 2 - Fig. 3 show the graphs of the dependence of the above efficiency indicators on the noise/signal ratio  $q$  (in terms of variances), where the left graph corresponds to the data for the correlation ones, and the right graph for the information SMs. As a mismatch, for simplicity, a parallel image shift was selected. We also note that the choice of parallel shift of images as frame-wise deformations doesn't weakens the generalization of the examination, since for any set of strain parameters the result of their effect on each point of the image can be recalculated through the Euclidean mismatch distance [16] into the vector of its shift relative to the initial position. That is, the result of any deformations can be represented by the PDF of such shifts.

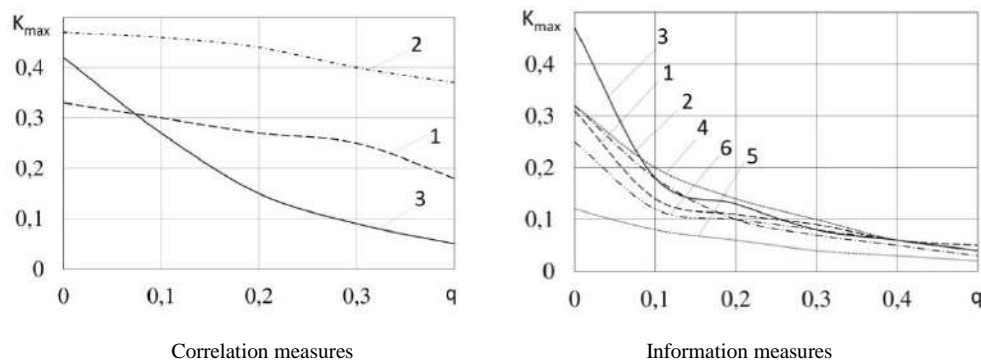


Fig. 2. Maximum slope of SM.

Analysis of fig. 2 shows that by the criterion of the maximum slope among the correlation SM the best result is provided by the Tanimoto coefficient whose maximum slope decreases most slowly with increasing noise intensity. The next by the result is the Kendall coefficient. CC at large noise is much inferior to them, however, with  $q < 0.07$  the maximum slope CC exceeds the  $K_{max}$  of Kendall coefficient. Investigations of information SMs showed that the potentially high convergence rate of the parameters being evaluated is provided by the I-measure of the F-information. The next most effective M-measure of F-information, which in large noises is more effective than the I-measure. Shannon and Tsallis MI, as well as the entropy of the JPD give close characteristics. Essentially they are inferior to the excluding F-information.

According to the effective range criterion (fig. 3), the best results among the correlation SMs also gave the use of the Kendall rank correlation coefficient. Then, with a margin of up to 40%, the coefficient of Tanimoto and CC, respectively. Among the information measures for small noise ( $q < 0.15$ ) the best results were shown by the I-measure of the F-information and the entropy of the joint JPD. This is followed by Tsallis and Shannon MI and the M-measure of F-information. For large noise ( $q > 0.25$ ), the M-measure and the I-measure of the F-information provide a larger effective range. Approximately the same parameters show the entropy of the JPD, Shannon and Tsallis MI. We note that the slope of the I-measure of the F-

information decreases significantly more rapidly with increasing noise than in the M-measure of the F-information. The worst results were shown by the excluding F-information.

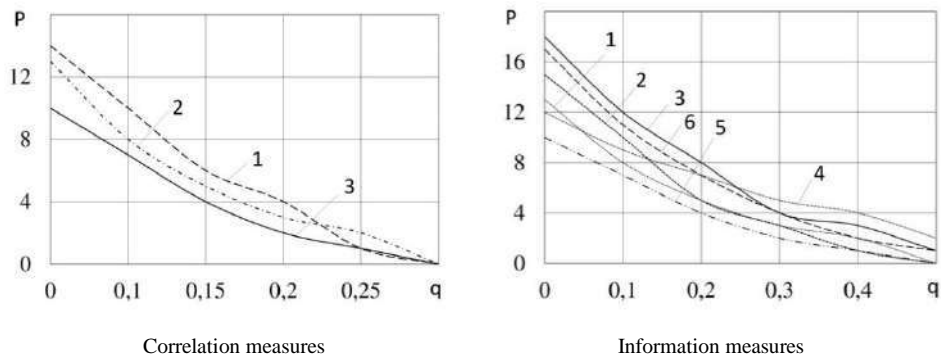


Fig. 3. Effective range of SM.

## 5. Conclusion

There are many SM of images on the basis of which recursive algorithms for estimating spatial deformations in a sequence of images can be synthesized. However, criteria that allow a priori evaluation of the potential efficiency of algorithms by the SM function are not well studied. A number of correlation (CC, Tanimoto coefficient and Kendall rank correlation coefficient) and informational (Tsallis and Shannon mutual information, F-information measures and joint probability distribution energy) SM are investigated based on the criteria of maximum slope that determines the potential rate of convergence of deformation parameter estimates and effective range, which refers to the subdomain of the domain of deformation parameters, in which the required indicators of accuracy and reliability of estimation are provided.

Studies have shown that, according to the criterion of maximum slope from correlation SM, the best result is provided by the Tanimoto coefficient, the maximum slope of which decreases most slowly with increasing noise intensity, and among the information I-measure of F-information. The worst results were shown by the CC and the excluding F-information, respectively.

By the criterion of the effective range, the best results among the correlation SMs were provided by the Kendall rank correlation coefficient. Then, respectively, the Tanimoto and CC coefficients. Among the information measures, the I-measure and the M-measure of the F-information showed greater resistance to noise. In this case, the slope of the I-measure of the F-information decreases significantly faster with increasing noise than in the M-measure.

## Acknowledgement

The reported study was supported by RFBR and Government of Ulyanovsk region, project № 16-41-732053, and RFBR grant № 15-41-02087.

## References

- [1] Gonzalez RC, Woods RE. Digital image processing. New Jersey: Prentice Hall, 2002; 793 p.
- [2] Goshtasby AA. Image registration. Principles, tools and methods: Advances in Computer Vision and Pattern Recognition. Springer, 2012; 441 p. DOI: 10.1007/978-1-4471-2458-0.
- [3] Theodoridis S, Koutroumbas K. Pattern Recognition, 4th edn. New York: Academic Press, 2009; 984 p.
- [4] Vasiliev, K.K. Statistical analysis of multidimensional images / K.K. Vasiliev, V.R. Krashennnikov. - Ulyanovsk: UIGTU, 2007. – 170 p.
- [5] Krashennnikov VR. The fundamentals of image processing theory. Ulyanovsk: UIGTU, 2003; 152 p.
- [6] Tashlinskiy AG. Estimation of image matching parameters for image sequences. Ulyanovsk: UIGTU, 2000; 131 p.
- [7] Brown LG. A survey of image registration techniques. ACM Computing surveys 1992; 24: 325–376. DOI: 10.1145/146370.146374.
- [8] D'Agostino E, Maes F, Vandermeulen D, Suetens P. An information theoretic approach for non-rigid image registration using voxel class probabilities. Med Image Anal. 2006; 6(3): 413–431. DOI: 10.1007/978-3-540-39701-4\_13.
- [9] De Castro E, Morandi C. Registration of translated and rotated images using finite Fourier transform. IEEE Transactions on Pattern Analysis and Machine Intelligence 1987; 9(5): 700–703. DOI: 10.1109/TPAMI.1987.4767966/.
- [10] Kendall MG. A new measure of rank correlation. Biometrika 1938; 30: 81–93. DOI: 10.2307/2332226.
- [11] Sevim YA, Atasoy A. Performance comparison of new nonparametric independent component analysis algorithm for different entropic indexes. Turkish Journal of Electrical Engineering & Computer Sciences 2012; 20: 287–297. DOI:10.3906/elk-1004-1.
- [12] Tashlinskii AG. Computational expenditure reduction in pseudo-gradient image parameter estimation. Lecture Notes in Computer Science 2003; 2658: 456–462. DOI: 10.1007/3-540-44862-4\_48.
- [13] Tashlinskii AG. Pseudogradient Estimation of Digital Images Interframe Geometrical Deformations. Vision Systems: Segmentation & Pattern Recognition. Vienna, Austria: I-Tech Education and Publishing, 2007: 465–494. DOI: 10.5772/4975.
- [14] Tashlinskii AG. Optimization of goal function pseudogradient in the problem of interframe geometrical deformations estimation. Pattern Recognition Techniques, Technology and Applications. Vienna, Austria: I-Tech. 2008: 249–280. DOI: 10.5772/4975.
- [15] Voronov SV, Tashlinskii AG. Efficiency analysis of information theoretic measures in image registration. Pattern recognition and image analysis 2016; 26(3): 502–505. DOI: 10.1134/S1054661816030226.
- [16] Voronov SV. The use of mutual information as objective function for image parameters' estimation. Radiotekhnika 2014; 7: 88–94.
- [17] Tashlinskiy AG, Tikhonov VO. Method for error estimation for stochastic gradient estimation of multidimensional processes. Izvestiya vuzov, seriya "Radioelektronika" 2001; 44(9): 75–80.

# Optimal bandwidth selection in geographically weighted factor analysis for education monitoring problems

A.Timofeeva<sup>1</sup>, K.Tesselkina<sup>1</sup>

<sup>1</sup>Novosibirsk State Technical University, 20 Prospekt K. Marksa, 630073, Novosibirsk, Russia

---

## Abstract

Geographically weighted models are widely used for analyzing the spatial data. There is a problem with spatial data processing of extracting a potentially lower number of unobserved variables while a set of correlated variables is observed. The factor analysis is commonly used to overcome this problem. A bandwidth selection is a main difficulty during the identification a spatial heterogeneity of factor loadings. In the paper an original bandwidth selection criterion is proposed. It is based on the testing the difference between factor loadings of global and geographically weighted model. Using the simulated data it is shown that the criterion proposed allows to define accurately the appropriate number of nearest neighbors. The proposed approach is used to analyze real data on performance metrics of Russian universities.

*Keywords:* spatial data; geographically weighted factor analysis; bandwidth selection; factor loadings; nearest neighbors

---

## 1. Introduction

Large amounts of spatial data that need to be processed is stored in modern geographic information systems. Thus, the issue of reducing the dimension of the attribute space occurs frequently. The most popular approaches in this field are the method of the principal components analysis (PCA) and exploratory factor analysis (EFA).

Both the PCA and EFA, require homogeneity of observed data. This assumption is violated in cases where the observations depend on geographical factors. Thus, both similar and different degree of dependency between observed variables and latent factors in different geographic regions can be observed. Often some of its spatial properties are ignored in analyzing data process and standard methods for reducing dimension are used. However, such spatial effects are often important for a better understanding of investigated process, and PCA in this case may be replaced by geographically weighted PCA [1], when we want to explain a certain spatial heterogeneity in the data.

At the same time, the idea of building a geographically weighted model does not transferred so easy from the method of principal component to factor analysis. They have a number of substantial differences, and in fact, the purpose their use is different, in particular, factor loadings play the important role in the interpretation of the EFA results, reflecting the impact on the observed variables. For example, if there is a system of parameters, which are exposed to the same latent factor, the loadings on the main factor show the degree of impact on the indicators.

In this paper, we used the idea of building a local model of EFA for geographically nearest neighbors (adaptive bandwidth). However, there is the problem of choosing the number of nearest neighbors. Authors of paper [2] that is devoted to geographically weighted PCA, propose criterion ‘goodness of fit’, based on the minimization of the residual sum of squares. This makes sense, since the aim of principal component analysis is to present indicators in the space of smaller dimension with the least loss of information. Here, for the factor analysis, we suggest another criterion, which takes into account the significance of the differences of factor loadings of locally weighted and global models. This is more consistent with the aim of factor analysis.

Further, the principal component analysis and factor analysis are described in more detail and explained the difference between them. Essence of geographical weighting is presented, the problem of bandwidth selection is described. A new criterion for selecting the number of nearest neighbors is proposed. Advantages of this approach were demonstrated by comparison with the existing criterion of goodness of fit in the simulation study.

## 2. Geographically weighted variable reduction

PCA and EFA are both variable reduction techniques and sometimes erroneously considered as the same statistical method. However, there are distinct differences between PCA and EFA. Further, mathematical description of these approaches is given, and we explain how they are adapted to spatial data analysis.

### 2.1. Principal component analysis

There are  $n$  observations of  $m$  variables, so a data matrix  $X$  contains  $n$  rows and  $m$  columns. The columns in  $X$  are normalized with zero mean and unit variance. Then  $R = X^T X$  is the correlation matrix for  $X$ . The matrix  $X^T$  denotes the transpose of  $X$ . The matrix  $R$  is a real symmetric matrix and its factorization into a canonical form is

$$R = \Lambda \Phi \Lambda^T \quad (1)$$



where an orthogonal matrix  $A$  contains the eigenvectors of  $R$ , and  $\Phi$  is a diagonal matrix which entries are the eigenvalues of  $R$ . The eigenvalues of diagonal  $\Phi$  imply the variances of the corresponding principal components. The eigenvectors in  $A$  are column vectors and represent the loadings of each variable on the corresponding principal component.

If the number of principal components equal to the number of variables, the decomposition (1) perfectly reproduces the correlation matrix  $R$ . By reduction  $m$  variables in  $q$ -dimensional sub-space ( $q < m$ ) the correlation matrix is represented as

$$\hat{R}_q = A_q \Phi_q A_q^T$$

where  $A_q$  denotes the matrix  $A$  with the first  $q$  columns, i.e. the loadings on the first  $q$  principal component, and  $\Phi_q$  is a diagonal matrix which entries are the first  $q$  eigenvalues of  $R$ . The principal components are sorted in decreasing order of eigenvalues so the first principal components keep the most important information from the data set.

Component scores in  $q$ -dimensional sub-space are found by multiplying the original data matrix  $X$  by loading matrix  $A_q$ . The best rank  $q$  approximation to  $X$  is  $\hat{X}_q = X A_q A_q^T$ . A standard result in linear algebra states that

$$A A^T - A_q A_q^T = A_{(-q)} A_{(-q)}^T$$

where  $A_{(-q)}$  denotes the matrix  $A$  with the first  $q$  columns removed.

To assess the quality of the reconstitution of  $X$  with  $q$  components, the dissimilarity between  $X$  and  $\hat{X}_q$  is usually evaluated. The error matrix is

$$E = X - \hat{X}_q = X A A^T - X A_q A_q^T = X A_{(-q)} A_{(-q)}^T.$$

The most popular coefficient used for evaluating the quality of PCA model is the residual sum of squares [3]

$$RESS_q = \|X A_{(-q)} A_{(-q)}^T\| \quad (2)$$

where  $\|M\|$  is the square root of the sum of all the squared elements of the matrix  $M$ .

Thus, mathematically, PCA depends on the eigen-decomposition of positive semidefinite matrices. Its main goal is to extract the important information from the data using the correlation between the variables and to represent it as a set of orthogonal principal components in the sub-space of lower dimension.

## 2.2. Factor analysis

EFA model assumes that the relationship between the measured variables is due to the effect of some unobservable (latent) factors. The input information is a correlation matrix  $R$  for all variables. It can be represented as [4]

$$R = A \Phi A^T + \Psi \quad (3)$$

where  $A$  is a factor loading matrix reflecting the relationship between the variables and factors,  $\Phi$  is a correlation matrix of  $q$  factors,  $\Psi$  is a covariance matrix of the unique factors.

The presence of unique factors in EFA model (3) is the main difference from the model of PCA (1). It is due to the fact that extracted latent factors do not fully (with some errors) describe the correlation between the observed variables. Uniqueness is the variance that is 'unique' to the variable and not shared with other variables. The independence of unique factors is assumed, so the matrix  $\Psi$  is a diagonal with uniqueness on the diagonal.

The matrices  $A$ ,  $\Phi$ ,  $\Psi$  are estimated. In contrast to the PCA model the matrix  $A$  is of a particular interest, but loadings are not uniquely determined, so the rotation procedure is used, so that the resulting factor structure has a meaningful interpretation. With orthogonal rotation the independence between the latent factors is assumed. So matrix  $\Phi$  is given as identity. There are a number of factor extraction methods for estimating loadings and uniqueness, for example, principal factor solution, minimum residual, maximum-likelihood method.

Minimum residual method is based on ordinary least squares (OLS). The loss function is

$$F_{OLS}(A, \Psi) = \text{tr} \left( R - (A \Phi A^T + \Psi) \right)^2. \quad (4)$$

Here  $\text{tr}(M)$  is the trace of a square matrix  $M$ . The OLS-estimates  $\hat{A}$ ,  $\hat{\Psi}$  are arguments at which the minimum of loss function (4) is achieved.

## 2.3. Geographically weighted models

The usage of local weighting as part of regression estimation initially was proposed by [5]. This approach is widely used in spatial data analysis [6] and known as geographically weighted model.

Geographically weighted model identifies spatial differences in the relationship between factors by constructing a regression model at each control point for geographically closed observations. The proximity regulated by assigning larger weights to closest points and reducing weights for observations as they move away from the control point. Thus, the weight is determined as a function of distance from the control point to the objects. The regression is estimated over the local subregion which volume is determined by the weight function parameter (a bandwidth).

Regardless of the form of weight function specified, the local correlation matrix is

$$R^{(i)} = A^{(i)} \Phi^{(i)} A^{(i)T} \quad (5)$$

with respect to the local subregion of the  $i$ -th control point. The scores for the  $i$ -th control point on the  $m$  variables are  $\mathbf{x}^{(i)} A^{(i)}$  where  $\mathbf{x}^{(i)}$  is a vector of variable values at  $i$ -th control point.

Similarly geographically weighted EFA model is defined as

$$R^{(i)} = A^{(i)} \Phi^{(i)} A^{(i)T} + \Psi^{(i)} \quad (6)$$

and values of matrices  $A^{(i)}$ ,  $\Psi^{(i)}$  are estimated with respect to the local subregion of the  $i$ -th control point.

### 3. Criteria of bandwidth selection

The choice of bandwidth value has a decisive influence on the estimation quality [6]. If someone takes bandwidth too large, then almost all observations will be included in the model, so it will be coincidental to the global model without geographical weighting. Thus it will not be possible to describe the change of the explanatory factors impact depending on the spatial location of the objects. Otherwise, too small bandwidth leads to the overfitting problem: the model perfectly predicts the training data, but drastically fails on some new datasets.

The choice of a bandwidth parameter cannot be based on the common fitting indicators (like R-squared, the mean square error, etc.). So for optimal bandwidth selection the cross-validation technique is often used when the training and quality evaluation both are produced on the distinct sample data [7]. There are some other approaches to solve this problem, for instance, Akaike information criterion [7], and the Lagrange multiplier test [8].

In recent studies on the bandwidth selection [8, 9] for geographically weighted models two essentially different methods to the weighting function construction are considered:

- with a fixed local area radius;
- with a given number of nearest neighbors.

The second one is considered to be adaptive because it allows adjusting to varying density of the spatial location of the objects. Thus in the neighborhood of one control point the objects may be more concentrated than for the other points where they are more distant from each other. In such cases the latching of the local area radius leads to the fact that for some control points the regression will be estimated on a very large number of observations, for the others – on very small.

We selected the adaptive approach. Therefore, the task is to determine the optimal number of nearest neighbors, taking into account features of EFA.

#### 3.1. Goodness of fit statistic

The criterion ‘goodness of fit’ is proposed for geographically weighted PCA model construction (see [2]). The criterion is based on the minimization of the residual sum of squares. It is calculated by the formula (2) for a global model. For a local PCA model (5) the residual sum of squares at the  $i$ -th control point is

$$RESS_q^{(i)} = \left\| X^{(i)} A_{(-q)}^{(i)} A_{(-q)}^{(i)T} \right\|$$

where a superscript  $(i)$  denotes the values that are calculated in a local subregion of the  $i$ -th control point. The values of the residual sum of squares are summed for all the control points to calculate the goodness of fit statistics:

$$GOF = \sum_{i=1}^n RESS_q^{(i)} .$$

A set of control points can be selected in different ways. Leave-One-Out Cross-Validation (LOOCV) is the simplest procedure for cross-validation. It suggests that only one observation is selected as a control point from the data set, while other observations are considered as a training sample. The procedure is repeated until all the objects will be alternately selected as control points. The advantage of this approach is a computational speed. Often the model structure based on LOOCV leads to the overfitting problem and the forecast error underestimation [10]. In our case the wrong selection of bandwidth parameter may cause such problems.

The more complicated procedure is a Monte-Carlo cross-validation (MCCV) [10]. It assumes that the whole sample is separated randomly into training and check samples. Nevertheless, this choice could increase the computational time. Therefore, we have chosen LOOCV procedure.

### 3.2. Test the difference between global and local factor loadings

There are some problems with usage the cross-validation technique for evaluating the quality of EFA models. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. The task of variable reduction is not quite a prediction. Of course, we can use the loss function  $RESS_q$  for PCA, and  $F_{OLS}(\Lambda, \Psi)$  for EFA. But goodness of fit is not so important for EFA, the loadings are more interesting. For this reason a new criterion for bandwidth selection is proposed. It is based on testing the difference between global and local factor loadings.

A statistical inference for comparing global and local factor loadings is based on the information about mean values of loadings and their standard deviation. We need replications of sample data to get this information. One way to get it is to take a sample of the same size  $n$  from the rows of data matrix  $X$  with replacement. Let we have  $L$  replications of sample data. For each  $l$ th replication we estimate loading matrix  $A_l$  of global EFA model (3) and loading matrix  $A_l^{(i)}$  of the geographically weighted EFA model (6). We can calculate matrices containing the average values of loadings for all replications

$$\bar{A}_l = \frac{1}{L} \sum_{l=1}^L A_l, \bar{A}_l^{(i)} = \frac{1}{L} \sum_{l=1}^L A_l^{(i)}.$$

The  $k, j$  th elements of matrices  $\bar{A}_l$  and  $\bar{A}_l^{(i)}$  are denoted by  $m(\lambda_{kj})$  and  $m(\lambda_{kj}^{(i)})$ .

Similarly, we can calculate the variance of loadings for all replications. Let us denote them as  $v(\lambda_{kj})$  and  $v(\lambda_{kj}^{(i)})$ . So the test statistic comparing the means is well known. It is given by

$$SS_{kj}^{(i)} = \frac{m(\lambda_{kj}^{(i)}) - m(\lambda_{kj})}{\sqrt{1/L} \sqrt{v(\lambda_{kj}^{(i)}) + v(\lambda_{kj})}}. \quad (7)$$

We propose the significance test statistics for optimal bandwidth selection

$$SS = \frac{1}{n \cdot q \cdot m} \sum_{i=1}^n \sum_{j=1}^q \sum_{k=1}^m |SS_{kj}^{(i)}|. \quad (8)$$

The maximum value of the significance test statistics (8) corresponds to the optimal number of nearest neighbors. On the one hand it is evident that for the global model (maximum number of nearest neighbors) the numerator of (7) will be minimal, and vice versa. So we would expect that the geographically weighted EFA model with the smallest number of nearest neighbors will be the best by test statistics (7). But on the other hand a small number of nearest neighbors results in very large loadings variation. Thus, the denominator of (7) will increase with a decrease of the number of nearest neighbors. Essentially the significance test statistics (8) is a trade-off between differences in the average factor loadings and their variation.

To calculate statistics (8) we need to compare multiple EFA models. These comparisons require columns of factor loading matrices to be properly aligned. However, the most rotation criteria do not uniquely define a factor loading matrix. This is referred to an alignment problem [4]. The most popular method for aligning a factor loading matrix against another is to minimize the sum of squared differences of factor loadings in the two matrices. Further, in simulation study, this approach was used. We compared the dissimilarity of loading column of global EFA model and one of local models, initial and with the opposite sign. Column reflection (an operation when the signs of values in column are replaced with the opposite) of original column was carried out in cases when reflected loading column corresponded to less value of sum of squared deviations.

There are some problems with the calculation of the statistics (8). Firstly, it is necessary to conduct the  $L \cdot n$ -fold factor analysis. With a large number of control points and repeated replications, this procedure takes a very long time. A smaller number of control points can be taken to reduce the calculation time. Another way is to replicate the sample using the jackknife method. However, the decrease in the number of replications may result in loss of the quality of an optimal bandwidth selection.

Secondly, factors can be extracted using various methods. There is a problem with the starting values during the optimization of the log likelihood. The uniqueness is technically constrained to lie in  $[0, 1]$ , but there are some problems with near-zero values, and the optimization is typically done with a lower bound of 0.005. Sometimes it is unable to optimize the likelihood from certain starting value, because the algorithm does not converge. If we try to increase or decrease the lower bound for uniqueness during the optimization, it allows a solution to be converged. However, such lower bound selection is practically not efficient in the case of  $L \cdot n$ -fold factor analysis. So more simple factoring methods should be used.

The method of principal axes may be used in the cases when maximum likelihood solutions fail to converge. However, it is based on the iterative algorithm, so it works rather slowly. If the procedure of factor analysis is repeated many times, the speed

of implementation of the factoring procedure is very important. Therefore, optimization procedures are more preferable. In addition, they produce even better solutions for some examples. Minimum residual method based on OLS usually has no problems with convergence and tends to produce better solutions.

Further, two approaches to the bandwidth selection are compared in a simulation study. For identification of PCA and EFA models one can use statistical packages, for instance, the free software for statistical analysis R [11]. The function `princomp` {stats} performs a principal components analysis on the given numeric data matrix, function `efa`{EFAutilities} performs exploratory factor analysis. The algorithms of optimal bandwidth selection are implemented using R.

#### 4. Results of simulation study

The main purpose of the simulation study is concluded in comparison of approaches mentioned above in precision of the bandwidth selection. A simple one-factor model was chosen. The factor  $F$  is standard normally distributed. It affects three variables  $x_1, x_2, x_3$ . So the EFA model has the form

$$\begin{cases} x_1 = b_1 F + \varepsilon_1, \\ x_2 = b_2 F + \varepsilon_2, \\ x_3 = b_3 F + \varepsilon_3 \end{cases} \quad (9)$$

where  $b_1, b_2, b_3$  are factor loadings,  $\varepsilon_1, \varepsilon_2, \varepsilon_3$  are random errors. The simulated random error  $\varepsilon_i$  was chosen as a normally distributed variable with variance that equals to  $0.8^2 - b_i^2$ .

The case with certain local centers of object's concentration was considered for modeling spatial heterogeneity. The whole number of those centers was equal to six, and each of them was represented by a circle with the same number of observations. All center's locations were chosen randomly within the unit square  $[0,1]^2$ . The sample size was  $n = 300$ . The true value of number of nearest neighbors was 50. Different spatial location of objects towards the centers was set in two ways.

In Model 1, the radius of a circle with homogeneous observations was fixed. It was set to 0.05.

Model 2 has a distance from the objects to the center of the local area multiplied by a coefficient  $1 + 0.2K$ , where  $K$  is a serial number of the area,  $K = 1, \dots, 6$ . The number of objects belonging to the  $K$ -th region was set to  $29 + 6K$ . This ensures that there are areas with varying density of spatial location of objects.

Equal loadings were set for all objects in the same local area. Their values are shown in Table 1.

Table 1. The true values of factor loadings.

Factor loadings	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
$b_1$	0.37	0.32	0.25	0.57	0.58	0.71
$b_2$	0.43	0.18	0.49	0.16	0.32	0.28
$b_3$	0.2	0.5	0.26	0.27	0.1	0.01

The total number of experiments was equal to 50 for each model. The values of both statistics  $GOF$  and  $SS$  were calculated on the simulated data for fixed number of nearest neighbors. The number of nearest neighbors was set from 20 to 150 with the step of 10. The optimal number of nearest neighbors based on  $GOF$  statistics was selected as argument of the goodness of fit statistics minimum constructed as a result of the LOOCV procedure. The optimal number of nearest neighbors based on  $SS$  statistics was selected as argument of the significance test statistics maximum constructed as a result of the MCCV procedure. The number  $L$  of replications of sample data was set to 100. The size of random sample with replacement was  $n = 300$ . The final results are presented in Table 2.

Table 2. The optimal number of nearest neighbors.

	Based on $GOF$ statistics				Based on $SS$ statistics			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
Model 1	42.5	75	84.2	137.5	30	40	50.6	67.5
Model 2	52.5	70	83.8	120	22.5	40	49.2	60

It is clearly seen that the significance test statistics proposed determines the number of nearest neighbors more accurately for both model 1 and model 2. The goodness of fit criterion gives an average value of optimal number of nearest neighbors of about 80, while the significance test statistics reaches a maximum value when the average number of nearest neighbors almost equal to

50. Thus, the use of the goodness of fit criterion leads to an explicit overestimation of the bandwidth. In addition, SS statistics also provide a more accurate determination of the number of nearest neighbors. The interquartile range of the optimal bandwidth for it is 30.5. While the optimal number of nearest neighbors, according to the GOF criterion, has an interquartile range of 95 and 67.5, that is 2-3 times more than the results of use the second criterion.

Figure 1 shows the resulting values of statistics for one of the samples. The true number of nearest neighbors for model 1 that equals 50 is indicated by a dashed line. For model 2 the true number of nearest neighbors varies according to local area from 35 to 65 with the step of 6. The average number of nearest neighbors is 50.

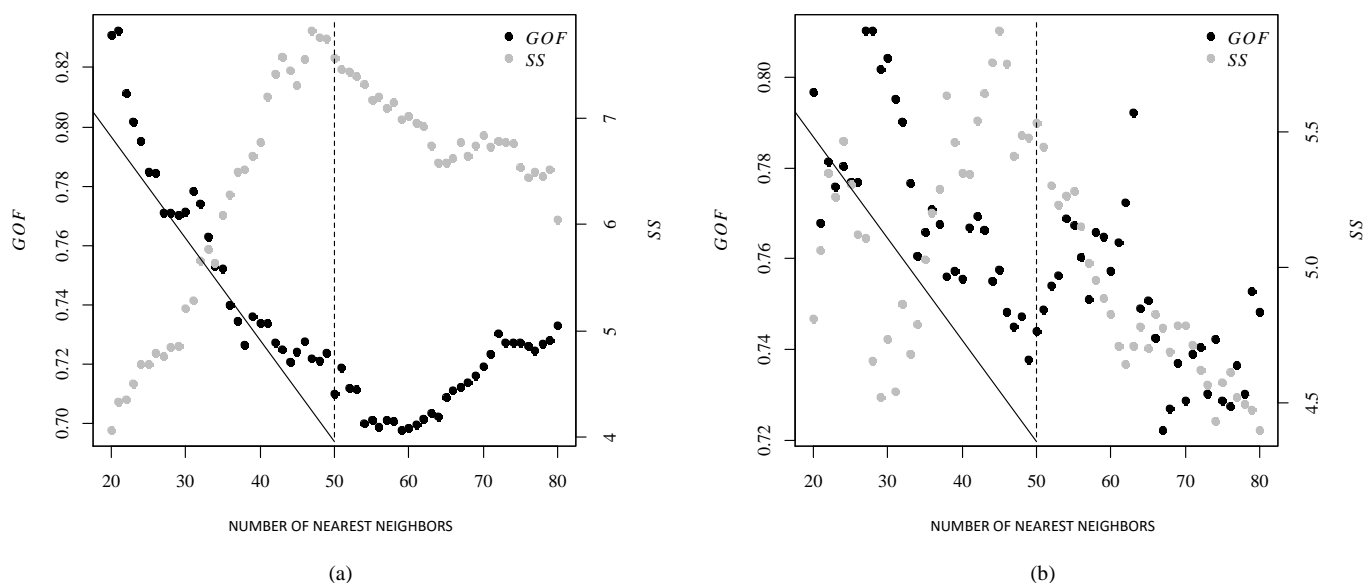


Fig. 1. The dependence on the goodness of fit statistics and the significance test statistics on the number of nearest neighbors for model 1 (a) and model 2 (b).

It should be noted that values of the goodness of fit statistics vary greatly. We see a lot of local minima and maxima. This fact complicates the use of optimization routines to find the best values of the bandwidth. At the same time, the dependence of the significance test statistics on the number of nearest neighbors appears smoother. This fact allows us to develop more effective optimization algorithms than direct-search method on the grid.

## 5. Application to educational monitoring

The Ministry of Education and Science of the Russian Federation initiated monitoring of the effectiveness of universities in 2012. Since then, all Russian universities are obliged to provide information on their activities on a number of indicators. The decision on the effectiveness of the university is made depending on whether the university is able to reach thresholds for most indicators. The leadership of different universities can differently determine the priority indicators. It is interesting to determine the structure of universities' efficiency taking into account regional differences in their activities.

We will rely on the model (9) for determining the structure of performance indicators. We are interested in the factor  $F$  that is the overall efficiency of the university. So the observed indicators of activities can be interpreted as

- $x_1$  is a financial and economic activity: income of the educational organization from all sources per one NDP;
- $x_2$  is a level of wages of the teaching staff: the ratio of the salary of PPP to the average wage for the economy of the region;
- $x_3$  is an employment of graduates: the proportion of graduates who have found employment during the calendar year following the year of release, in the total number of graduates of the educational organization who have studied the main educational programs of higher education.

Coefficients  $b_1, b_2, b_3$  show the extent to which a particular performance indicator determines effectiveness.

The data from monitoring the effectiveness of educational institutions of higher education for 2015, downloaded from the pages of each individual university, were used as an information base [12]. 571 universities are represented in the sample. They provided information on performance indicators, branches of universities are not included in the analysis. The optimal number of nearest neighbors was chosen based on the SS statistics. The MCCV procedure was used. Control points were located in administrative centers within six federal districts: Central Federal District (CFD), Northwestern Federal District (NWFD), Volga Federal District (VFD), Ural federal district (UFD), Siberian Federal District (SFD), Far Eastern Federal District (FEFD). A total of 67 control points are set. The number  $L$  of replications of sample data was set to 300. The size of random sample with replacement was 571. The optimal number of nearest neighbors was 119. The average values of loadings for all replications denoted by  $m(b_1)$ ,  $m(b_2)$ ,  $m(b_3)$  are presented graphically in Figure 2 using pie charts.



Fig. 2. Geographical variations of the factor loadings of universities' efficiency.

As can be seen from Fig. 2, the structure of indicators of the effectiveness of HEIs varies greatly depending on the territory. Thus, for the regions of Siberia and the Far East, the income of the educational organization has the greatest weight. For many European regions of the country, financial and economic activity does not have such a significant contribution. In most cases, the level of wages of scientific and pedagogical workers is most important in the formation of performance of universities. Only for one northern region of Russia (Arkhangelsk region), employment of graduates predominates in the structure of performance indicators. Consequently we can conclude that for most universities the key performance indicators are financial, while the employment of graduates as a result of educational activities does not play such a significant role.

## 6. Conclusion

In the paper we propose an original criterion to determine the bandwidth for geographically weighted factor analysis estimation. For this purpose the authors developed a software implementation using the statistical framework R. The investigation of the accuracy of the criterion that determines the optimal number of neighbors is based on the results of the experiments. They show that proposed significance test statistics determines the optimal number of nearest neighbors more accurately. Furthermore the dependence of significance test statistics on the number of nearest neighbors appears smoother. This makes it possible to develop more effective optimization algorithms for automatic bandwidth selection. The proposed approach is used for education monitoring problems.

## Acknowledgements

This research has been supported by the Ministry of Education and Science of the Russian Federation as part of the state task (project No 2.7996.2017/BCh).

## References

- [1] Lloyd CD. Analysing population characteristics using geographically weighted principal components analysis: a case study of Northern Ireland in 2001. *Computers Environment and Urban Systems* 2010; 34: 389–399.
- [2] Harris P, Brunsdon C, Charlton M. Geographically Weighted Principal Components Analysis. *International Journal of Geographical Information Science* 2011; 25(10): 1717–36.
- [3] Abdi H, Williams LJ. Principal component analysis // *Wiley interdisciplinary reviews: computational statistics* 2010; 2(4): 433–459.
- [4] Zhang G. Estimating standard errors in exploratory factor analysis. *Multivariate behavioral research* 2014; 49(4): 339–353.
- [5] Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots // *Journal of the American statistical association* 1979; 74(368): 829–836.
- [6] Brunsdon C, Fotheringham AS, Charlton ME. Geographically Weighted Regression: a Method for Exploring Spatial Nonstationarity. *Geographical Analysis* 1996; 28(4): 281–298.
- [7] Farber S, Páez A. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems* 2007; 9(4): 371–396.
- [8] Cho SH, Lambert DM, Chen Z. Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. *Applied Economics Letters* 2010; 17(8): 767–772.
- [9] Guo L, Ma Z, Zhang L. Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Canadian Journal of Forest Research* 2008; 38(9): 2526–2534.
- [10] Xu QS, Liang YZ. Monte Carlo Cross Validation. *Chemometrics and Intelligent Laboratory Systems* 2001; 56(1): 1–11.
- [11] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL: <http://www.R-project.org/> (15.05.2017).
- [12] Information and analytical materials on the results of monitoring the effectiveness of educational institutions of higher education. URL: <http://indicators.miccedu.ru/monitoring/2015/> (15.05.2017).

# A learning based feature point detector

A. Verichev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We propose a learning-based image feature points detector. Instead of giving an explicit definition for feature point we apply the methods of machine learning to infer it inductively using a representative training set. This allows for a flexible tuning of the proposed detector to a specific problem that is described by a training set of desired responses. To increase feature points' repeatability and robustness to various image transformations the feature space of the learning algorithm includes raw image moments and image moment invariants. Experiments demonstrate high flexibility in tuning the detector to a specific task, acceptable repeatability of the feature points and robustness to various image transformations.

*Keywords:* image feature points; image feature points detector; image moments; image moment invariants; machine learning

---

## 1. Introduction

Feature point is a piece of information which is relevant to solving a certain application-related computational task. Feature points find their use in numerous applications such as image stitching, stereo correspondence, locating and tracking of a moving object, object detection and recognition, and others [1, 2]. The ubiquitous usage of feature points is a direct consequence of their properties [3]:

- *Repeatability:* Given two images of the same object or scene, a high percentage of the features detected on the scene visible in both images should be found in both images.
- *Informativeness:* The intensity patterns underlying the detected features should show a lot of variation.
- *Locality:* The features should be local, so as to reduce the probability of occlusion and to allow simple model approximations of the geometric and photometric deformations between two images.
- *Quantity:* The number of detected features should be sufficiently large, such that a reasonable number of features are detected even on small objects.
- *Accuracy:* The detected features should be accurately localized.
- *Efficiency:* The detection of features in a new image should allow for time-critical applications.

Algorithms and methods that detect image feature points by making local decisions are called feature points detector. An abundance of image feature points detectors is known, most of which are based on a certain criterion - a heuristics that implicitly defines what a term feature point constitutes. Generally these heuristics can be classified into three categories [4]:

- *Gradient-based:* A majority of image feature points detectors is based on computation of gradients of intensity function, for example Förstner [5], Harris [6], Shi-Tomasi [7].
- *Template-based:* Feature points are found by comparing the intensity of surrounding pixels with that of center pixels which is governed by some template. The well-known template-based detectors are SUSAN [8], FAST [9], AGAST [10].
- *Contour-based:* A feature point is defined as the intersecting point of two adjacent edge lines, examples are DoG-curve [11], ANDD [12].

However, formulating a heuristics for an image feature points detector requires a well-formed application-dependent definition of the term feature point, which in turn requires some level of expertise in the application domain. Moreover, a strictly stated criterion, although sharpening performance, diminishes its flexibility to adjust to a particular problem, which renders all the possible usages outside the destined application moot.

The goal of this work is to dispense with defining the term feature point altogether and focus on the properties we wish the feature points to possess. With that goal in mind we resort to machine learning methods. Image raw moments and image moment invariants are used along with some other local characteristics of image points to form a feature space of a learning algorithm. The detector is trained to solve a specific problem on a relevant and carefully collected training set. This effectively defines the term feature point implicitly, since it's inductively inferred from the training examples.

The proposed method is described in full detail in section 2, along with the learning algorithm, its feature space and the procedures for collecting training and test sets. Evaluation criteria of a trained detector's performance and the results of experimental evaluation are described in section 3. We conclude with a discussion of these results.

## 2. Proposed method

The proposed learning-based feature points detector is based on the idea of transforming detection task into a classification task as suggested in [13], which boils down to training the detector's classifier on a set of the desired responses.

### 2.1. Feature space

The first step towards constructing our detector is to define the classifier's feature space, which is an  $\mathbb{R}^{15}$  vector space. Each pixel of an image  $I[x, y]$  is mapped to a certain vector in this feature space using a locally defined operator  $P^{9 \times 9} \rightarrow \mathbb{R}^{15}$ , where  $P = \{n: 0 \leq n < 256\}$  is a set of intensities of a grayscale image. The features of the feature space are described below.

The first two features are standard deviation of a standardized local area,  $\phi_1$ , and standard deviation divided by the norm of the local area,  $\phi_2$ :

$$\phi_1 = \sqrt{\frac{1}{80} \sum_{i=-4}^4 \sum_{j=-4}^4 \frac{1}{n^2} (I[x+i, y+j] - \bar{I})^2}, \quad (1)$$

$$\phi_2 = \frac{\phi_1}{n},$$

where norm  $n$  and local mean  $\bar{I}$  are defined:

$$\bar{I} = \frac{1}{81} \sum_{i=-4}^4 \sum_{j=-4}^4 I[x+i, y+j],$$

$$n = \sqrt{\sum_{i=-4}^4 \sum_{j=-4}^4 (I[x+i, y+j])^2}.$$

The use of these features is motivated by their sensitivity to monotonous and textured areas.

The next four features are chosen to be central image moments of a local image area:  $\phi_{t+3} = \mu_{tt}$ ,  $0 \leq t \leq 3$ . The central moments are defined [14]:

$$\mu_{ij} = \sum_{k=-4}^4 \sum_{l=-4}^4 k^i \cdot l^j \cdot \frac{1}{81} I[x+k, y+l]. \quad (2)$$

To induce invariance to rotation transformations the following Hu invariant image moments and Flusser moments are used [15, 16]:

$$\begin{aligned} \phi_7 &= \mu_{20} + \mu_{02}, \\ \phi_8 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2, \\ \phi_9 &= (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2, \\ \phi_{10} &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2, \\ \phi_{11} &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2], \\ \phi_{12} &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}), \\ \phi_{13} &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] - (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2], \\ \phi_{14} &= \mu_{11}[(\mu_{30} + \mu_{12})^2 - (\mu_{03} + \mu_{21})^2] - (\mu_{20} - \mu_{02})(\mu_{30} + \mu_{12})(\mu_{03} + \mu_{21}). \end{aligned} \quad (3)$$

Moments calculation is an intensive computational task that requires of a lot of operations. To reduce the number of arithmetical operations we apply the recursive method of moments calculation based on the use of integer factorial polynomials [16].

The last feature that characterizes misalignment of centre of local area and its centre of mass is defined:

$$\phi_{15} = \sqrt{(x_c - x)^2 + (y_c - y)^2}, \quad (4)$$

where  $x_c = \mu_{10}/\mu_{00}$  and  $y_c = \mu_{01}/\mu_{00}$ .

The set of the features  $\phi_i$ ,  $1 \leq i \leq 15$ , defined by (1) - (4), with a usual addition and scalar multiplication operations form the feature vector space.



## 2.2. Tuning the detector

### 2.2.1. Collecting a training set

Tuning the detector requires a training set that consists of the desired detector's responses. Depending on the application there are various ways the set can be obtained:

- manually, involving experts of the domain;
- automatically, using well-known feature points detectors such as Harris or Canny;
- combining the two.

In case there is a human involvement of any kind it is inevitable for a training set to contain a so called training noise [17]. Besides, in a typical scenario a number of feature points is small compared to the other points. To alleviate these negative effects the neighbouring points of the feature points can be considered feature points as well.

Provided an application requires high level of robustness to certain transformations, a training set can be enlarged to contain the so called virtual examples [18]. To this end every image used to form a training set is transformed according to some transformation. Since the parameters of that transformation are known, the elements of the original image can be mapped onto the transformed image, which makes it possible to extract feature vectors of the points of the transformed image that correspond to the feature points of the original image. These new feature vectors are the virtual examples that convey information about various effects the transformation have on the feature vectors.

### 2.2.2. Training a classifier

With a training set at hand we can pose and solve a supervised learning problem. Since the number of the feature vectors in a training set is typically quite large we chose to apply nonparametric probability density estimation approach. Let  $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$  denote training set, where  $\mathbf{x}_i$  is a feature vector,  $y_i$  is its label,  $y_i \in \{C_1, C_2\}$ .  $C_1$  corresponds to feature points and  $C_2$  corresponds to the other points. Then, an estimation of conditional probability density function is defined as follows:

$$\hat{p}(\mathbf{x}|C_i) \propto \sum_{j=1}^N [y_j = C_i] K\left(\frac{\|\mathbf{x}-\mathbf{x}_j\|}{h}\right), \quad (5)$$

where  $K$  is a kernel function,  $h$  is kernel's width parameter. By the Bayes' Theorem:

$$\hat{p}(C_i|\mathbf{x}) \propto \hat{p}(\mathbf{x}|C_i) \cdot \hat{\pi}_i, \quad (6)$$

where  $\hat{\pi}_i$  is an estimate of prior probability of  $i^{\text{th}}$  class:

$$\hat{\pi}_i = \frac{1}{N} \sum_{j=1}^N [y_j = C_i]. \quad (7)$$

Define a *characteristic function* of a feature point  $l(\mathbf{x})$ :

$$l(\mathbf{x}) = \ln(\hat{p}(C_1|\mathbf{x})) - \ln(\hat{p}(C_2|\mathbf{x})). \quad (8)$$

In order to smooth the detector's response we filter the characteristic function  $l(\mathbf{x})$  using a local peak filter. The peak filter suppresses non-maximal values in a local  $3 \times 3$  neighbourhood of the point  $\mathbf{x}$ :

$$\tilde{l}(\mathbf{x}) = \begin{cases} l(\mathbf{x}), & l(\mathbf{x}) > l(\mathbf{g}) + \delta \quad \forall \mathbf{g} \in W \setminus \{\mathbf{x}\} \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

where  $W$  is a set of all feature vectors from the local neighbourhood,  $\delta$  is some threshold.

From (8) and (9) we infer the decision rule:

$$y(\mathbf{x}) = \begin{cases} C_1, & \tilde{l}(\mathbf{x}) > t = \ln\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right) \\ C_2, & \text{otherwise} \end{cases} \quad (10)$$

## 3. Experimental evaluation

### 3.1. Experimental setup

To experimentally evaluate the proposed detector we built a set of images. The set contains a series of 10 overlapping images of 6 different scenes, 60 images in total. Figure 1 shows three images of one of these scenes. Each of the 6 groups of images was split in relation 8:2 to form training set  $D$  and test set  $C$ , respectively. We chose to use Harris [6] corner detector to detect feature points. The training set was enlarged by the virtual examples as described in section 2.2.1 and the transformations that were applied are described in section 3.3.



Fig. 1. Example images of a scene.

### 3.2. Evaluation of training accuracy

Let  $V = \{(x_i, y_i)\}_{i=1}^N$  be training or test set. The primary criterion of detector's performance on the set  $V$  is its *accuracy*:

$$A(V) = \frac{1}{N} \sum_{i=1}^N [y(x_i) = y_i]. \quad (11)$$

Besides the accuracy two more criteria are used: *precision*  $P$  and *recall*  $R$  [19]. Precision is the fraction of relevant instances over the retrieved instances, while recall is the fraction of relevant instances among the retrieved ones over the total number of relevant instances in the set.

Let  $FP$ ,  $FN$  and  $TP$  denote false positives, false negatives and true positives, respectively. Then,

$$\begin{aligned} FP(V) &= \sum_{i=1}^N [y(x_i) = C_1] \cdot [y_i = C_2], \\ FN(V) &= \sum_{i=1}^N [y(x_i) = C_2] \cdot [y_i = C_1], \\ TP(V) &= \sum_{i=1}^N [y(x_i) = y_i]. \end{aligned} \quad (12)$$

Precision and recall are defined:

$$\begin{aligned} P(V) &= \frac{TP(V)}{TP(V) + FP(V)}, \\ R(V) &= \frac{TP(V)}{TP(V) + FN(V)}. \end{aligned} \quad (13)$$

The proposed detector was first trained on the training set. Accuracy, precision and recall were evaluated on the training set  $D$  and test set  $C$ . The results are shown in table 1. Taking into account a fairly large size of the sets, the data suggests an adequate quality of training.

### 3.3. Repeatability evaluation of the detector

As mentioned in introduction, repeatability is one of the most important properties of the feature points. Along with its importance, repeatability allows for an objective and qualitative evaluation. Hence, we used repeatability to evaluate the performance of the proposed detector.

Table 1. Accuracy, precision and recall of the trained detector .

	$A(D)$	$P(D)$	$R(D)$
Training set, $D$	0.997	0.905	0.960
Test set, $C$	0.9766	0.730	0.580

The procedure for repeatability evaluation is outlined below.

- An original image is used to find a set of feature points  $P_o$ .
- The original image is transformed by one of the transformations (cf. the next list below).
- The transformed image is used to find a set of feature points  $P_t$ .
- Since parameters of the transformation are known, coordinates of the points  $P_o$  of the original image can be mapped onto the transformed image. Thus, the points in the set  $P_o$  are mapped onto the transformed image, forming a set  $P_m$ .
- The sets  $P_m$  and  $P_t$  are matched. Two points  $a \in P_m$  and  $b \in P_t$  are considered equal if  $a \in V_\varepsilon(b)$ ,  $\varepsilon = 2.0$ .

- As a result of the comparison performed in the previous step we find three sets of points:  $P_{TP}$  are the points found on both sets,  $P_{FP}$  are new points that were not found on the original image but were found on the transformed image,  $P_{FN}$  are the missed points that were found on the original image and were not found on the transformed image. The cardinalities of these sets are, respectively,  $TP$ ,  $FP$  and  $FN$  values of the proposed detector. These values are used to calculate the detector's accuracy, precision and recall.

To evaluate repeatability we used the following transformations of the images:

- rotation by angle  $\alpha$ ,  $-45^\circ \leq \alpha \leq 45^\circ$ ,  $\alpha$  is increased by  $3^\circ$ ;
- sub-pixel shift by  $t$ ,  $0.25 \leq t \leq 0.75$ ,  $t$  is increased by 0.05;
- scaling by  $s$ ,  $0.5 \leq s < 1.5$ ,  $s$  is increased by 0.1

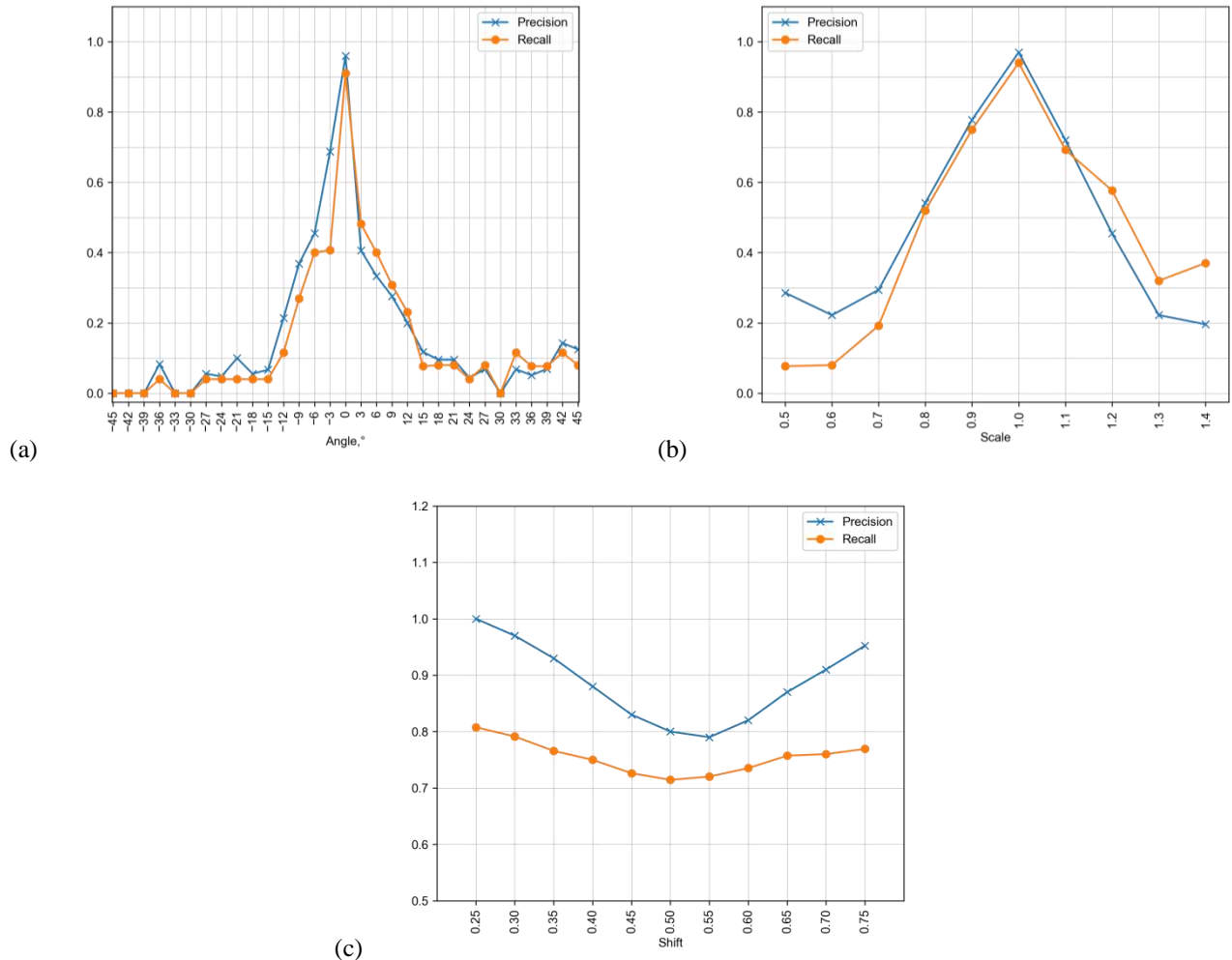


Fig. 2. Repeatability of the detector evaluated for various transformations: (a) rotation, (b) scaling, (c) translation.

The results of the repeatability evaluation of the proposed detector that was trained on the training set  $D$  are shown on fig. 2. The detector's performance can be considered adequate on rotated images for  $-9^\circ < \alpha < 9^\circ$  and on scaled images for  $0.8 \leq s \leq 1.2$ . The performance on shifted images is high for the whole range of the parameter  $t$ .

#### 4. Conclusion

In this paper we investigated a relatively new approach to feature point detection. Contrary to the standard approach to the problem, we didn't formulate any heuristics-based definition of the term feature point but tried to infer it inductively using the methods of machine learning and a representative training set. This enabled us to tune the proposed detector to a specific problem at hand. The results of the experimental evaluation of the detector verify that such a tuning is in fact possible. Moreover, the detector showed acceptable robustness to rotation and scaling transformation, and high robustness to sub-pixel shift transformation. This suggests a great potential of the learning-based approach to feature points detection.

#### Acknowledgements

The reported study was funded by RFBR according to the research project №17-29-03190-ofi.

## References

- [1] Szeliski R. *Computer Vision: Algorithms and Applications*. London: Springer, 2011; 812 p.
- [2] Denisova AY, Myasnikov VV. Anomaly detection for hyperspectral imaginary. *Computer Optics* 2014; 38(2): 287–296.
- [3] Tuytelaars T, Mikolajczyk R. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision* 2008; 3(3): 177–280. DOI: 10.1561/06000000017.
- [4] Li Y, Wang S, Tian Q, Ding X. A survey of recent advances in visual feature detection. *Neurocomputing* 2015; 149: 736–751. DOI: 10.1016/j.neucom.2014.08.003.
- [5] Förstner W, Gülch E. A fast operator for detection and precise location of distinct points, corners and centres of circular features. *Proc. ISPRS intercommission conference on fast processing of photogrammetric data* 1998; 281–305.
- [6] Harris C, Stephens M. A combined corner and edge detector. *Alvey vision conference* 1988; 15(50): 147–151.
- [7] Shi J, Tomasi C. Good features to track. *Proc. Intl Conf. on Comp. Vis. and Pat. Recog (CVPR)* 1994; 593–600.
- [8] Smith SM, Brady J.M. SUSAN – A new approach to low level image processing. *International Journal of Computer Vision* 1997; 23(1): 45–78. DOI: 10.1023/A:1007963824710.
- [9] Rosten E, Drummond T. Machine learning for high-speed corner detection. *European Conference on Computer Vision* 2006; 430–443. DOI: 10.1007/11744023\_34.
- [10] Mair E, Hager GD, Burschka D, Suppa M, Hirzinger G. Adaptive and generic corner detection based on the accelerated segment test. *European conference on Computer Vision* 2010; 183–196. DOI: 10.1007/978-3-642-15552-9\_14.
- [11] Zhang X, Wang HA, Smith WB, Ling X, Lovell BC, Yang D. Corner detection based on gradient correlation matrices of planar curves. *Pattern Recognition* 2010; 43(4): 1207–1223. DOI: 10.1016/j.patcog.2009.10.017.
- [12] Shui PL, Zhang WC. Corner detection and classification using anisotropic directional derivative representations. *IEEE Transactions on Image Processing* 2013; 22(8): 3204–3218. DOI: 10.1109/TIP.2013.2259834.
- [13] Chernov AV, Myasnikov VV, Sergeyev VV. Fast Method for Local Image Processing and Analysis. *Pattern Recognition and Image Analysis* 1999; 9(2): 237–238.
- [14] Flusser J, Suk T. Pattern recognition by affine moment invariants. *Pattern Recognition and Image Analysis* 1993; 26(1): 167–174. DOI: 10.1016/0031-3203(93)90098-H.
- [15] Hu MK. Visual pattern recognition by moment invariants. *IRE transactions on information theory* 1962; 8(2): 179–187. DOI: 10.1109/TIT.1962.1057692.
- [16] Myasnikov VV. Constructing efficient linear local features in image processing and analysis problems. *Automation and Remote Control* 2010; 72(3): 514–527. DOI: 10.1134/S0005117910030124.
- [17] Theodoridis S. *Machine learning: a Bayesian and optimization perspective*. San Diego: Academic Press, 2015; 1062 p.
- [18] Alpaydin E. *Introduction to machine learning*. Cambridge: MIT press, 2014; 584 p.
- [19] Hastie T, Tibshirani R, Frieman J. *Elements of statistical learning: data mining, inference, and prediction*. London: Springer, 2011; 745 p.

# Combined method for calculating the disparity value on stereo images in problems of stereo-range metering

A.N. Volkovich<sup>1</sup>

<sup>1</sup>United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Surganova str. 6, 220012, Minsk, Republic of Belarus

---

## Abstract

The paper considers the solution of the problem of restoring three-dimensional information based on stereo images. An original combined approach to disparity calculation is proposed, as well as a variant of solving the problem of the heterogeneity of the initial data in calculating the actual metric parameters.

*Key words:* stereo images; disparity maps; gradient operators; satellite photographs; long-range systems; optical systems

---

## 1. Introduction

The tasks of comparing and image search are the main tasks in computer vision. The quality of the search, low sensitivity to distortions - the fundamental requirements for search algorithms. There are many approaches to solve such tasks, as well as a wide range of technical solutions to the problems of range-finding using lasers and other systems. At the same time there are a number of problems that imply the impossibility of active far-range systems use. In the described project is planned to develop and implement effective methods of image sections searching based on characteristics of pixel's local neighborhoods.

The relevance of the project is ensured by the need to develop methods for solving labor-consuming probabilistic-geometric problems of digital image processing. The complexity of these tasks is growing in connection to information and computing technologies development. New research methods of solution and specific applied problems are combined to the modular principle of components of all the systems being created. It makes possibilities of successful implementation of investigating problematic and applied problems.

## 2. Current state and features of the range-finding systems

The task of determining the distance to the object is extremely urgent in such areas as geodesy, military science, navigation and computer vision. Rangefinders are used to determine the distance.

Range-finding devices are divided into active and passive. The initial development of the range-finding devices was among the passive systems but in recent years the most widespread got active range-finders due to the simplicity of their implementation, their unpretentious use and the rather high accuracy of measurements.

The principle of active type rangefinder functioning consists in time measuring spent by the sent signal from the rangefinder to the object and back. The speed of the signal propagation is considered as known. Also there are active range-finders estimating changes in the parameters of the reflected signal (phase or power).

It should be noted that there are a number of problems in which the use of active range finders is difficult or not possible. These tasks include using of rangefinders on low-visibility platforms. The use of emitters leads their unmasking as well as long-distance measuring require installation of high-power emitters that could be dangerous for users.

Measurement of distances by passive range finders is based on determining the height  $h$  of an isosceles triangle  $ABC$  on the known side  $AB = l$  (stereo-base) and opposite acute angle  $b$  (so-called parallactic angle). At small angles  $\beta$  (expressed in radians)  $h = l/b$ . One of the quantities,  $l$  or  $b$ , is usually a constant, and the other is a variable (measured). By this feature, rangefinders with a constant angle and range finders with a constant base are distinguished. In general, based on the distance between the observation points  $l$  (stereo-base) and the angle of displacement  $\alpha$ , the distance to the object is calculated:

$$h = \frac{l}{2 \sin \frac{\beta}{2}}$$

For a long time, the use of stereovision principles in systems has not been considered in practice due to the relative complexity of implementation and poor quality of digital images. In recent years, the quality parameters of shooting equipment increased significantly, which suggests the possibility of developing such systems. In addition, the stereovision system can possess in addition to low-visibility (due to the implementation of the passive calculation technique) also by a functional allowing the measuring on post-production. It is possible to use stereo-pairs and recalculate the parallax for a multitude of image points in the overlapping area and the sharply displayed image space in comparison with active range finder systems which allows to obtain the distance-range information only at the time of shooting.

Based on the above facts, it seems extremely relevant to develop both theoretical methods in the field of stereo-range metering, as well as technical hardware-software solutions.

### 3. Passive systems of range-finding

A digital image obtained with a passive stereo system carries only color or brightness information and does not possess any additional data.

Images of the real world include a narrow set of colors or luminances, and therefore, when solving the problem of determining conjugate identification is not for individual points, but for fragments of images. Thus, it is extremely important for a point taken in one image to know, where at the second image is its conjugate and how to compare these fragments correctly.

A comparison of the neighborhoods of conjugate points does not yield to strict formalization. It is based on the problem of identifying images of fragments of the three-dimensional world from images. This task can hardly be adequately described in a formal way. Significant differences in views lead to the appearance of projective and brightness distortions when shooting. It is of fundamental importance that these differences depend not only on the geometry of shooting-system, but also on the geometric and physical characteristics of the surface itself. The location of the light source influences to the surface affects the light distribution. The position of the surface elements and their properties determine the amount of energy that enters the camera lenses and the local differences in the brightness of the conjugate fragments of images.

The significance of the differences depends on the difference in the viewing angles. The more this difference (in particular, the larger base), the less similar the images becomes. Therefore, all methods of comparing neighborhoods of conjugate points rely more or less on a formal approach rather than on the character of images. Also important is the possibility of their preliminary processing, reduction to an epipolar stereopair, construction of efficient descriptors of neighborhoods of conjugate points for accuracy and speed for comparison.

In an idealized situation, the values of the similarity function in the scanning process along the line should represent a one-moment peak value for the desired pixel when the zero similarity value is returned for all other pixels (neighborhoods) of the line.

During processing real graphics data, such combination of values returned by the function of similarity is impossible. But processing initial data with sufficient information to identify a local region, the graph of the function retains a sufficiently clear extremum, which allows to identify the desired pixel of the image.

The main task in the construction of the disparity map is the selection of a variant for comparison of regions in which the extremum of values of the similarity function will be most pronounced. This involves defining for the point some characteristics that would uniquely characterize the point of the image. Moreover, the conjugate point on the reference image had identical or maximally similar values of similar characteristics.

The final step of calculating disparity is the aggregation of total or averaged values. When aggregating, as the resulting disparity for the point  $p$  of the base image, the value of  $d$ , is chosen, where the minimum value of the cost is reached. In the real situation, it is possible to find several values of  $d$  with the same or minimum values of disparity (especially when averaging the values). The problem of possible ambiguity arises in most methods of constructing disparity maps, which is associated with optical, mechanical, electronic features of cameras. A solution to multi-valuedness can be the introduction of certain conditions on the value of disparity, for example, the largest, average or smallest possible.

### 4. Combined method of stereo reconstruction

Usually, in practice, measures are taken based on the sum of the absolute differences or the sum of the squares of the differences. Both functions (summation over a given window) allow you to calculate the cost effectively enough when the corresponding pixel of the conjugate image has the closest intensity value. These functions are extremely sensitive to the quality and parameters of the original data (exposure, glare, overexposure, underexposure, matrix noise, random emissions).

In the process of carrying out a computational experiment on real-world images, it was determined that correlation methods of comparing local parts of images give stable results on textured areas and extremely low accuracy in homogeneous areas (there are no explicit contrasts). During analyze of "alive" systems, we can conclude that in nature the distance to a homogeneous object is also poorly localized and its binding to the boundaries due to reflex saccadic movements.

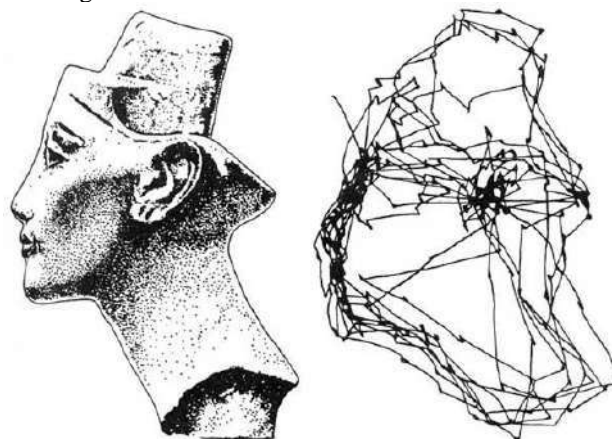


Fig. 1. Recording the movement of the eye (scanning while viewing the head of Nefertiti) according to Yarbus, 1965.

It can be argued that a comprehensive approach to the problem of ranging is required, including both consideration of the possibilities of increasing the uniqueness of the means of identification, and the development of a combined approach using different techniques to the neighborhoods of points, depending on their local characteristics.

Work with digital stereo images allowed us to determine the following combined method: correlation areas with the maximum uniqueness of points should be applied to areas of images that have contrast objects in their area (brightness differences), and to homogeneous ones - to bind to remote contrast boundaries by building a set of vectors on several directions.

Thus, in the preprocessing phase, it becomes important to compile a calculation map based on the proximity to the contrasting areas. In order to classify a point as being on a brightness difference, the brightness change associated with a given point must be substantially greater than the change in brightness at the background point. In connection with the specifics of local calculations, the way to determine the "essential" values to establish a threshold. In turn, the concepts of the first and second derivatives are used for the quantitative expression of the brightness variation.

The definition of an image point as a drop point occurs if its two-dimensional derivative of the first order exceeds a certain predetermined threshold. The calculation of the first derivative of a digital image is based on various discrete approximations of a two-dimensional gradient. The direction of the gradient vector coincides with the direction of the maximum rate of change of the function  $f$  at the point  $(x, y)$ .

$$\begin{array}{ccc} z1 & z2 & z3 \\ z4 & z5 & z6 \\ z7 & z8 & z9 \end{array}$$

The calculation of the gradient of the image consists in obtaining the values of the partial derivatives  $Gx = df/dx$   $Gy = df/dy$  for each point. One of the methods for finding the first partial derivatives  $Gx$   $Gy$  at a particular point is to apply the following gradient Sobel operator:

$$\begin{aligned} Gx &= (z7 + 2 * z8 + z9) - (z1 + 2 * z2 + z3) \\ Gy &= (z3 + 2 * z6 + z9) - (z1 + 2 * z4 + z7) \end{aligned}$$

It is necessary to determine the appropriate masks for the Sobel operator, which identifies horizontal and vertical contours (brightness differences) for convolution with the original image. It is also possible to change the above formulas that give the maximum response for contours directed diagonally. Additional pairs of Sobel's masks for detecting gaps in diagonal directions can be defined as:

$$\begin{array}{ccc} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{array}$$

for points lying on the diagonal edge -45 degrees;

$$\begin{array}{ccc} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{array}$$

for points lying on the diagonal edge +45 degrees.

For each of the masks the sum of the coefficients equals zero. That means that these operators will validate zero response on the areas of constant brightness, which is characteristic of the differential operator. The masks considered are used to obtain the gradient components  $Gx$  and  $Gy$ . To calculate the magnitude of the gradient these components must be used together:

$$Gradient = |Gx| + |Gy|$$

Calculation map is to be constructed on the computed gradient map base. This process implies the recognition of the area as low-textured in the event that in the search window around the point less than 15% of the area is occupied by contrasts.

Direct calculation of disparity occurs in several stages. At the first stage, high-textured sections are processed using the correlation functions of similarity measures, such as, for example, Euclidean distance, cross-correlation, etc.

In the world practice only brightness information is usually used as criteria for comparing image points. The disadvantage of this approach is the color interpretation multiplicity for points with the same brightness value. In addition, one should take into account the fact that the perception of color and monochrome images is uneven. This feature is taken into account in the methods of degradation of the color model of the image to 256 shades of gray due to the introduction of coefficients applied to the respective channels.

$$Y = 0.299 * R + 0.587 * G + 0.114 * B$$

Most of the images were initially formed by a color sensor in color. Therefore, in order to increase efficiency, the use of color information is seen as obvious.

Working with three components of color can be represented as a "cloud" of points in three-dimensional space with the axes corresponding to the color channels of the image. However, the RGB space is not orthogonal, due to the specificity of the human visual analyzer, which has a different number of rods and cones that are susceptible to a particular color.

Since the correlation function of a three-dimensional space is a measure of correlation functions that use the Euclidean distance, which is correctly calculated in orthogonal systems, one should perform orthogonalization of the space RGB into the space XYZ.

The representation of the RGB base colors, according to the ITU recommendations, in the XYZ space has the following correction factors:

$$\begin{aligned} Red: x &= 0,64 \quad y = 0,33 \\ Green: x &= 0,29 \quad y = 0,60 \\ Blue: x &= 0,15 \quad y = 0,06 \end{aligned}$$

Therefore the transformation system for translating colors between RGB and XYZ systems can be represented in the following form:

$$X = 0,431 * R + 0,342 * G + 0,178 * B$$

$$Y = 0,222 * R + 0,707 * G + 0,071 * B$$

$$Z = 0,020 * R + 0,130 * G + 0,939 * B$$

After reduction of spaces, operations that are valid for orthogonal systems to points can be applied. Using the "color" image processing increases the potential uniqueness of the point 1.72 times (the maximum distance between the luminance values in the gray scale is 255 units, color values 416 units).

After this step is performed the distances from the points in the low-textured regions to the nearest contours in several directions are calculated. A group of multidimensional characteristic vectors is formed, to which a similar approach is applied, as well as to vectors with color information.

As the stages are completed, the disparity map is filled. Due to the fact that all operations are in strict accordance with the calculation card, auto-aggregation of the results of different stages into a single map occurs.

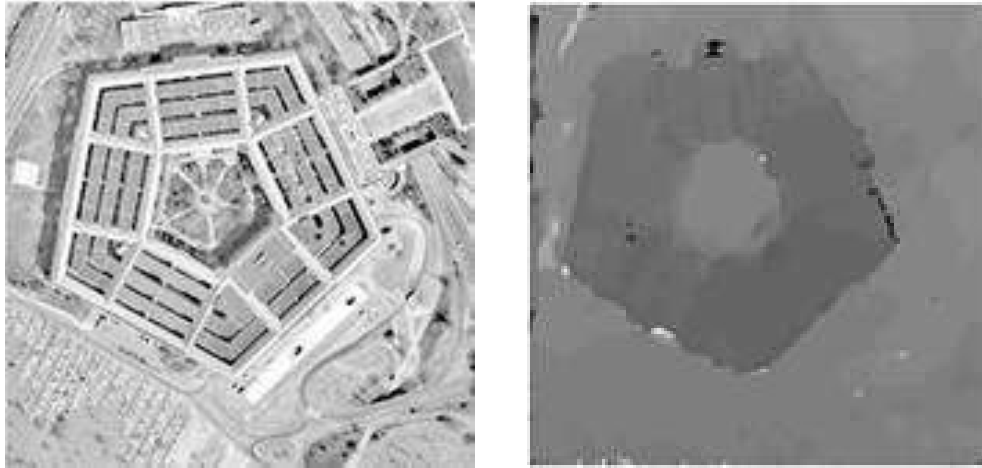


Fig. 2. Image of stereopair and map of disparity.

Despite the multi-stage implementation, this algorithm has a large number of cyclic stereotyped locally independent operations, which makes it possible to parallelize the algorithm.

## 5. Calculation of the distance to the object on the basis of heterogeneous initial data

Generalized the principle of determining the position of points in space on the basis of disparity data has been repeatedly described in the literature. Suppose two cameras  $L$  and  $R$  are installed in such a way that their  $X$ -axes are collinear, and the  $Y$  and  $Z$ , axes are parallel. The centers of the cameras are displaced relative to each other by an amount  $b$ , corresponding to the base of the stereoscopic system. When observing a certain point of the space  $P$  the point  $P_l$  is formed on the left image, and on the right  $P_r$ .

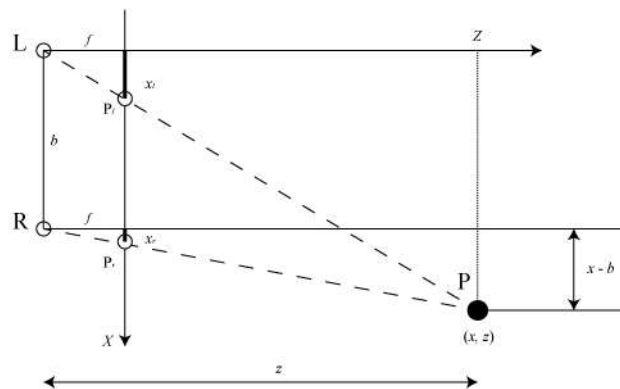


Fig. 3. Geometric model of a stereoscopic system.

Considering the similarity of two pairs of triangles, we obtain the equations:

$$\frac{z}{f} = \frac{x}{x_l} \quad \frac{z}{f} = \frac{x-b}{x_r} \quad \frac{z}{f} = \frac{y}{y_l} = \frac{y}{y_r}$$

It should be noted that by construction, the coordinates of the image points  $y_l$  and  $y_r$  can be considered the same, which corresponds to the rectified system with a rigid connection between the photosystems (3D camera, human visual system). Given this property, it is possible to transform the system of equations for the explicit expression of the coordinates  $x$ ,  $y$ ,  $z$  of  $P$  in real space on the basis of the coordinates of the projections of points on stereopair images:

$$z = fb/(x_l - x_r) \quad x = \frac{x_l z}{f} = b + \frac{x_r z}{f} \quad y = \frac{y_l z}{f} = \frac{y_r z}{f}$$



The solution of the system of equations allows one to uniquely calculate the position of a point in space.

Unfortunately, the given system of equations is not applicable for digital reconstruction, because there is a mixture of different systems of dimension: disparity value in pixel distance, focal length and base in metric units. However, it is possible to calculate the distance to the point using the disparity value, the angle of the lens alignment and the base of the stereo system.

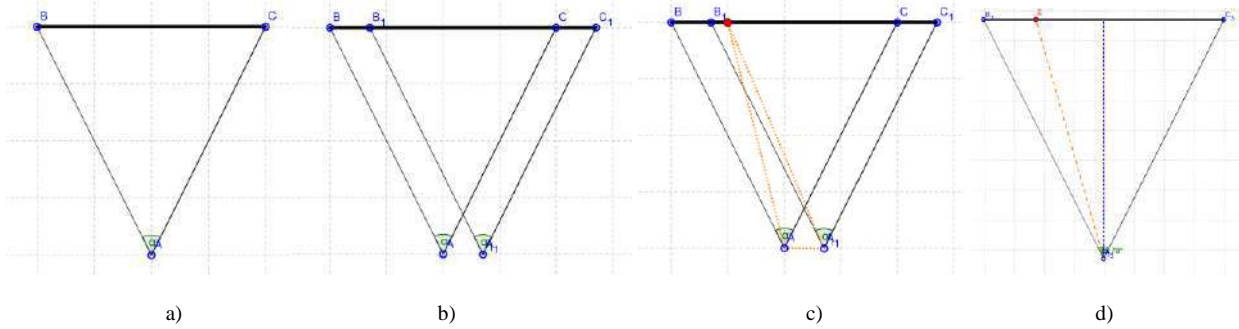


Fig.4. Geometrical model of calculating distance on the lens alignment.

The stereovision system can be represented in the following form:

- A, A<sub>1</sub> – observation point;
- BC – left image;
- B<sub>1</sub>C<sub>1</sub> – right image;
- BC<sub>1</sub> – zone of overlap;
- α – horizontal lens opening angle.

The calculation of the distances to the object is the problem of solving the triangle B<sub>1</sub>AA<sub>1</sub>.

Within the system, the base of the triangle is known - the base of the stereo system. Angles at the base can be calculated through the angle of the lens opening.

The calculation of the viewing angle to the object of interest for the left camera is performed as follows (Fig. 4c):

$$\angle EAA_1 = \beta = 90 \pm \arctg\left(\frac{\operatorname{tg} \frac{\alpha}{2} * \left(\frac{BC}{2} - BE\right)}{\frac{BC}{2}}\right)$$

Where:

- β – angle of the desired triangle;
- α – angle of the horizontal lens opening;
- BE – the X- coordinate of the image point;
- BC – the width of the image (Fig 4d).

Знак «+» используется в системе при  $BE < \frac{BC}{2}$ , «-» при  $BE > \frac{BC}{2}$  соответственно, а при  $BE = \frac{BC}{2}$  принимаем  $\angle EAA_1 = \beta = 90$

The angle of sight is calculated in the same way as the correction for the fact that the sign «-» is used in the system for  $BE < \frac{BC}{2}$ , «+» for  $BE > \frac{BC}{2}$  and for  $BE = \frac{BC}{2}$  and  $\angle EA_1A = \beta_1 = 90$ .

The third angle can be obtained by the formula:

$$\beta_2 = 180 - \beta - \beta_1$$

By the sine theorem, it is possible to determine the lengths of the sides A<sub>1</sub>E and AE

$$AE = AA_1 \frac{\sin \beta}{\sin \beta_2} \text{ and } A_1E = AA_1 \frac{\sin \beta_1}{\sin \beta_2}$$

Due to the possible inclination of AE relative to the horizontal plane of the system, it is necessary to bring EA into the plane of the system

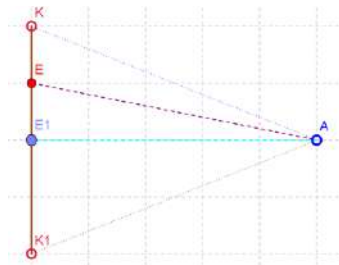


Fig. 5. Geometrical model of vertical declination of the system.

In situation of  $KE < KK_1/2$

$$\vartheta = \arctg\left(\frac{\operatorname{tg} \frac{\alpha}{2} * \left(\frac{KK_1}{2} - AE\right)}{\frac{KK_1}{2}}\right)$$

In situation of  $KE > KK_1/2$

$$\vartheta = \arctg\left(\frac{\operatorname{tg} \frac{\alpha}{2} * \left(AE - \frac{KK_1}{2}\right)}{\frac{KK_1}{2}}\right)$$

In situation of  $BE = BC/2$

$$\begin{aligned}\vartheta &= 0 \\ \vartheta &= \angle EAE_1 \\ AE_1 &= AE \cos \vartheta\end{aligned}$$

As a result, the distance from the reference (left) camera to the object and the angle relative to the base (plane of the matrices) of the system are obtained.

It should be noted that increasing the measured distance increases the sensitivity of the system to the accuracy of the alignment of the system and the quality of the images. Since the angles at the base of the system take values close to  $90^\circ$ . This leads to the fact that the values of trigonometric functions change very dynamically.

## 6. Conclusion

During the research, the author made a study of the existing methods for processing stereo images in the tasks of stereo reconstruction. The algorithms functioning patterns are revealed and the causes of their unstable work are determined. A combined image processing technique that takes into account the characteristics of local image sections is proposed. Additionally author examined the problem of the heterogeneity of the initial data necessary for obtaining metric information of three-dimensional objects in the field of interests.

The algorithm developed and described was implemented by the author in the form of a program library, which can later be used in a wide range of applications. Due to the fact that the matrix of distances to image points can be translated into a specific coordinate system for one or another application system. It should also be noted that the organization of the user's access to the functions allows for more flexible use of the library.

In addition, the described technique has found its application in a number of software and hardware and software developments that are carried out at the United Institute of Informatics Problems of the National Academy of Sciences of Belarus. Specifically, as element of the mobile topogeodetic system and as program library for ERS system.

## References

- [1] Borodach A, Tuzikov A. Automatic determination of matching points on two images. Proceedings of the 9th International Conference "Pattern Recognition and Information Processing", 22-24 May, Minsk, Belarus 2007; 1: 49–53.
- [2] Shapiro L. Computer vision. Moscow: BINOM. Laboratory of Knowledge, 2006; 752 p. (in Russian)
- [3] Volkovich AN. Use of color characteristics in the construction of disparity maps. Materials of the International Congress of ROPI-2011. Nizhny Novgorod: UNN 2011; 64: 112–117. (in Russian)
- [4] Lyakhovsky VV, Volkovich AN, Zhuk DV, Tuzikov AV. Sistem automatic reconstruction of three-dimensional scenes for several images. Materials of the V Belorussian Space Congress, October 25-27, Minsk 2011; 2: 129–133.
- [5] Zhuk DV, Tuzikov AV. Reconstruction of a three-dimensional model using two digital images. Informatics 2006; 1: 16–26.
- [6] Shulgovsky VV. Fundamentals of Neurophysiology. URL: <http://www.braintools.ru/rubric/information/from-books/fundamentals-of-neurophysiology> (01.02.2017).

# Increasing the energy efficiency of OFDM systems using differential signal conversion

G.S. Voronkov<sup>1</sup>, I.V. Kuznetsov<sup>1</sup>, A.Kh. Sultanov<sup>1</sup>

<sup>1</sup> Ufa State Aviation Technical University, 12, K. Marx St., 450000, Ufa, Russia

---

## Abstract

A method for improving the energy efficiency of OFDM systems based on differential transformation and extrapolation is considered. The possible structure of the extrapolator is analyzed, the implementation of an extrapolator based on the Kalman - Bucy filter and the Wiener filter is considered. The results of carried out simulation confirming the effectiveness of the proposed scheme are given.

*Keywords:* OFDM; dynamic range; compression; extrapolation; energy efficiency; differential method

---

## 1. Introduction

One of the main directions of the development of telecommunication systems nowadays is the data rate increase. Quadrature modulation and orthogonal frequency division multiplexing are used to solve this problem. Mobile networks of 4th and 5th generation [1,2] and modern satellite communications [3] can be referred as the examples of this. But data rate increase requires signal-to-noise ratio increase while keeping channel parameters unchanged. This cause transmitting power increase. The power increase, in turn, makes the process of microwave end amplifiers development and leads the devices power consumption increase which decreases their recharge interval. Various methods are currently used to reduce the radiated power. Base station (BTS) manages the mobile terminal (MT) power using control channel in 2d generation mobile networks (GSM, DCS). BTS measures the received from MT signal strength during the communication session and adjusts MT radiated power. This is an iterative process, MT power is being reduced gradually until optimal threshold level [4]. When degradation of communication quality is detected, BTS increases MT radiated power gradually in the same way until its maximum. This method obviously requires dedicated control channel. Using automatic control theory terms, it is feedback path control.

Another power control procedure is used in 3d and 4th (UMTS, LTE) generation mobile networks. Mobile terminals in these networks transmit and receive signals in the same frequency channel in one BTS cell sector, so MT output power is varied not from its maximum value to the minimum, but on the contrary – from the minimum to the optimal value that guarantees the predefined communication quality (to reduce electromagnetic disturbance for the other MT). Output power is being changed according the base station commands [1,5]. It is important to note that the method described doesn't reduce transmitter maximum output power but only adjusts its according to the signal propagation conditions to provide predefined communication quality.

Some methods allow maximum output power decrease, that provides recharge interval and life utility [6] growth. For example, multi-dimensional signal constellation method is known. It allows to reduce output power by 1 dB [7], but requires quadrature modulators of consisting telecommunication system replacing, because it changes constellation forming algorithm. So, this method increases the complexity of devices.

It is suggested to use differential transformation of OFDM signals to reduce a transmitter output power without communication quality decrease and device complexity increase. Differential transformation means OFDM signal dynamic range decrease by extrapolating of its values. In this case, the extrapolator transfer function must be synthesized considering the signal properties and the channel noise. Differential transformation allows to reduce OFDM band signal power while keeping communication system noise stability unchanged, that is considered in this paper.

## 2. The extrapolation method

Two methods of differential transformation can be offered in general: using input control (“input method”) or output control (“output method”). OFDM signal generating and receiving general scheme is given in Fig. 1.

The extrapolator parameters should be known on the receiving side to receive transformed signal correctly. These parameters are suggested to be transferred using secondary communication channel to reduce the calculation amount. If extrapolator parameters are changed slow compared with signal, secondary channel may be considered as a lossless channel, so transmitter and receiver extrapolators are equal.

In the differential transformation “input method” schema design Kalman – Bucy filter can be used as the extrapolator. Such schema advantages are:

- 1) considering channel noise directly in the model;
- 2) both stationary and non-stationary signals processing;
- 3) solving the problem in digital form.

The schema disadvantages are:

- 1) significant prediction error in the initial stages of signal observation;
- 2) problem of stability guaranteeing.

Another disadvantage is watch and state equations system solving necessity, which is somewhat complicated. Taking into consider these disadvantages, it is suggested to use the schema of differential transformation based on “output method” and to synthesize extrapolator transfer function based on Wiener-Hopf equation solution. It is also suggested to reduce the number of schema elements using one common extrapolator with coordination function to process both signals of in-phase and quadrature channels as it is shown in [5]. The calculation model used for this case shown in Fig. 2.

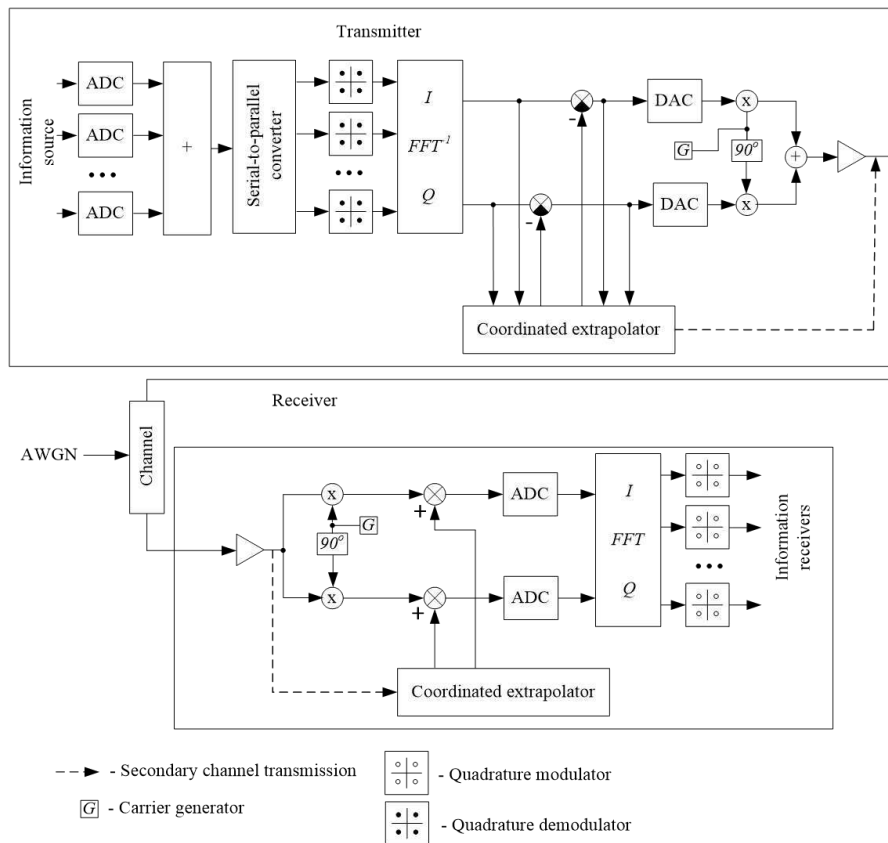


Fig. 1. OFDM signal with differential transformation generating and receiving general.

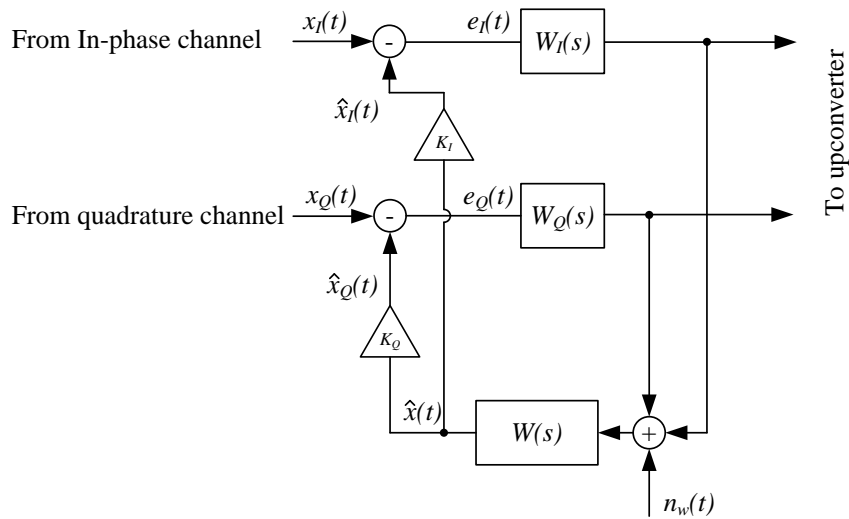


Fig. 2. OFDM-signal transmitter with coordinated extrapolator calculation model

Functions designated as  $x_I(t)$ ,  $x_Q(t)$  are OFDM band signal in-phase and quadrature components on the invers Fourier transform unit output,  $\hat{x}_I(t)$ ,  $\hat{x}_Q(t)$  – extrapolated values of the corresponding functions,  $e_I(t)$ ,  $e_Q(t)$  – difference signals. Channel noise  $n_w(t)$ , which spectral power density (SPD)  $\Phi_{NN}$  is assumed to be known, is also taken in consideration in the proposed extrapolator schema transfer function  $W(s)$  ( $s$  – complex Laplace variable) synthesis process. Transfer functions designated as  $W_I(s)$  and  $W_Q(s)$  are introduced to describe the transformation and delay processes in the in-phase and quadrature channels and are also known. The transfer function synthesis is described in [4, 5]. In general, the extrapolator transfer function can be represented in the form below:

$$W(s) = \frac{1}{\Phi_{zz}^+(s)} \left[ \frac{\Phi_{zx}(s)}{\Phi_{zz}^-(s)} \right]_+$$

where  $\Phi(s)$  – cross-spectral density of signals designated in the index;

$x(t)$  – OFDM band signal in-phase or quadrature (according to the index) component;

$z(t)$  – OFDM band signal in-phase or quadrature component and channel noise mixed signal,  $z(t) = x(t) + n_w(t)$ .

The solution for the special case of PSD that represented in the form below was obtained in [5]:

$$\Phi_{xx}(s) = \frac{1}{a_1^2 (\alpha^2 - s^2)},$$

$$\Phi_{mm}(s) = \frac{1}{a_2^2}.$$

Extrapolator transfer function in the case described is:

$$W(s) = \frac{K_1}{1 - K_1 + \tau s},$$

where

$$K_1 = \frac{a_2^2}{\sqrt{a_1^2 \alpha^2 + a_2^2} (a_1 \alpha + \sqrt{a_1^2 \alpha^2 + a_2^2})},$$

$$\tau = \frac{a_1}{\sqrt{a_1^2 \alpha^2 + a_2^2}}.$$

Simulation modelling was done for the extrapolator parameters described above.

### 3. Simulation modelling

The described solution simulation modeling was carried out in the MatLab. Channel parameters are: channel bandwidth is equal 8 MHz, OFDM subcarriers number – 16. The channel is modeled as a medium with additive white Gaussian noise (AWGN), which power determines the signal-to-noise ratio at the reception at 22 dB. Multipath propagation and radio blackout aren't considered in the model. These baselines and assumptions are correct when the method proposed is being used for the satellite communication channel or line-of-sight channel between pilotless vehicle and terrestrial management and its control center. QPSK is used for quadrature modulation. Earth remote sensing system picture is used as the source of information. Picture size is 512x512 pixels, image format is BMP without compression, color mode is grayscale. The source picture and its histogram are given in Fig.3

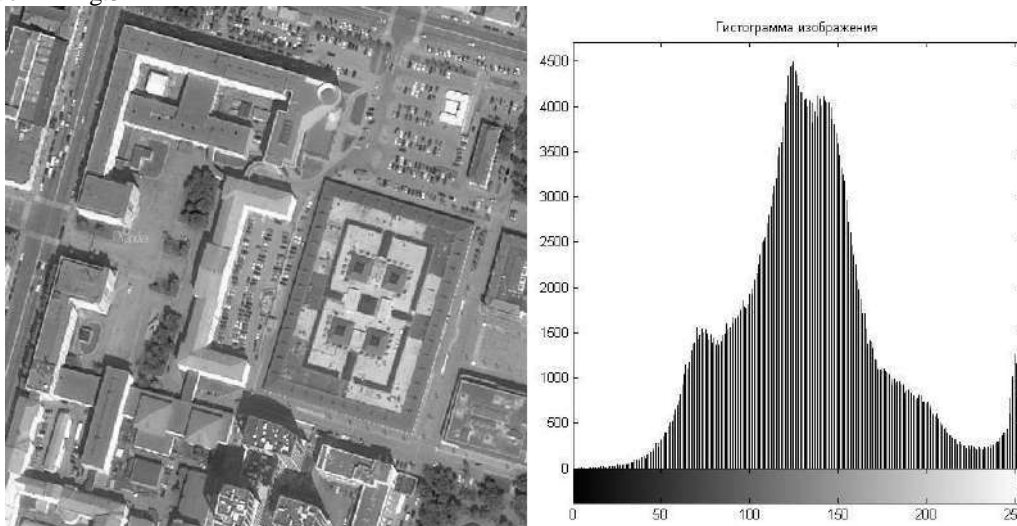


Fig. 3. The source image and its histogram.

The graphic file is divided into 32x32 pixel fragments, then each of the resulting 256 fragments is being converted to a binary format, after which the QPSK symbols are generated. These symbols are subjected to inverse fast Fourier transform, which allows to obtain the first type of band-pass signal (signal 1). The signal generated is being convoluted with the extrapolator impulse responses, the difference between the original signal and its extrapolated value is calculated, thus, a second type of band signal is generated - a signal compressed in a differential scheme (signal 2). In-phase and quadrature signal shapes before upconversion for one of the fragments are given in Fig.4. Signal 1 and signal 2 average powers are calculated after that. Third type of signal (signal 3) is being formed with decreasing signal 1 amplitude until this signal power be equal signal 2 power. Transmission along the AWGN channel is simulated after that. The received band signal is being transformed using fast Fourier

transform. Since 3 variants of the band signal were previously obtained, it is possible to compare QPSK signal constellations obtained by processing different band signals and to estimate the symbol error. The received signal constellations for one of the fragments transmission case are shown in Fig. 5, from left to right: a signal constellation for signal 1, signal 2 and signal 3.

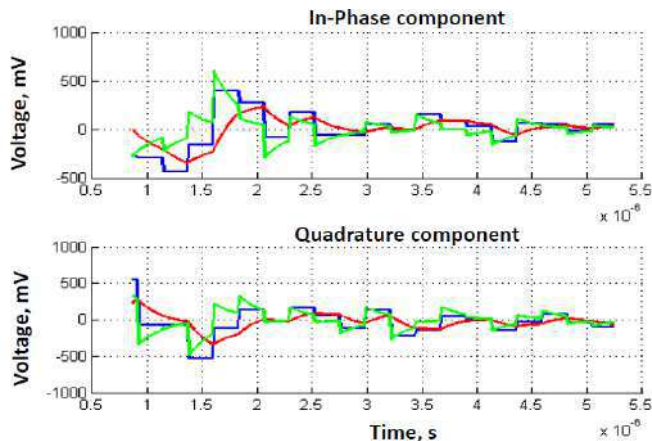


Fig. 4. Signal shapes before upconversion. Blue – uncompressed signal; Red – extrapolated signal; Green – their difference.

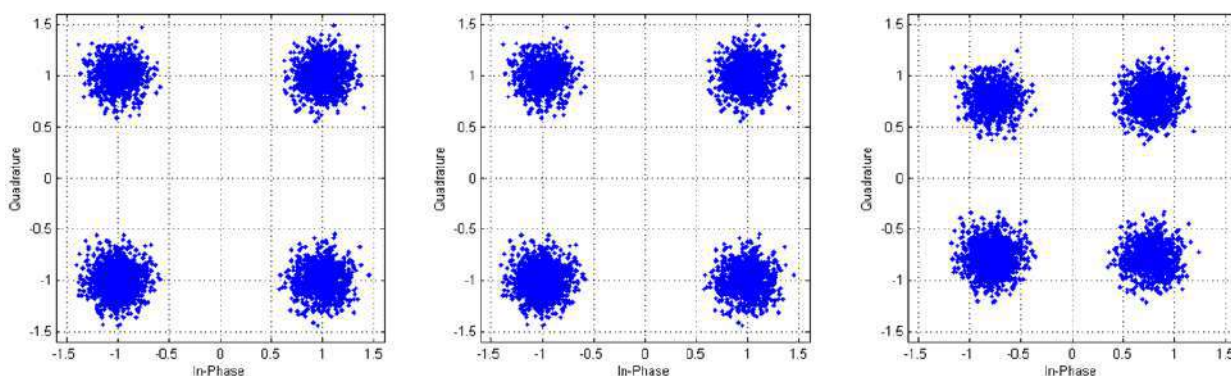


Fig. 5. Received signal constellations.

Since the original image was divided into 256 fragments, there were obtained 256 power gain values for the differential transformation using as a result of modeling. The histogram of compression levels is shown in Fig. 6. The horizontal axis shows the compression ratios, dB, vertical axis - the number of information parcels.

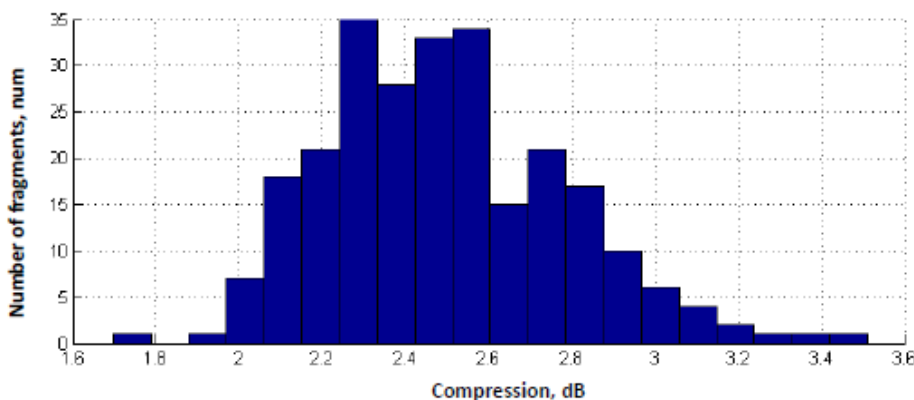


Fig. 6. Dynamic range reduction histogram.

According to the results of the simulation, the differential conversion made it possible to reduce the power of the band signal by 2.49 dB relative to the original signal. A corresponding reduction in transmitting power without extrapolation leads to an increase in the symbol error. For the uncompressed signal and for the signal subjected to differential transformation, the average symbol error ratio was about  $0.000947 \text{ s}^{-1}$ , while for the signal with a reduced power the average symbol error ratio was about  $0.0638 \text{ s}^{-1}$ . The result can also be interpreted in a different way: the use of differential transformation makes it possible to reduce the required signal-to-noise ratio at reception by an average of 2.49 dB without degrading the communication quality.

#### 4. Conclusion

The computational model of OFDM signal differential transformation of based on its extrapolation is presented in the paper. Proposed solution simulation for a short-range radio channel with AWGN without signal multipath propagation and radio

blackout is made. The high value of the signal-to-noise ratio given in the experiment is explained by the absence of noise-proof coding in the simulation program. The results of the simulation confirm that the proposed method of differential transformation allows to reduce the amplitude and reduce the average power of the band signal without impairing the noise immunity of the system. Thus, the proposed OFDM signal generation scheme allows to increase the communication system energy efficiency without degrading communication quality.

## Acknowledgment

This work is supported by the Ministry of Education and Science of Russian Federation under the Basic part of the State assignment for higher education organizations 8.5701.2017/BCh.

## Reference

- [1] LTE for UMTS: OFDMA and SC-FDMA Based Radio Access. Edited by Harri Holma and Antti Toskala. John Wiley & Sons Ltd, 2009.
- [2] Fundamentals of 5G Mobile Networks. Edited by Jonathan Rodriguez. John Wiley & Sons Ltd, 2015.
- [3] ITU-R M.2047-0. Detailed specifications of the satellite radio interfaces of International Mobile Telecommunications-Advanced (IMT-Advanced).
- [4] Gromakov JuA. Mobile radio standards and systems. ECO-TRENDS. Moscow, 1998. (in Russian)
- [5] GSM, GPRS and EDGE performance. Evolution Towards 3G/UMTS. Edited by Timo Halonen, Javier Romero, Juan Melero. John Wiley & Sons Ltd, 2003.
- [6] Filatov PE. Increasing the energy-deficient multi-channel communication systems efficiency based on coordinated signal conversion. Applied electro-dynamics, photonics and living systems -2016. International Scientific and Technical Conference, 2016; 143–148. (in Russian)
- [7] Markiewicz Tomasz G. An Energy Efficient QAM Modulation with Multidimensional Signal Constellation. International Journal of Electronics and Telecommunications 2016; 62(2): 159–165. DOI: 0.1515/eletel-2016-0022.
- [8] Kuznetsov IV, Voronkov GS, Sultanov AKh, Antonov VV. Differential OFDM-converter for energy deficient communication system based on coordinated signal predictor design. Radioengineering 2016; 12: 59–63.

# Complex Matrix Model for Data and Knowledge Representation for Road-Climatic Zoning of the Territories and the Results of Its Approbation

A. Yankovskaya<sup>1,2,3,4</sup>, A. Sukhorukov<sup>2</sup>

<sup>1</sup>Tomsk State University of Architecture and Building, 634003, Tomsk, Russia

<sup>2</sup>National Research Tomsk State University, 634050, Tomsk, Russia

<sup>3</sup>National Research Tomsk Polytechnic University, 634050, Tomsk, Russia

<sup>4</sup>Tomsk State University of Control Systems and Radioelectronics, 634050, Tomsk, Russia

---

## Abstract

Complex matrix model of data and knowledge representation is proposed for solution of a road-climatic zoning of the territories problem using an intelligent system. This model consists of: 1) an extended matrix model, which includes extended description and distinguishing matrices (the extension is realized by the way of including of additional columns into the description matrix) for the territories under investigation, 2) description and distinguishing matrices of highly qualified experts' knowledge and 3) a partial matrix model, consisting of an extended description matrix of the territories under investigation (recognition). For the first time original approbation results of intelligent data and knowledge analysis on the base of intelligent instrumental software IMSLOG are given. The system is designed and developed in intelligent systems laboratory of the Tomsk State University of Architecture and Building to solve the problem of road-climatic zoning.

*Keywords:* complex matrix model; intelligent data analysis; approbation results; geocomplex, road-climatic zoning

---

## 1. Introduction

An urgent need in intelligent systems (IS) application for a number of problem areas is not in doubt. Among the basic IS applications are those given in the monograph [1]: medicine, engineering, transport system and others. Among the fundamental IS components we distinguish the data and knowledge base. In this paper we will concentrate our efforts on the base construction.

When developing the design standards for the highways we should take into account the regional features of the geographic territories. It is performed through the method of road-climatic zoning. The method serves as a basis for the development of building regulations, directives and guidelines valid in Russia [2,3], China [4,5], USA [6], Germany [7], Great Britain, Sweden [8] and in other countries, including such neighboring countries as Kazakhstan [9], Belorussia [10], Kyrgyzstan [11]. According to the building regulations [2,3] the Russian Federation territory is zone differentiated and divided into 5 road-climatic zones, which are differing sufficiently in terms of the complexes of nature-climatic and geoengineering conditions. In their turn, the zones are divided into 9 subzones due to the road industry standards [9] and into 13 subzones due to the set of rules [3]. Depending on the position of the road section under design in one or another zone and subzone the road designers make technical decisions, providing safe and convenient traffic according to the requirement stated in [2,3].

A number of researchers in their papers [12–15] point out that the existing special position of the zones and subzones boundaries does not allow to provide the level of operational reliability of the highways due to the operability criterion since the position of the zones and subzones are not substantiated sufficiently. This situation is especially inherent to Western Siberia and Far East. That leads to an increase in financial and labor resources for maintaining and restoring the required technical condition of highways. Thus the research of the new approaches to the road-climatic zoning design is rather actually. The specificity of data and knowledge to solve the problem of the road-climatic zoning requires the new methods of the data and knowledge representation. Taking into consideration is proposed to choose intelligent instrumental software IMSLOG (IIS IMSLOG) [16,17] for IS construction of road-climatic zoning of territories (IS RCZT).

Hereafter we give a description of a complex matrix model for the data and knowledge representation for the IS RCZT. The IS RCZT is based on test methods of pattern recognition and cognitive graphics tools.

## 2. Complex Matrix Model for Data and Knowledge Representation

For the first time we suggest to represent the complex model by 3 types of the following matrix models of data and knowledge representation [18,19]:

1. An extended matrix model including an extended matrix of descriptions and a matrix of distinguishing. The extension is performed due to additional columns introduction to the matrix of descriptions [19]. The matrix of descriptions sets the objects description within the space of characteristic features. The additional columns correspond to compulsory features: zones, subzones, road districts, supporting point of the investigated territories. The present paper deals with the research results on Western Siberia territory. The columns of the extended matrix of descriptions correspond to characteristic features, represented by the 3 groups of factors. Those factors constitute the geographical complex of the territory: zonal, intrazonal and regional factors. The columns of the matrix of distinguishing correspond to the zones, subzones and road districts. We use integer features in the model.



2. A matrix of expert knowledge description without compulsory features and a matrix of distinguishing. The columns of the matrix of distinguishing, as well as those of the matrix of distinguishing of the extended matrix model, correspond to the zones, subzones and road districts. Here we also use integer features.

3. A partial matrix model, consisting only of the extended matrix of description of the investigated territories under recognition.

Now we concentrate on the elements description of the matrices under consideration. The integer values of the characteristic features including grouped ones and the compulsory features are the elements of the extended matrix of objects' description ( $Q^e$ ). The group characteristic features are split into the features of the integer values, which correspond to a certain partition intervals of the feature under study. A column of the matrix  $Q^e$  corresponds to each characteristic features. A row of the matrix  $Q^e$  corresponds to the stronghold for which the values of characteristic features are determined. Thus, the element of the matrix  $Q^e$  is the value of the integer characteristic features, including compulsory one. This feature correspond to a certain supporting point [18]. Note that the compulsory features are not used in regulations revealing. They are implemented only for the mapping of the zones, subzones and road districts.

Integer values of the classification features of three types are the elements of the matrix of distinguishing ( $R^e$ ). We restrict our study to the diagnostic matrix of distinguishing. For the matrix under study each subsequent column splits the previous one into the classes of equivalency. Due to the methodology given in [20,21], we will use the three classification features of diagnostic type. The 1st feature corresponds to the zones, the 2nd one – to the subzones, the 3rd one – to the road districts.

We note that we need the 2nd matrix model due to the incomplete information on the zones, subzones and road districts. Such information is contained in the 1st matrix model. The rows of the matrix of description and the matrix of distinguishing are fulfilled by the highly qualified experts in the problem area. Matrix fulfillment with data is performed by the colleagues of the Automobile roads building department of the Tomsk State University of Architecture and Building.

The learning sample is represented by the extended matrix model. In the learning sample some combinations of the classification features could not be represented. Therefore, the dimensions of the matrix of description and the matrix of distinguishing could exceeded sufficiently the dimensions of the matrix of the extended matrix model, fulfilled beforehand. This is due to the absence of a number of combinations of the classification features values in the learning sample.

The extended matrix of description of the territories under recognition for the partial matrix model is fulfilled by the highly qualified experts. They use the reference data and the data acquired during field and/or laboratory research. The research results could be transmitted both to the system's users and to the enterprises, interested in the road-climatic zoning research results.

The decision making about the supporting point correspondence to a certain zone, subzone and road district we perform using the two aforementioned matrix representations (extended matrix and the one based on the expert knowledge) based on the rules of the total decision making with use of IS RCZT. The architecture IS RCZT is presented in the publication [22].

### 3. Data and knowledge structuring. Bases of a database and knowledge construction

The basis of the information technology of road-climatic zoning of territories is IS RCZT. To create the IS we united our efforts with our colleagues. Together with specialists in the cognitive science and experts in road-climatic zoning we have structured the data and knowledge on road-climatic zoning. The structuring has been performed based on the complex matrix model of data and knowledge representation, described in section 2.

A list of characteristic features with indicating their values for the matrices of description is given in Table 1. The characteristic features are grouped ones beginning with characteristic features  $z_{10}$ . Symbolic characteristic features, intervals of the integer characteristic features partitions as well as the real characteristic features are coded by numbers. The number 20 (limiter) is used only for the sake of size reduction of the data and knowledge matrix representation. In table 1, the value of an integer feature is not greater than 8.

Table 1. A list of characteristic features excluding compulsory ones.

Characteristic feature	Code	Intervals of values
Vegetation type	$z_1$	1 – tundra vegetation; 2 – forest-tundra vegetation; 3 – forest vegetation (northern taiga, with propagation of permafrost soils); 4 – forest vegetation (middle taiga); 5 – forest vegetation (southern taiga); 6 – forest-steppe vegetation; 7 – steppe vegetation; 8 – desert and desert steppe vegetation
Terrain relief	$z_2$	1 – flat terrain with a relative elevation of the relief (RER) up to 25 m; 2 – hilly with RER from 25 m up to 200 m; 3 – mountainous (low mountains terrain) with RER from 200 m up to 500 m, and with a prevailing slope gradient (PSG) from 5° up to 10°; 4 – mountainous (mid-mountain terrain) with RER from 500 m up to 1000 m, PSG from 10° up to 25°, and elevation above sea level of about 1000–2000 m; 5 – mountainous (highland terrain) with RER from 1000 m, PSG more than 25°, and elevation above the sea level more than 2000 m
Calculated soil moisture (CSM), p.u.	$z_3$	1 – low soil moisture with CSM up to 0.4; 2 – normal soil moisture with CSM from 0.4, up to 0.6; 3 – increased soil moisture with CSM from 0.6, up to 0.8; 4 – waterlogged soil with CSM from 0.8, up to 1
Evaporation from the land surface, mm/year	$z_4$	1 – extremely low, from 100 mm up to 150 mm (arctic deserts); 2 – very low, from 150 mm up to 200 mm (Siberian tundra provinces); 3 – low, from 200 mm up to 400 mm; 4 – average, from 400 mm up to 600 mm (taiga, central and central black earth regions of Russia, Krasnodar region); 5 – increased, from 600 mm up to 700 mm (mixed forests); 6 – high evaporation, from 700 mm up to 800 mm; 7 – very high evaporation, from 800 mm up to 900 mm (steppers); 8 – extremely high, from 900 mm up to 1000 mm (semi-deserts and deserts)

Continuation table 1.

Characteristic feature	Code	Intervals of values
Syelyaninov's hydrothermic coefficient	z <sub>5</sub>	1 – redundant moistening of the soil with SHC from 1,4 to 5; 2 – significant moistening of the soil in particular years with SHC from 1 to 1,4; 3 – insufficient moistening of the soil with SHC from 0,5 to 1; 4 – dry regions with SHC up to 0.5
A number of days with negative air temperature	z <sub>6</sub>	1 – low from 141 to 198; 2 – medium from 199 to 246; 3 – high from 247 to 315
Snow cover height (SCH), mm	z <sub>7</sub>	1 – snowless regions with SCH up to 300; 2 – little snow cover regions with SCH from 300 to 500; 3 – medium snow cover regions with SCH from 500 to 700; 4 – high snow cover regions with SCH from 700 to 1000; 5 – exclusive high snow cover regions with SCH from 1000 to 2900
Soil frost depth (SFD), cm	z <sub>8</sub>	1 – small frost depth with SFD from 50 to 180; 2 – medium frost depth with SFD from 180 to 220; 3 – high frost depth with SFD from 220 to 260; 4 – very high frost depth with SFD from 260 to 300; excessive frost depth with SFD from 300 to 600
Soil type according to natural condition I zone	z <sub>9</sub>	1 – continuous distribution of the frozen soils for many years; 2 – continuous in general of the frozen soils for many years; 3 – predominately island distribution of the frozen soils for many years
Average air temperature for many years, °C	z <sub>10</sub>	1 – extremely low temperature with AAT from –15.5 to –10.0; 2 – very low temperature with AAT from –10.0 to –6.0; 3 – low temperature with AAT from –6.0 to –2.0; 4 – medium temperature with AAT from –2.0 to 2.0; 5 – high temperature with AAT from 2.0 to 6.0; 6 – very high temperature with AAT from 6.0 to 10.0; 7 – extremely high temperature with AAT from 10.0 to 14.2
Average minimum air temperature, °C	z <sub>11</sub>	1 – extremely low temperature less than –40.0; 2 – very low temperature from –39.9 to –32.0; 3 – low temperature from –31.9 to –24.0; 4 – medium temperature from –23.9 to –16.0; 5 – high temperature from –15.9 to –8.0; 6 – very high temperature from –7.9 to 0.0; 7 – extremely high temperature above 0.0
Average annual maximum air temperature, °C	z <sub>12</sub>	1 – extremely low temperature from 0 to 4; 2 – very low temperature from 4 to 7; 3 – low temperature from 8 to 11; 4 – medium temperature from 12 to 15; 5 – high temperature from 16 to 19; 6 – very high temperature from 20 to 23; 7 – extremely high temperature above 24
Annual precipitation, mm	z <sub>13</sub>	1 – low less than 250; 2 – medium from 251 to 500; 3 – high from 501 to 1000; 4 – very high above 1000
Annual precipitation for the cold season, mm	z <sub>14</sub>	1 – low less than 60; 2 – medium from 61 to 150; 3 – high from 151 to 405; 4 – very high above 405
Annual precipitation for the warm season, mm	z <sub>15</sub>	1 – low less than 190; 2 – medium from 191 to 340; 3 – high from 341 to 600; 4 – very high above 600
Soil humidity on the liquid limit, p.u.	z <sub>16</sub>	1 – low from 0.29 to 0.33; 2 – medium from 0.34 to 0.38; 3 – high from 0.39 to 0.43
Soil humidity on the plastic limit, p.u.	z <sub>17</sub>	1 – low from 0.20 to 0.23; 2 – medium from 0.24 to 0.26; 3 – high from 0.27 to 0.30
Plasticity index, %	z <sub>18</sub>	1 – non-cohesive soil (sand, etc.) from 0 to 1; 2 – clay sand from 1 to 7; 3 – light clay loam from 7 to 12; 4 – heavy clay loam from 12 to 17; 5 – light clay from 17 to 27; 6 – heavy clay from 27 and above
Grain-size composition of the clay sands, sand grain content, mass %	z <sub>19</sub>	1 – clay sand above 50; 2 – pulverescent clay sand less than 50
Grain-size composition of the clay sands, sand grain content, mass %	z <sub>20</sub>	1 – low from 70.540 to 73.279; 2 – medium from 73.280 to 76.019; 3 – high from 76.020 to 78.76
Grain-size composition of the clay sands, clay grain content, mass %	z <sub>21</sub>	1 – low from 7.120 to 9.150; 2 – medium from 8.160 to 11.199; 3 – high from 11.200 to 13.240
Grain-size composition of the clay loams, sand grain content, mass%	z <sub>22</sub>	1 – sandy clay loam over 40; 2 – pulverescent clay loam less than 50
Grain-size composition of the clay loams, pulverescent grains content, mass %	z <sub>23</sub>	1 – low from 72.310 to 75.589; 2 – medium from 77.489 to 75.590; 3 – high from 77.490 to 77.540
Grain-size composition of the clay loams, clay grains content, mass %	z <sub>24</sub>	1 – low from 18.400 to 18.455; 2 – medium from 18.456 to 20.510; 3 – high from 20.511 to 23.870
Grain-size composition of the clays, sand grain content, mass %	z <sub>25</sub>	1 – sandy clay over 40; 2 – pulverescent clay, less than 50
	z <sub>26</sub>	1 – low from 68.954 to 70.080; 2 – medium from 70.081 to 71.205; 3 – high from 71.343 to 72.329

Continuation table 1.

Characteristic feature	Code	Intervals of values
Grain-size composition of the clays, pulverescent grain content, mass %	$z_{27}$	1 – low from 23.871 to 24.895; 2 – medium from 25.896 to 27.920; 3 – high from 27.921 to 29.945

To the above mentioned characteristic features we add 4 compulsory features (zone, subzone, road district, supporting point). The compulsory features are applied to 3 zones only, since Western Siberian territory has been investigated partly. We pointed out 1 subzone and 1 road district in the 1st zone, 2 subzones and 7 road districts – in the 2nd zone and 2 subzones and 3 road districts – in the 3rd zone.

Illustrating example of matrices  $Q^e$ ,  $R^e$  and  $R'$  descriptions is given in Fig. 1. The matrices correspond to partial knowledge description. We use only a part of the characteristic features space and its values.

	$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	$z_6$	$z_7$	$z_8$	$z_9$		$k_1$	$k_2$	$k_3$			
$Q^e =$	1	5	1	3	3	3	3	4	4	$R^e =$	2	1	1	$R' =$	1	1
	1	4	1	4	3	3	3	4	4		2	1	1		1	2
	1	4	1	4	3	3	3	4	4		2	1	1		1	3
	1	4	1	3	3	3	3	5	4		2	1	2		2	4
	2	4	1	3	3	2	3	5	4		2	1	2		2	5
	1	4	1	3	3	2	3	5	4		2	1	2		2	6
	1	4	1	3	3	3	3	5	4		2	1	2		2	7
	1	4	1	4	3	3	3	5	4		2	1	2		2	8
	2	4	1	3	3	3	3	5	4		2	1	2		2	9
	2	4	1	3	3	3	3	5	4		2	1	2		2	10
	...	...	...	...	...	...	...	...	...		...	...	...		...	...
	3	4	1	3	3	3	3	5	3		3	1	3		8	24
	2	4	1	3	3	2	3	5	4		2	3	2		9	25
	2	4	1	4	3	3	3	5	3		2	3	2		9	26
	4	4	1	3	2	2	2	5	0		2	3	2		9	27
	3	4	2	3	2	2	3	5	3		2	3	2		9	28
	2	4	2	3	3	3	3	5	0		2	2	2		10	29
	4	4	2	3	3	2	3	5	3		3	2	1		11	30
	4	4	2	3	2	2	3	5	3		3	2	1		11	31
	4	4	2	0	3	3	3	5	0		3	2	1		11	32
2	4	2	3	3	3	3	5	2	2	2	1	12	33			

Fig. 1. Fragments of the extended matrices of description and distinguishing.

For the matrix model, filled in by the experts, the expert knowledge on the four zones, all the subzones and all the road districts are included. The fragments of description and distinguishing the matrices are represented in Fig. 2.

There are examples of usage of some visualization tools including cognitive graphics tools. The free-distributed open-street maps (OSM) [23] with information layer overlay for the presentation of common information are proposed. Information layer presents road regions with borders and some information about its. This information is a number of zone and subzone which are determined for road region. The proposed visualization tool is presented on Fig. 3.

In doing so note that for the mapping of decision-making results with usage of cognitive graphics tools we use 3-simplex for the zones representation and 2-simplex for subzones representation in case when the number of subzones equals 3 [24].

The information layer is denoted by number 1. It is a transparent layer over the map. The thin black lines separate the different road regions. The different color tones are used for labeling the different zones (red color tone is used for zone 2, blue color tone is used for zone 3). Each color of the road region in every zone is unique color gradation from zone base color given from color transformation in the hue-saturation-bright palette (HSB palette). The wide black lines are used to separate the different zones. Hatching over road region shows subzone type. Only 3 hatching types are used and only 2 types from them presented on Fig. 3. Description for all used colors and hatching is presented in the legend (see Fig. 3) and it is denoted by number 2. There is a list of all used hatchings and presented subzones in the upper part of the legend. The list of all used zones and correlated road regions is in the bottom part of the legend.

The information window for a road region is shown after click on a road region presentation. This window contains full information about a region. This information contains the road region name, 3-simplex and 2-simplex as information about proximity to specific zone (left 3-simplex) and subzone (right 2-simplex). The OUI (road region) is displayed as the circle with a big radius. Objects of learning sample are displayed as circles with smaller radiuses. The distance from the object OUI to an edge is directly-proportional to proximity of the object to the pattern corresponding to the edge. Distances of an OUI to edges are displayed as color lines. Color of an OUI (or objects from a learning sample) is mapped to the pattern which belongs to the concrete object. Mathematical foundations of the visualization of these objects with use of n-simplex are given in [25,26].

$Q =$									$R^e =$			$R =$	
$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	$z_6$	$z_7$	$z_8$	$z_9$	$k_1$	$k_2$	$k_3$		
2	4	2	4	3	3	2	5	4	2	1	1	1	1
2	4	2	4	3	3	2	5	4	2	1	1	1	2
2	4	2	4	3	3	2	5	4	2	1	1	1	3
2	4	2	4	3	3	2	5	4	2	1	2	2	4
2	4	2	4	3	3	2	5	4	2	1	2	2	5
2	4	2	4	3	3	2	5	4	2	1	2	2	6
2	4	2	4	3	3	2	5	4	2	1	2	2	7
2	4	2	4	3	3	2	5	4	2	1	2	2	8
2	4	2	4	3	3	2	5	4	2	1	2	2	9
2	4	2	4	3	3	2	5	4	2	1	2	2	10
1	4	1	0	3	3	3	4	5	2	1	2	2	...
1	4	1	0	3	3	3	4	5	2	1	2	2	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	4	4	2	2	2	2	6	2	3	1	3	8	207
4	4	4	2	2	2	2	6	2	2	3	2	9	208
4	4	4	2	2	2	2	6	2	2	3	2	9	209
4	4	4	2	2	2	2	6	2	2	3	2	9	210
4	4	4	2	2	2	2	6	2	2	3	2	9	211
4	4	4	2	2	2	2	6	2	2	2	2	10	212
4	4	4	2	2	2	2	6	2	3	2	1	11	213
4	4	4	2	2	2	2	6	2	3	2	1	11	214
4	4	4	2	2	2	2	6	2	3	2	1	11	215
4	4	4	2	2	2	2	6	2	2	2	1	12	216

Fig. 2. Fragments of the matrices of description and distinguishing, filled in by highly qualified experts.

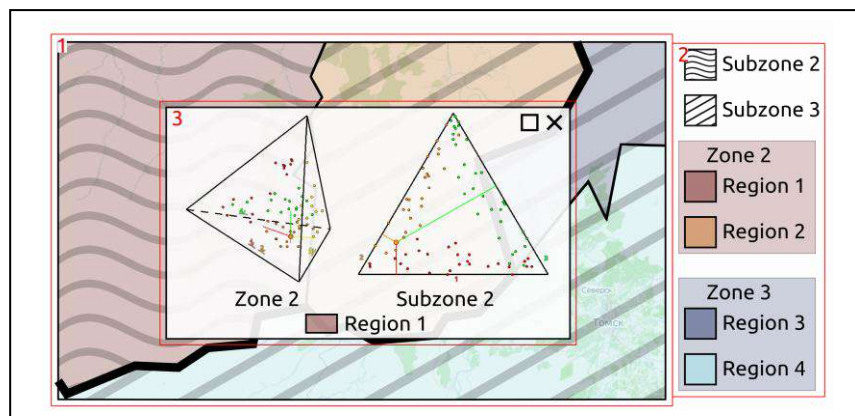


Fig. 3. Visualization tool for representation of the map with zoning results.

We revealed the different types of regularities on the basis of algorithms proposed by A. Yankovskaya and realized in IS RCZT. The revealed regularities allowed to reduce the features space from 27 to 11. That, in turn, has led to reduction of quantity of revealed feature values on 59 %.

We also verified the decisions-making using the generated supporting point descriptions proposed by A. Yankovskaya. The research results showed the IS RCZT development will lead to reduce significantly the expenditure and the cost of field and laboratory works on the territories under investigation. That, in its turn, will essentially reduce time expenses the specialists of road branch for the identification of zone, subzone and road district of the territory under investigation.

The proposed approach on road-climatic zoning of territories will allow to provide the required level of operational reliability of the highways.

## 4. Conclusion

The analysis of domestic and foreign standards of designing and building of highways is given. The advisability of creation intelligent systems road-climatic zoning of territories is substantiated.

For the first time we proposed the complex matrix model of data and knowledge representation for road-climatic zoning. It has allowed to carry out structurization of the data and knowledge on the road-climatic zoning. Complex matrix model is represented by the 3 matrix models: the extended matrix model that includes the extended matrix of description and the matrix of distinguishing; the matrix of knowledge description and the matrix of distinguishing filled with highly qualified experts; the partial matrix model consisting of the extended matrix of description of the territories under study.

It is created a prototype of the intelligent system of road-climatic zoning. On the basis of extended matrix representation it is created data and knowledge base using the research results on natural and climatic conditions of Western Siberian regions. The base of data and knowledge was created highly qualified experts.

For the 1<sup>st</sup> and the 2<sup>nd</sup> matrix representations of data and knowledge are revealed and is eliminated at a finding of intersections of objects descriptions from different patterns.

Results of a research prototype approbation of road-climatic zoning of the territories intelligent system have shown as follows: reduction on 59 % of necessary number of characteristic features for decision-making on reference of territory part under study to this or that zone, a subzone and road district. Application IS RCZT will decrease significantly the expenditure and cost on the field and laboratory research of the territories under study. That will also save the reduce time expenses of the road branch specialists.

The proposed approach on road-climatic zoning of the territories will allow to provide demanded level of operational reliability of again under construction and reconstructed highways and first of all in regions with the poorly developed network of highways.

## Acknowledgements

The research was funded by RFBR grant (project No. 14-07-00673a and No. 16-07-0859a). The authors are grateful to V. Efimenko, Doctor of Science; S. Efimenko, Doctor of Science; M. Badina, Candidate of Science for the information on road-climatic zoning of West Siberian regions; to V. Churilin, senior lecturer for the data and knowledge base fulfilling; to R. Ametov, Deputy director of the Information Technologies Center of the Tomsk State University of Architecture and Building and A. Yamshanov, Junior research fellow, assistant of the Tomsk State University of Control Systems and Radioelectronics for the IIS IMSLOG development and for revealing the regularities within the developed data and knowledge base on road climatic zoning; to S. Kitler, the executor of the project No 16-07-00859a for experiments conducting.

## References

- [1] GavriloVA TA, Kudryavcev DV, Muromcev DI. Knowledge Engineering. Models and methods. St.P.: Publishing company "Lan", 2016; 324 p. (in Russian)
- [2] Highways: SP 34.13330.2012. M.: Ministerstvo regional'nogo razvitiya RF, 2013; 106 p. (in Russian)
- [3] Design of flexible pavement: ODN 218.046-01. M.: Informavtodor, 2001; 145 p. (in Russian)
- [4] Code of Practice for Highway Routes of the People's Republic of China: JTG D20-2006. People's Communications Press, 2006.
- [5] Chao Li, Yu-lan Wang, Jin-liang Xu. Research on Geographic Information System of Natural Zoning for Highway. Applied Mechanics and Materials 2013; 353-356: 3502-3506.
- [6] "Filing system" of physiographic units helps to resolve local design criteria. Highway Res. News 1973; 51: 42-60.
- [7] Richtlinien für die Standartisierung des Oberbaues von Verkehrsflächen: RStO 01. Köln.: FGSV-Verlag, 2001.
- [8] Groney D. The design and performance of road pavements. London: Transport and road research laboratory, 1977; 673 p.
- [9] Highways: SNiP RK 3.03-09-2006. Astana: Proektnaya akademiya "KAZGOR", 2014; 51 p. (in Russian)
- [10] Highways. Flexible pavement. Design rules: TKP 45-3.03-112-2008. Minsk.: Ministroiarhitekturi, 2009; 86 p. (in Russian)
- [11] Design. Highways: SNiP RK 32-01:2004. Bishkek: Goskomarhstroi pri Pravitel'stve Kirizskoi Respubliki, 2004; 85 p. (in Russian)
- [12] Efimenko SV, Efimenko VN, Afinogenov AO. The Outline of Road Building Climatic Zoning in Western Siberia. Vestnik TSUAB 2013; 4(1-3): 78-84.
- [13] Efimenko SV, Badina MV. Road zoning of Western Siberia: monography. Tomsk: Publishing of TSUAB 2014; 244 p. (in Russian)
- [14] Ushakov VV, Efimenko VN, Vishnevskiy AV. Road-climatic zoning of highway "Amur" Chita – Khabarovsk under the terms of the construction and operation. Highways 2007; 5: 77-79. (in Russian)
- [15] Yarmolinskiy VA. Khabarovsk territory zoning in snow cleaning of roads. Vestnik TSUAB 2014; 5: 152-158. (in Russian)
- [16] Yankovskaya AE, Gedike AI, Ametov RV, Bleikher AM. IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition. Pattern Recognition and Image Analysis 2003; 13(2): 243-246.
- [17] Yankovskaya AE, Gedike AI, Ametov RV. Construction of applied intelligent systems on the base of software tool IMSLOG-2002. Vestnik TSU. Application 2002; 1(II): 185-190. (in Russian)
- [18] Yankovskaya AE, Efimenko VN, Efimenko SV, Cherepanov DN. Application of matrix models for creation of intelligent information technology in sphere of the state and municipal management. Fuzzy systems and soft computing: Proceedings of 6<sup>th</sup> Russian science-practical conference. St.P.: Politehnika-servis 2014; 2: 118-127. (in Russian)
- [19] Yankovskaya A, Cherepanov D, Selivanikova O. Data and Knowledge Base on the Basis of the Expanded Matrix Model of Their Representation for the Intelligent System of Road-Climatic Zoning of Territories. IOP Conf. Series: Materials Science and Engineering 2016; 142: 012041.
- [20] Efimenko VN, Efimenko SV, Sukhorukov AV. Accounting for natural-climatic conditions in the design of roads in Western Siberia. Sciences in Cold and Arid Regions 2015; 7(4): 307-315.
- [21] Efimenko SV. Territorial homogeneity of geographic complexes in design of automobile roads. Vestnik TSUAB 2015; 3: 226-236. (in Russian)
- [22] Yankovskaya AE, Ametov RV. Architecture of intelligent system oriented on road-climatic zoning of territories. Proceedings of the Congress on intelligent systems and information technologies. Taganrog: Publishing UFU 2016; 1: 98-104. (in Russian)

- [23] Yankovskaya A, Yamshanov A. Bases of intelligent system creation of decision making support on road-climatic zoning. Pattern Recognition and Information Processing (PRIP'2014): Proceedings of the 12<sup>th</sup> International Conference. Minsk: UIIP NASB 2014: 311–315.
- [24] Yankovskaya A, Yamshanov A. Family of 2-simplex cognitive tools and their application for decision-making and its justifications. Computer Science & Information Technology (CS & IT) 2016; 6(1): 63–76.
- [25] Yankovskaya A, Krivdyuk N. Cognitive Graphics Tool Based on 3-Simplex for Decision-Making and Substantiation of Decisions in Intelligent System. Proceedings of the IASTED International Conference Technology for Education and Learning (TEL 2013). Marina del Rey, USA, 2013: 463–469.
- [26] Yankovskaya AE, Yamshanov AV, Krivdyuk NM. Application of Cognitive Graphics Tools in Intelligent Systems. IJEIT 2014; 3(7): 58–65.

**Table of Contents**  
High-Performance Computing

1 Parallel calculations in the construction of the kinetic model of benzylidene benzylamine synthesis I.V. Akhmetov, I.M. Gubaydullin.....	1-4
DOI: 10.18287/1613-0073-2017-1902-1-4	
2. Modeling and Simulation of the interaction between oil and rotating gear within Final drive volume E. Avdeev, V. Ovchinnikov.....	5-9
DOI: 10.18287/1613-0073-2017-1902-5-9	
3. Numerical modeling of the labyrinth seal taking into account vibrations of the gas transmittal unit rotor in aeroelastic formulation L.N. Butymova, V.Ya. Modorskii.....	10-17
DOI: 10.18287/1613-0073-2017-1902-10-17	
4. High-performance DTW-based signals comparison for the brain electroencephalograms analysis A.I. Makarova, V.V. Sulimova.....	18-24
DOI: 10.18287/1613-0073-2017-1902-18-24	
5. Software for heterogeneous computer systems and structures of data processing systems with increased performance A.A. Kolpakov, Ju.A. Kropotov.....	25-31
DOI: 10.18287/1613-0073-2017-1902-25-31	
6 Advanced mixing audio streams for heterogeneous computer systems in telecommunications A.A. Kolpakov, Ju.A. Kropotov.....	32-36
DOI: 10.18287/1613-0073-2017-1902-32-36	
7. The algorithm for a video panorama construction and its software implementation using CUDA technology I.A. Kudinov, O.V. Pavlov, I.S. Kholopov, M.Yu. Khramov.....	37-42
DOI: 10.18287/1613-0073-2017-1902-37-42	
8. Various algorithms, calculating distances of DNA sequences, and some computational recommendations for use such algorithms B.F. Melnikov, S.V. Pivneva, M.A.Trifonov.....	43-50
DOI: 10.18287/1613-0073-2017-1902-43-50	
9. Development of parallel implementation of the informative areas generation method in the spatial spectrum domain N. Kravtsova, R. Paringer, A. Kupriyanov.....	51-54
DOI: 10.18287/1613-0073-2017-1902-51-54	
10. Numerical simulation of motion of dust particles in an accelerator path A.V. Piyakov, D.V. Rodin, M.A. Rodina, A.M. Telegin.....	55-61
DOI: 10.18287/1613-0073-2017-1902-55-61	
11. Performance Analysis of a Simple Runtime System for Actor Programming in C++ S.V. Vostokin, E.G. Skoryupina.....	62-67
DOI: 10.18287/1613-0073-2017-1902-62-67	
12. Application of the pyramid method in diffrence solution D'Alembert equations of graphic processor with the use Matlab L.V. Yablokova, D.L. Golovashkin.....	68-70
DOI: 10.18287/1613-0073-2017-1902-68-70	
13. S.B. Popov, Doctor of Engineering (Commemorating the 60 <sup>th</sup> Birth Anniversary) V.O. Sokolov.....	71-75
DOI: 10.18287/1613-0073-2017-1902-71-75	

# Preface

Vladimir Fursov<sup>1</sup>, Yegor Goshin<sup>1</sup>, Vladimir Voevodin<sup>2</sup>, Dmitry Savelyev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Lomonosov Moscow State University, 2nd Education Building, Faculty CMC, GSP-1, Leninskie Gory, 119991, Moscow, Russia

---

Session «High-Performance Computing» was held at the 3rd International Conference on Information Technology and Nanotechnology - 2017 (ITNT-2017) in Samara, Russia, April 25–27, 2017 (<http://ru.itnt-conf.org/itnt17ru/>). At this session relevant studies of development and implementation of high-performance computing algorithms were presented and discussed.

The goal of the ITNT-2017 Conference was to discuss problems of fundamental and applied research in information technology and nanotechnology, including but not limited to:

- Computer Optics;
- Diffractive Nanophotonics;
- Image Processing;
- High-performance Computing;
- Computer Vision;
- Mathematical Modeling;
- Data Science.

Scientists from Austria, Belarus, Bulgaria, Denmark, Germany, Great Britain, India, Iraq, Mexico, Moldova, Russia, Spain, USA, and Finland presented over 330 reports at the ITNT-2017 Conference.

The most significant studies presented at the Conference will be published in *Procedia Engineering* (Elsevier BV). This volume contains all the papers presented at «High-Performance Computing» session, not included in *Procedia Engineering*.

We are grateful to everybody who has contributed to the session and look forward to meeting you again at future events. Further, we thank all the authors who presented their studies, as well as the reviewers and the delegates. Moreover, we sincerely thank the team of organizers for making the session successful and this publication possible.

## Guest Editors

- Professor Vladimir Voevodin, MSU Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
- Professor Vladimir Fursov, Samara National Research University, Samara, Russia;
- Yegor Goshin, Samara National Research University, Samara, Russia;
- Denis Kudryashov, Samara National Research University, Samara, Russia.

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov – Samara National Research University, Russia

## Vice-chairs

- Vladimir Bogatyrev – Samara National Research University, Russia
- Nikolay Kazanskiy – Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Russia
- Eduard Kolomiets – Samara National Research University, Russia
- Alexander Kupriyanov – Samara National Research University, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Russia



# Parallel calculations in the construction of the kinetic model of benzylidene benzylamine synthesis

I.V. Akhmetov<sup>1</sup>, I.M. Gubaydullin<sup>1,2</sup>

<sup>1</sup>Ufa State Technological Petroleum University, Kosmonavtov street 1, 450062, Ufa, Russia

<sup>2</sup>Institute of Petrochemistry and Catalysis Russian Science Academy, Prospect Oktyabrya 141, 450075, Ufa, Russia

---

## Abstract

In this paper, a kinetic model of the benzylidene benzylamine synthesis reaction has been developed. The optimal rate constants for the stages and activation energies are found. When searching for kinetic parameters, the OpenMP parallel computing technology was used. An effective number of flows are determined, in which the solution of inverse problems is most effective.

*Keywords:* kinetic model; rate constants; inverse kinetic problem; parallel calculations; OpenMP

---

## 1. Introduction

The catalytic reaction of the synthesis of the N-benzylidene benzylamine aromatic compound has a wide range of applications. N-benzylidene benzylamine is known as an indicator for the quantitative determination of organolithium compounds by the titrimetric method and is the starting compound for the synthesis of a number of heterocycles [1, 2].

To study the mechanism of the synthesis reaction, it is necessary to construct a kinetic model, the solution of the inverse kinetic problems for which is complicated, because it is the most difficult and time consuming stage of kinetic model development.

Usage of parallel calculations is becoming more and more popular as a method of mathematical processing of experimental data because of the increasing complexity of obtained detailed information on chemical reactions.

The inverse problems of chemical kinetics refer to such physico-chemical problems that involve a significant amount of computations [3]. High-performance computing systems usage fundamentally changed the possibilities of complex chemical processes analysis: a detailed analysis of complex kinetic models with a large amount of experimental information has become available; The time for kinetic models construction has been reduced in many times; The accuracy of decisions has increased.

At present, solutions of inverse kinetic problems are proposed with the use of parallel calculations on cluster systems and graphics processors. Computer systems with multicore processors are actively introduced, the advantages of which are availability and ease of use, which expands the possibilities of their application in scientific researches [4, 5].

In this paper we propose a method for kinetic parameters searching using parallel calculations technology on multi-core systems for kinetic models construction of chemical reactions on metal complex catalysis with the aim of studying time reducing and mastering new catalytic processes.

## 2. Kinetic model

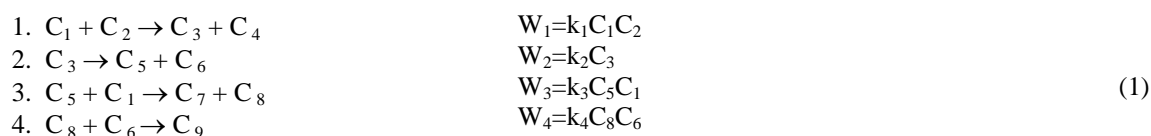
To understand the physical and chemical nature of the catalytic reaction, the subsequent mathematical modeling of the catalytic process and the definition of the conditions for its industrial realization, it is necessary first of all to develop its kinetic and mathematical models [6].

The fundamental basis for catalytic processes modeling is, first of all, the detailed studies of the physical and chemical nature of chemical reactions, since the quantitative characteristics obtained in the practical experiment and refined in numerical experiments will allow to develop kinetic models that will become a reliable basis for subsequent research.

The kinetic model of the process is a set of elementary stages, reactions and equations characterizing the dependence of the rate of chemical transformation on the reaction parameters: pressure, temperature, reagent concentrations, etc. Such dependencies are determined on the basis of experimental data obtained in the practical experiment while changing reaction parameters at the range of industrial conditions. The model developed in this way is the first level of the catalytic reactor model and the basis for later solving static and dynamic problems which arise in the development of technological processes.

The development of kinetic models, which is given in this paper, is based on experimental data of the benzylidene benzylamine synthesis which was obtained in the laboratory of hydrocarbon chemistry in IPC RAS. During the experiments, a new original method to include carboxyl group in pyrrols compounds, based on interaction between pyrrols and  $\text{CCl}_4\text{.CH}_3\text{OH}$ -catalyst system.

Based on the analysis of the experimental data and the results of tit mathematical processing, the following scheme of chemical transformations and the corresponding kinetic equations (1)-(2) are proposed:



$C_i$  – concentration of the components, mol/l:  $C_1=C_7H_9N$  – benzylamine,  $C_2=CCl_4$  – carbon tetrachloride,  $C_3=C_7H_8NCl$  – chlorobenzylamine,  $C_4=CHCl_3$  – chloroform,  $C_5=C_7H_7N$  – 1- phenylmethaneimine,  $C_6=HCl$  – hydrochloric acid,  $C_7=C_{14}H_{13}N$  – benzilidenbenzilamin,  $C_8=NH_3$  – ammonia,  $C_9=NH_4Cl$  – ammonium chloride;  $k_j$  – kinetic rate constant of  $j$ -th reaction,  $l \cdot mol^{-1} \cdot h^{-1}$  ( $j=1, 3, 4$ ),  $h^{-1}$  ( $j=2$ ),  $W_j$  –  $j$ -th rate of reaction,  $mol/(l \cdot h)$ .

The kinetic equations of the transformation scheme (1) are analyzed in accordance to the law mass action. The correct description of the laboratory reactor with a stirrer is the ideal mixing model:

$$\frac{d\bar{N}}{dt} = F_N, \quad F_N = \frac{1}{V_o} \sum_{j=1}^J \delta_j \omega_j, \quad \delta_j = \sum_{i=1}^I v_{ij} \quad (2)$$

$$\frac{dX_i}{dt} = \frac{F_i - X_i F_N}{\bar{N}} \quad (3)$$

with the initial conditions:  $t = 0, X_i = X_i^o, \bar{N} = 1$ , where  $\bar{N} = C/C_o$  – relative change in the number of moles of the reaction mixture;  $C$  и  $C_o$  – the molar density and its initial value, mol/l;  $X_i=C_i/C$  – concentration of the components, the mole fractions;  $V_o$  – the volume of the reaction space, l;  $\omega_j=W_j/C_o$  – given reaction rate,  $h^{-1}$ ;  $J$  – number of stages of chemical transformation;  $I$  – the number of components.

The right sides of the systems of equations (2)-(3) have the following form:

$$F_1 = -\omega_1 - \omega_3; F_2 = -\omega_1; F_3 = \omega_1 - \omega_2; F_4 = \omega_1; F_5 = \omega_2 - \omega_3; F_6 = \omega_2 - \omega_4; F_7 = \omega_1; F_8 = \omega_3 - \omega_4; F_9 = \omega_4; F_n = \omega_2 - \omega_4.$$

### 3. Usage of parallel calculations

For the parallel solution of the inverse kinetic problem the most effective method is genetic algorithm which is based on the idea of breeding, borrowed from biology, that is, the preferential multiplication of the fittest individuals [7]. The practical application of the genetic algorithm in all known cases led to positive results [8]. It is shown that the genetic algorithm, unlike the gradient methods of minimization, is a universal method for searching for an optimum regardless of the complexity of the functions [8]. The sequence of operations that form the basis of the genetic is described below.

At the first step of the algorithm, an initial population is created randomly, consisting of  $N$  individuals ( $N$  points in the space of kinetic parameters, each point has  $m$  coordinates – parameter values). At the stage of mutation, the individuals change in accordance with a predetermined mutation operation, in which the coordinate/parabolic descent from the points of space was taken. At the stage of selection, a certain proportion of the whole population is selected, which will remain "alive" at this stage of evolution. The probability of survival of an individual depends on the value of the fittest function for this individual; as a function of the fittest  $s$  is the residual functional. The proportion of surviving  $s$  is a parameter of the genetic algorithm, and according to the results of selection from  $N$  individuals of the population, the total population will include  $sN$  individuals. In the case under consideration,  $s = 1/2$ . When forming a new generation, a crossing is used – to produce a descendant, two parents are needed. To form a new point in the parameter space, one point from the "survivors" and one of the "dying" are selected as parents, and the crossing is done by choosing  $m/2$  coordinates from the first point and the remaining ones from the second point; while the descendant inherits the traits of both parents. Specimens for reproduction are selected from the entire population, and not from surviving elements at the first step in order to exclude the possibility of population degradation. This set of actions is repeated iteratively, so the "evolutionary process", which lasts for several life cycles (generations), is modeled until the criterion for stopping the algorithm is fulfilled, which is any of the conditions:

- 1) finding a global or suboptimal solution;
- 2) the exhaustion of the number of generations released for evolution;
- 3) the exhaustion of the time allowed for evolution.

Parallelization of the calculation process takes place at the stage of initial filling, when the given pseudo-random points in the parameter space are uniformly distributed over the flows. Each mutation is mutated independently; The data is exchanged at the selection stage. At the same time, the autonomous work time of processes significantly exceeds the time of internuclear interactions, which determines the effectiveness of this algorithm (Fig. 1).

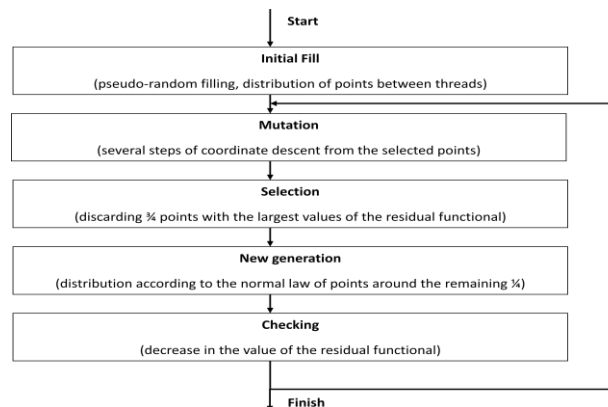


Fig. 1. Genetic algorithm.

#### 4. Results of computational experiments

The numerical values of the found rate constants of the stages and activation energies for the synthesis of benzylidene benzylamine are presented in Table 1.

Table 1. Kinetic parameters for synthesis of benzilidenbenzilamina.

Kinetic constants at 23°C, h <sup>-1</sup>		E <sub>i</sub> , kcal/mol
k <sub>1</sub>	1.5×10 <sup>-2</sup>	10.6
k <sub>2</sub>	4.7	7.7
k <sub>3</sub>	13.4	1.6
k <sub>4</sub>	0.6	0.4

Estimating the efficiency of parallelization when testing a program on a computational cluster showed us that the parallel program works efficiently with increasing number of threads. When solving the inverse kinetic problem for the benzylidene benzylamine synthesis reaction from all experimental data, the total calculation time was 60 hours at 1 core, 3.5 hours using 18 fluxes (Fig. 2 and Fig. 3).

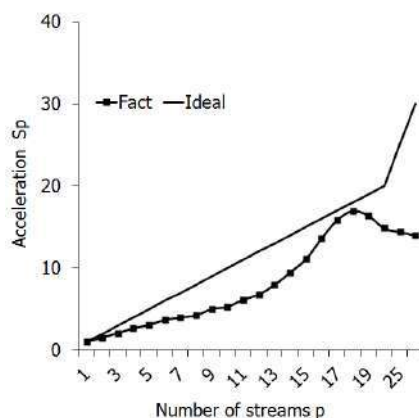


Fig. 2. Efficiency.

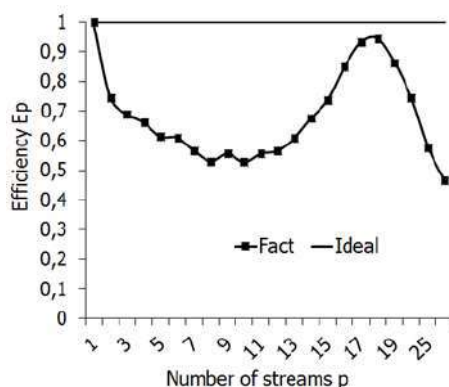


Fig. 3. Accelerate.

The adequacy of the constructed model with usage of parallel calculation of the display by comparing the calculated and experimental data on the yield of the desired product, benzylidene benzylamine (Fig. 4).

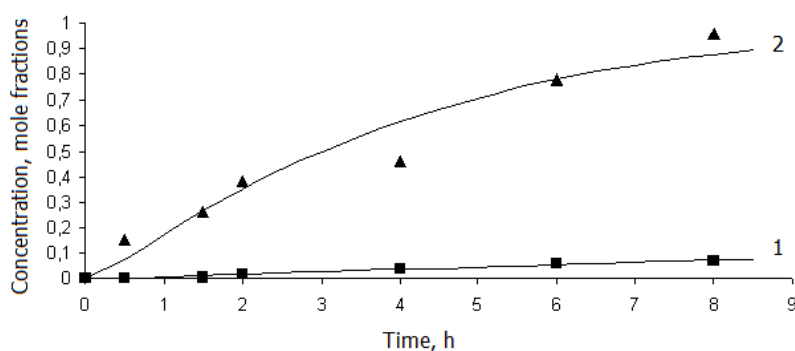


Fig. 4. Comparison of the calculated and experimental temperatures of 85 C°(2) and 23 C°(1).

## 5. Conclusion

Thus, an algorithm has been developed for using multi-core computing systems to solve the inverse problems of chemical kinetics. The method is implemented as a software package that includes a database of kinetic studies, sequential and parallel algorithms for solving systems of ordinary differential equations, implemented on single-core and multi-core computing systems.

The informational and analytical systems have been developed, the successful application of which in the development of kinetic models of reactions of aromatic and heterocyclic compounds synthesis of has shown the universality of the proposed system approach to solving inverse kinetic problems.

The system allows users to easily adapt to the development of kinetic models of various reactions due to the formation of new blocks in the database of experimental data, the selection or addition of new methods of data processing, the construction of mathematical models of the objects of varying complexity.

Based on the developed approach, a kinetic model for the benzylidene benzylamine synthesis reaction was constructed. It is shown that with the use of the information system, the computing process can be accelerated approximately in 18 times.

## References

- [1] Khusnutdinov RI, Bayguzin AR, Aminov RI. Synthesis of N-benzylamine benzilidenbenzilamina under the action of iron catalysts in CCl<sub>4</sub>. *Journal of Organic Chemistry* 2012; 48 (8): 1063– 1065.
- [2] Khusnutdinov RI, Baiguzina AR, Mukminov RR, Akhmetov IV, Gubaidullin IM, Spivak SI, Dzhemilev UM. New synthesis of pyrrole-2-carboxylic and pyrrole-2,5-dicarboxylic acid esters in the presence of iron-containing catalysts. *Russian Journal of Organic Chemistry* 2010; 46(7): 1053– 1059.
- [3] Akhmetov IV, Gubaydullin IM. Analysis of methods for solving inverse problems of chemical kinetics with the use of parallel computing. PCT 2016 - Proceedings of the 10th Annual International Scientific Conference on Parallel Computing Technologies. *CEUR Workshop Proceedings* 2016; 402– 410.
- [4] Akhmetov IV. Multinuclearity in inverse kinetic problems. Scientific service in the Internet: search for new solutions Proceedings of the International Supercomputer Conference 2012; 656– 661.
- [5] Akhmetov IV. Development of kinetic models for reactions of synthesis of aromatic and heterocyclic compounds based on multicore computing systems. *Parallel Computing Technologies* 2013. Proceedings of the International Scientific Conference 2013; 268– 277.
- [6] Slinko MG. Modeling of chemical reactors. Novosibirsk: Science, 1968; 96 p.
- [7] Nikitin AV, Nikitina LI. Evolutionary model of optimization of modular associative memory for data flow machines based on genetic algorithm. *Programming* 2002; 6: 31– 42.
- [8] Chernyshev O, Borisov A. Comparative analysis of solving optimization problems by genetic and gradient methods. *Transport and Telecommunication* 2007; 8(1): 40–52.

# Modeling and Simulation of the interaction between oil and rotating gear within Final drive volume

E. Avdeev<sup>1,2</sup>, V. Ovchinnikov<sup>2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Laduga Automotive Engineering, 71 Mozhayskoe shosse, 143000, Odintsovo, Russia

---

## Abstract

An adaptive mesh refinement (AMR) method based on discretization matrix metric is described. The computational algorithm is implemented using OpenFOAM parallel library. This open C++ library provides data structures and routines to work with the finite volume method and adaptive mesh. The method was used for oil flow in final drive simulation. For more efficient use of computing resources, we decided to use an approach based on the use of adaptive mesh refinement/coarsening. Adaptive mesh refinement approach showed greater efficiency in cases of low rotational frequencies and less efficiency in case of high frequency.

*Keywords:* mesh adaptation; lubrication modeling; final drive modeling; bearing modeling

---

## 1. Introduction

In this paper we consider a design of automobile final drive. One of the problems arising during the automobile final drive design is the lubricity problem. In particular, the authors of this paper solved the problem of the oil flow simulation created by rotating gear wheel of final drive. Then, the calculation results are transferred to design engineer, who will correct the shape of the final drive body accordingly that the oil flow will reach the stuffing box (see Fig.1).

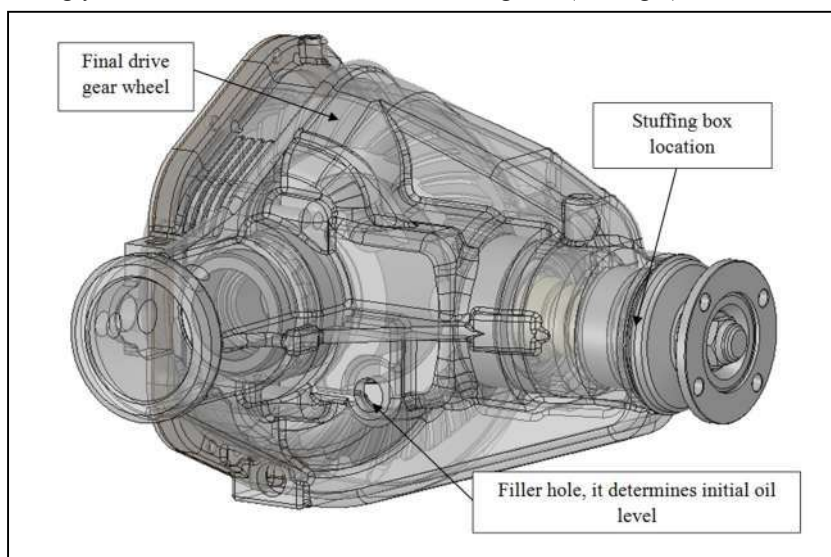


Fig. 1. The original geometry and the basic elements of the final drive.

To simulate oil flow we decided to use a two-phase liquid-air model without taking into account the compressibility, heat transfer and miscibility. For phase separation we use VOF method, such as Lemfeld [1], Chunfeng [2]. Oil flow simulation is quite complex and requires a lot of computational time. For more efficient use of computing resources, we decided to use an approach based on the use of adaptive mesh refinement/coarsening (Adaptive Mesh Refinement - AMR). AMR procedure implemented in the OpenFOAM library [3], but the original library does not support AMR for rotating mesh. In this paper, we conducted the OpenFOAM library modification for final drive lubricity modeling.

## 2. Mesh adaptation method

Currently, the mesh adaptation technologies are widely used in numerical problems solving. There is a large amount of literature which deals with dynamic mesh and mesh adaptation methods. One of the first works on dynamic mesh application were investigations of Miller [4] and Yanenko[5]. Mesh adaptation methods usually based on minimization of some selected functional. It is achieved by refinement or coarsening of mesh elements (h-adaptation) or mesh nodes moving (p-adaptation).

Mesh adaptation allows to reduce computational cost, to correct mesh in more complex areas, to handle moving surfaces, phase transitions and other areas of high gradients. Mesh adaptation approaches was successfully implemented in many commercial and non-commercial software packages such as Star-CCM+, FlowVision, Abaqus, Ansys, OpenFOAM. In this study we used an OpenFOAM open library, which has complete modules for AMR implementation.

For AMR configuration in OpenFOAM user need to provide following information:

- mesh update frequency (update mesh on every first, second or subsequent iteration);
- scalar field, whose values will be used for the mesh refinement/coarsening;
- field values interval, defined by minimum and maximum values, at which we want to refine mesh;
- field threshold value, below which we want to start mesh coarsening;
- maximum cells refinement level relative to initial mesh cells;
- the maximum allowable mesh cells amount.

In this work as scalar field we use field, based on discretization matrix eigenvalues estimation. This method described in more detail in [6].

Adaptive mesh refinement algorithm includes following steps:

1. Discretization matrix  $\mathbf{A}$  initialization.
2. Matrix  $\mathbf{M} = \mathbf{I} - \mathbf{A}$  calculation.
3. Eigenvalues estimation matrix calculation.

$$\mathbf{F}_i = |m_{ii}| + \sum_{i \neq j} |m_{ij}|,$$

where  $m_{ii}$  and  $m_{ij}$  diagonal and off-diagonal elements of matrix  $\mathbf{M}$ .

4. Mesh cells refinement/coarsening, based on matrix  $\mathbf{F}$ .

Current version of OpenFOAM-v4.1 does not allow to use mesh adaptation (implemented by dynamicRefineFvMesh class) and rotation of the mesh (implemented by solidBodyMotionFvMesh class) simultaneously. Therefore, to achieve the desired functionality, we have created a new C++ class solidBodyMotion dynamicRefineFvMesh by virtual inheritance. The sources available at [7].

### 3. Discussed Problems

In this section we briefly consider the mathematical model, which describes oil distribution during final drive gear wheel rotation.

Oil distribution is described by the following equations [8]:

$$\frac{\partial \alpha_\varphi \overline{U}_\varphi}{\partial t} + \nabla \cdot (\alpha_\varphi \overline{U}_\varphi \overline{U}_\varphi) + \nabla \cdot (\alpha_\varphi \overline{R}_\varphi^{eff}) = -\frac{\alpha_\varphi}{\rho_\varphi} \nabla \overline{p} + \alpha_\varphi \mathbf{g} + \frac{\overline{M}_\varphi}{\rho_\varphi} \quad (1)$$

$$\frac{\partial \alpha_\varphi}{\partial t} + \nabla \cdot (\overline{U}_\varphi \alpha_\varphi) = 0 \quad (2)$$

where  $\varphi$  – phase,  $\alpha$  – phase fraction,  $\overline{R}_\varphi^{eff}$  is combined Reynolds (turbulent) and viscous stress,  $\overline{M}_\varphi$  – averaged inter-phase momentum transfer term,  $\overline{U}_\varphi$  – averaged transport velocity,  $p$  – pressure,  $t$  – time discretization step size,  $\mathbf{g}$  – acceleration due to gravity,  $\rho_\varphi$  – phase density.

Combining equation (2) for two phases with  $\varphi = a$  and  $b$  yields the volumetric continuity equation for the mixture, which will be utilized to formulate an implicit equation for the pressure. The volumetric continuity equation reads:

$$\nabla \cdot \overline{U} = 0 \quad (3)$$

where  $\overline{U} = a_\alpha \overline{U}_\alpha + a_b \overline{U}_b$ .

The averaged equations representing the conservation of mass and momentum for each phase.

After the technical requirements analysis we decide to perform lubrication modeling for the following selected shaft rotational frequencies: 551, 800, 1600, 2400 rpm. These frequencies set describes final drive basic operating modes.

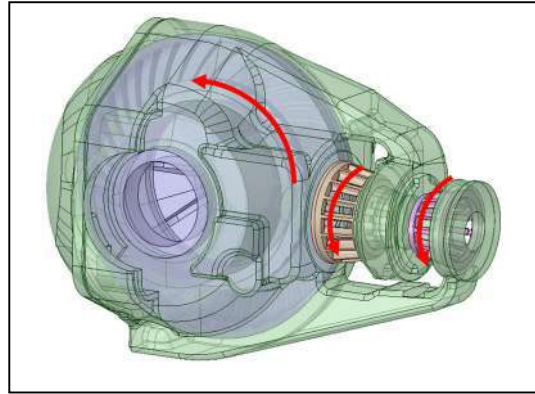


Fig. 2. The final drive internal volume, gear wheel and bearings rotation directions.

#### 4. OpenFOAM library parallelism

The method of parallel computing used in OpenFOAM is based on the computational domain mesh and fields decomposition into separate parts, every single part is assigned to a separate computing core. Algorithms parallelization is built-in OpenFOAM parallel library. Thus, the parallel calculation process includes the following steps: mesh and fields decomposition; parallel solver run; postprocessing after mesh and fields reconstruction or right in the decomposed form. When post-processing cases that have been run in parallel the user has two options:

- reconstruction of the mesh and field data to recreate the complete domain and fields, which can be post-processed as normal;
- post-processing each segment of decomposed domain individually.

A decomposed OpenFOAM case is run in parallel using the openMPI implementation of MPI. openMPI can be run on a local multiprocessor machine very simply but when running on machines across a network, a file must be created that contains the host names of the machines. The file can be given any name and located at any path. OpenFOAM also allows to use other MPI implementation libraries.

All computations are performed on cluster “Sergey Korolev”. In particular, we used two blade server types:

- HS22 blade servers, each of them has 2x CPU: Intel Xeon X5560, 4 cores;
- HS23 blade servers, each of them has 2x CPU: Intel Xeon E5-2665, 8 cores.

Execution time comparison for this server types showed on Figure 3.

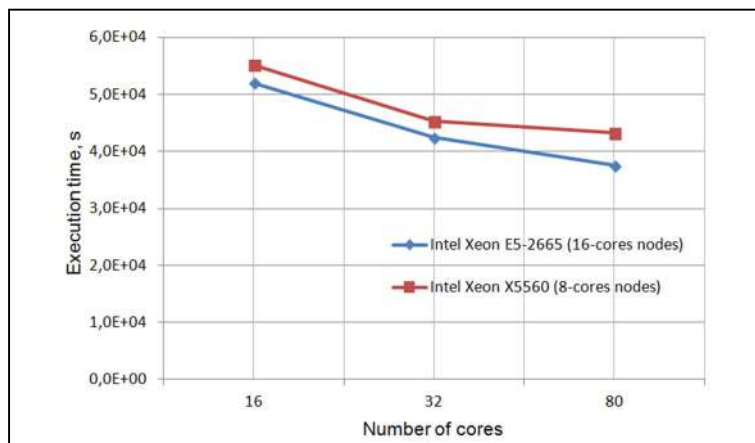


Fig. 3. Execution time for different number of cores and nodes types.

#### 5. Numerical simulation results

Figure 4 shows the oil-air free surface for wheel rotational frequency 2400 rpm, time = 0.278 s.

Figure 5 shows the oil distribution for wheel rotational frequency 551 rev/min, time  $t = 1.7$  s. It can be seen that in this case the oil flow reaches the stuffing box location.

Adaptive mesh refinement more effective in areas of constant oil flow form, less effective in areas with stochastic oil flow behavior.

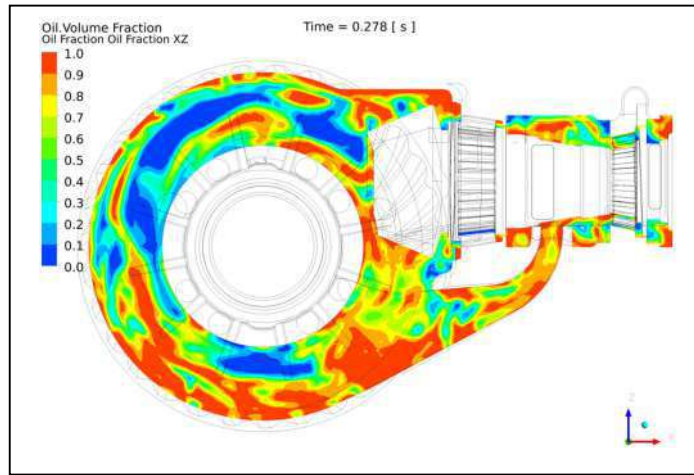


Fig. 4. Oil- distribution for wheel rotational frequency 2400 rpm, time  $t = 0.278$  s.

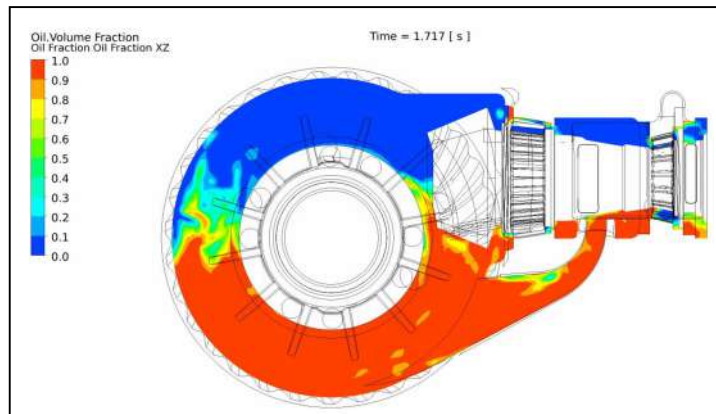


Fig. 5. Oil distribution for wheel rotational frequency 551 rev/min, time  $t = 1.7$  s.

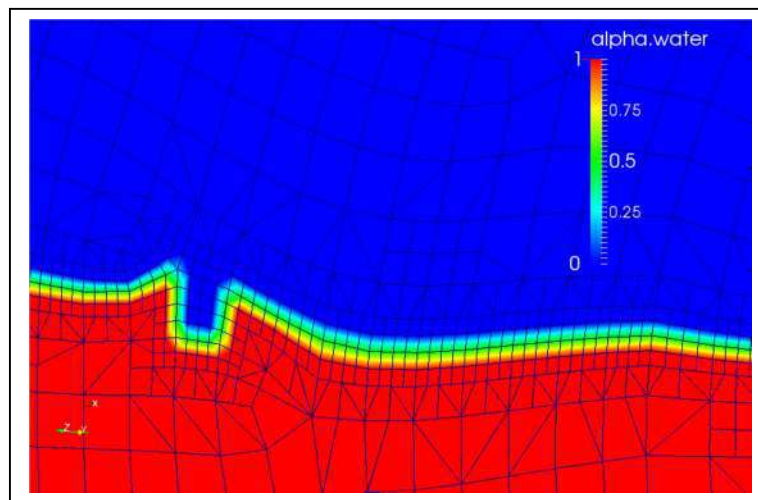


Fig. 6. Mesh fragment, wheel rotational frequency 551 rev/min, time  $t = 1e-6$  s.

Figure 6 shows mesh fragment for case of wheel rotational frequency 551 rev/min, time  $t = 1e-6$  s. More fine mesh formed in areas with a higher  $\alpha$  phase fraction gradient, which reduces task computational cost.

## 6. Conclusion

Algorithms parallelization performed by built-in features of OpenFOAM parallel library. An implemented by OpenFOAM library model has shown efficiency and stability. Adaptive mesh refinement along with the ability to use parallel computing also provides computational costs reduction compared to the static mesh.

Adaptive mesh refinement approach showed greater efficiency in cases of low rotational frequencies and less efficiency in case of high frequency. Computational complexity of the problem should be taken into account when deciding on the use of AMR. The additional costs of calculating the mesh quality metric field and mesh updating decrease efficiency of AMR using.



## Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation.

## References

- [1] Lemfeld F, Fran K, Unger J. Numerical simulations of unsteady oil flows in the gearboxes. *Journal of applied science in the thermodynamics and fluid mechanics* 2007; 1: 1–5.
- [2] Vande V, Vierendeels J, Dick E. Flow simulations in rotary volumetric pumps and compressors with the fictitious domain method. *Journal of Computational and Applied Mathematics* 2004; 168(1-2): 491–499.
- [3] OpenFOAM UserGuide. URL: <http://foam.sourceforge.net/docs/Guides-a4/OpenFOAMUserGuide-A4.pdf> (4.02.2017).
- [4] Miller K, Miller RN. Moving finite elements. *Journal on Numerical Analysis – SIAM* 1981; 18(6): 1019–1032.
- [5] Yanenko NN, Lisseikin VD, Kovenia VM. The method of the solution of gaz dynamical problems in moving meshes. *Computing Methods in Applied Sciences and Engineering* 1977; 91(2): 48–61.
- [6] Avdeev E, Fursov V, Ovchinnikov V. An adaptive mesh refinement in the finite volume method. *CEUR Workshop Proceedings* 2015; 1490: 234–241.
- [7] Final drive lubrication OpenFOAM case sources. URL: <https://github.com/j-avdeev/DriveLubrication> (4.02.2017).
- [8] Rusche H. *Computational Fluid Dynamics of Dispersed Two-Phase Flows at High Phase Fractions*. URL: <http://powerlab.fsb.hr/ped/kturbo/OpenFOAM/docs/HenrikRuschePhD2002.pdf> (4.02.2017).

# Numerical modeling of the labyrinth seal taking into account vibrations of the gas transmittal unit rotor in aeroelastic formulation

L.N. Butymova<sup>1</sup>, V.Ya. Modorskii<sup>1</sup>

<sup>1</sup>Perm National Research Polytechnic University, pr. Komsomolsky 29, 614099, Perm, Russia

---

## Abstract

The article deals with the issues related to the mutual influence of vibrations and gas dynamic processes in the labyrinth seals (LS) of gas transmittal unit compressors. The mutual influence of vibrational gas dynamic processes in LS and vibrations of the rotor is studied. Within the framework of the unified algorithm, a solution is obtained for an unsteady aeroelastic one-dimensional gas flow problem in a deformable LS. A new factor (the rotor diameter in the LS region), which affects the pulsation magnitude of the gas dynamic force in the LS, is revealed. Changing the diameter of the rotor, you can reduce vibration. In this case, it is possible to reduce the designated clearances in the LS and to reduce the leakage.

*Keywords:* aeroelasticity; rotor vibration; labyrinth seal; unified algorithm; stress; pressure; deformation; displacement

---

## 1. Introduction

To ensure a contactless connection between a rotating rotor and a stationary body in aircraft engines [16], high-pressure pumps [13, 14], etc., labyrinth seals (LS) are used. In seals of labyrinth type, the working medium is sealed by throttling it when moving through successively located constrictions and extensions.

The main task of the LS is to ensure tightness of the rotor, therefore expansions and constrictions of the flow in LS are usually considered in the direction parallel to the rotor axis. However, in order to ensure aerovibration resistance, it is necessary to take into account the processes of motion of the working medium that take place in the peripheral direction of the LS under the rotor vibrations. It should be noted that sequencing of the constrictions and expansions affects the oscillation amplitude in the gas-dynamic cavity between the LS and the rotor, and also increases the flow non-uniformity. Consequently, refusal to take these elements into account in aeroelastic calculation [15, 21-22] can give an additional margin from the point of view of reducing oscillations in LS, and, which is important for solving the related problems of continuum mechanics [19], reduce complexity and time of calculations.

Based on the said above, the LS calculation is replaced by calculating the gap seal, equivalent (with margin) to the labyrinth seal, if we consider the processes taking place in the LS circumferential direction.

As it is known, LS works at high temperatures and high rotation speeds. Under critical operating conditions the LS is effected by significant loads from the gas-dynamic flow as well as the LS influences the gas-dynamic flow. The impact of this process is ambiguous and requires more detailed research. Publications related to vibrational processes in LS and vibration of rotors, consider the influence of precession [11], geometric characteristics of LS [12] and in the most cases do not take into account the influence of gas-dynamic forces.

The gas dynamic processes that arise in LS at the rotor vibrations caused, for example, by technological imbalances, may lag behind the rotor oscillations. It is necessary to analyze the possibility of amplifying or weakening the rotor vibrations and dependence of these processes on the LS characteristics [9-10].

The classical formulation of the vibration problem takes into account the influence of structural [20], physical-mechanical and technological parameters on vibration, but does not consider the influence of gas dynamic loads.

When considering the problem of the vibration effect on gas-dynamic processes in the labyrinth seals of a centrifugal compressor model gas transmittal unit in a unidirectional dynamic related formulation it is possible to take into account the gas-dynamic factors [13, 17]. In addition, it becomes possible to calculate the oscillation parameters of gas-dynamic forces acting on the rotor [4-8].

## 2. Object of study

A physical model describing the LS operation in an aeroelastic formulation is developed. As a model, the LS scan with the width corresponding to the LS width is considered. In a unidirectional FSI-statement, the rotor motion is replaced by the movement of pistons located diametrically opposite, and moving with the specified amplitude and frequency (Fig.1). The scan length is equal to the circumference of the gas-dynamic cavity, which is aligned along the middle line between the rotating rotor and the stationary LS.

The model is quasi-two-dimensional, dynamic.

Thus, the rotor oscillations in the LS gap are modeled by a nonlinear dynamic quasi-two-dimensional gas dynamic scan model of the LS gap with movable boundaries. Extrusion of the gas as the gap is reduced and filling free volume with the gas as the gap is increased during the rotor oscillation in the LS is modeled by two pistons moving in harmonic order. Their oscillation frequency is the same and corresponds to the rotor oscillation frequency, calculated from the wave path equal to the length of the circumference arc around the rotor. The piston oscillation amplitudes are equal (formula 1). While calculating, the

oscillations of displacements, velocities, pressures and gas-dynamic force acting on the rotor in the LS area are recorded at the control points. The displacement of these oscillations ( $\varphi_U$ ) may take place with respect to two parameters: the gas-dynamic forces acting on the rotor in the LS area and the rotor displacements. With different phases ( $\varphi_U$ ) convergent, divergent and steady oscillatory processes can be observed.

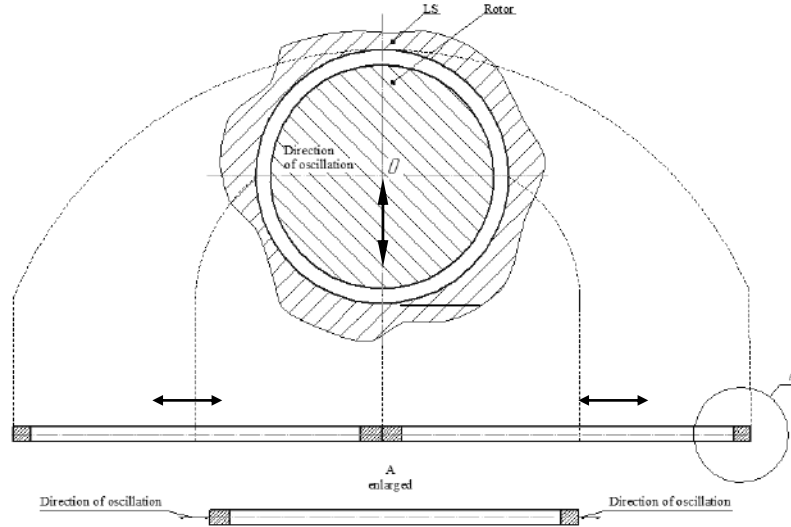


Fig. 1. Formation of the LS calculation scheme (unidirectional FSI-statement).

### 3. Mathematical model

The mathematical description of the gas-elastic process in this formulation includes the following relationships:

$$\frac{\partial \rho_{\Gamma}}{\partial t} + \frac{\partial \rho_{\Gamma} V_x}{\partial x} = 0 \quad (1)$$

$$\frac{\partial \rho_{\Gamma} V_{x\Gamma}}{\partial t} + \frac{\partial (\rho_{\Gamma} V_{x\Gamma} V_{x\Gamma})}{\partial x} + \frac{\partial P}{\partial x} = 0 \quad (2)$$

$$\frac{\partial \rho_{\Gamma} E}{\partial t} + \frac{\partial (\rho_{\Gamma} E V_{x\Gamma})}{\partial x} + \frac{\partial (P V_x)}{\partial x} = 0 \quad (3)$$

$$P = \rho_{\Gamma} (k-1) \left( E - V_{x\Gamma}^2 / 2 \right) \quad (4)$$

$$\frac{\partial \rho_K}{\partial t} + \frac{\partial (\rho_K V_{xK})}{\partial x} = 0 \quad (5)$$

$$\frac{\partial (\rho_K V_{xK})}{\partial t} - \frac{\partial \sigma_{xx}}{\partial x} = 0 \quad (6)$$

$$U_x = U_{x0} + \int_0^t V_{xK}(t) dt \quad (7)$$

$$\varepsilon_{xx} = \frac{\partial U_x}{\partial x} \quad (8)$$

$$\sigma_{xx} = E \varepsilon_{xx} \quad (9)$$

- initial conditions

at  $t = 0$  (Gas):

$$P = P_0, \quad \rho_{\Gamma} = \rho_{0\Gamma}, \quad E = P_0 / \rho_{0\Gamma} (k-1) \quad (10)$$

at  $t = 0$  (SSS):

$$\sigma_{xx} = 0, \quad \varepsilon_{xx} = 0, \quad U_{x0} = 0, \quad V_x = 0$$

Boundary conditions (SSS)

a) «rigid wall»

(«Sticking»)

$$V_{xK} = 0, \quad V_{x\Gamma} = 0, \quad (11)$$

a) gas-structure

$$\sigma_{xx} = -P_{\Gamma P} \quad V_{.xK} = V_{.xT}$$

The boundary condition for the piston motion:

$$V_{\text{left.piston}} = -V_0 \sin(\omega t); \quad V_{\text{right.piston}} = V_0 \sin(\omega t) \tag{12}$$

where  $V_0$  – amplitude of the piston oscillations,  $\omega$  – the piston oscillation frequency.

**4. Method of solution**

For solving gas dynamic tasks was used method of large particles. Using the same method for solving gas dynamic and stress-strain state tasks provide unity mesh for gas dynamic region and stress-strain state tasks. For this used unified system of differential equations to ensure coupled solving for elastic tasks and gas dynamic tasks. Thus we used method of large particles for calculations.

The main idea of method’s large particles consisted in splitting into physical processes of the initial non-stationary system of Euler equations which written in the form’s conservation laws. The space is modeled by particle system which coincides with cell’s Euler grid in the moment. If stationary solving is, we get it in process stabilized solving. So all process solving composed multiple repetitions of time steps.

Each computational cycle is divided into seven stages. The first three stages are designed to solve the gas dynamic tasks. The next four stages are designed to evaluate the parameters dynamic stress-strain state of the structure.

**5. Description of results**

*5.1. Analysis of the influence of geometric characteristics*

In a unidirectional aeroelastic formulation, the solution using equations (1-4) and initial and boundary conditions, yielded the following results.

When investigating the dependence of pressure fluctuations in the LS gas-dynamic cavity on the LS geometric characteristics the rotor diameters in the LS area varied and were equal to 65, 130, 195 or 260 mm. The working body of the gas-dynamic cavity is air, adiabatic exponent equals 1.4, density  $\rho = 1,29 \text{ kg / m}^3$ .

With an increase in the shaft diameter from 65mm to 260mm, the pressure oscillations in the gas-dynamic cavity are noted, which have a time periodic character (Fig. 2). The pressure amplitude is 101368.8 Pa with  $D = 65\text{mm}$ , with  $D = 130 \text{ mm}$  the pressure amplitude is 148792.2 Pa, and with  $D = 195\text{mm}$  the pressure amplitude is 107577.7 Pa, with  $D = 260\text{mm}$  the pressure amplitude is 101732.1Pa.

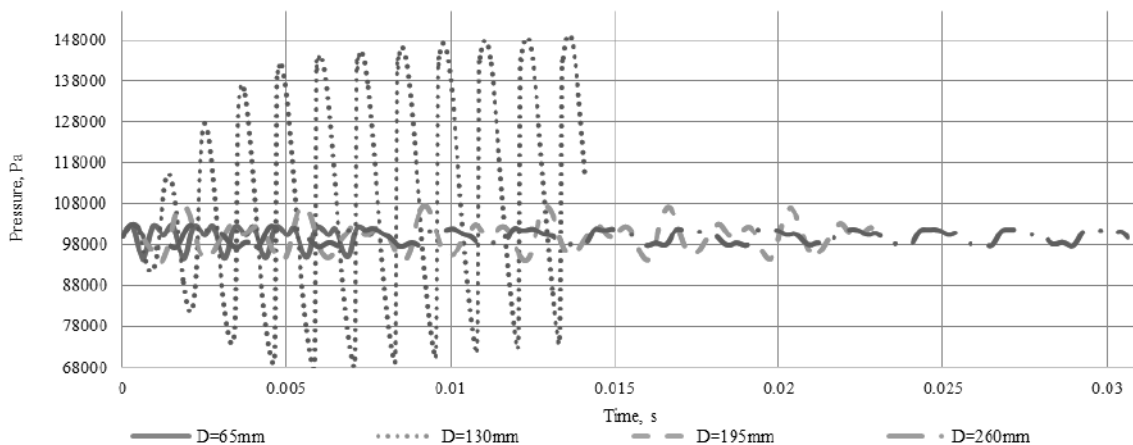


Fig. 2. Dependence of the gas dynamic pressure of the gap near LS on time at different geometric characteristics of LS.

Table 1. Pressure amplitude for LS different geometric characteristics.

Computational experiment	1	2	3	4
Rotor diameter D, mm	65	130	195	260
Maximum pressure amplitude $U_p$ , MPa	0.1	0.15	0.11	0.102

When studying the dependence of the temperature oscillations in the gas-dynamic cavity on the geometric characteristics of the LU, the diameters of the rotor in the zone of the LU varied and assumed values of 65, 130, 195 or 260 mm, the working fluid of the gas-dynamic cavity is air, the adiabatic index is 1.4,  $\rho = 1,29 \text{ kg/m}^3$ .

With an increase in shaft diameter from 65mm to 260mm, oscillations in the temperature of the gas-dynamic cavity are noted, which have a periodic character of the change in time (Fig. 3). The temperature amplitude is 272.4936K. At  $D = 65\text{mm}$ ,

at  $D = 130\text{mm}$  the temperature amplitude is  $323.6043\text{K}$ , and at  $D = 195\text{mm}$  the temperature amplitude is  $279.0668\text{K}$ , at  $D = 260\text{mm}$  the temperature amplitude is  $274.4728\text{K}$

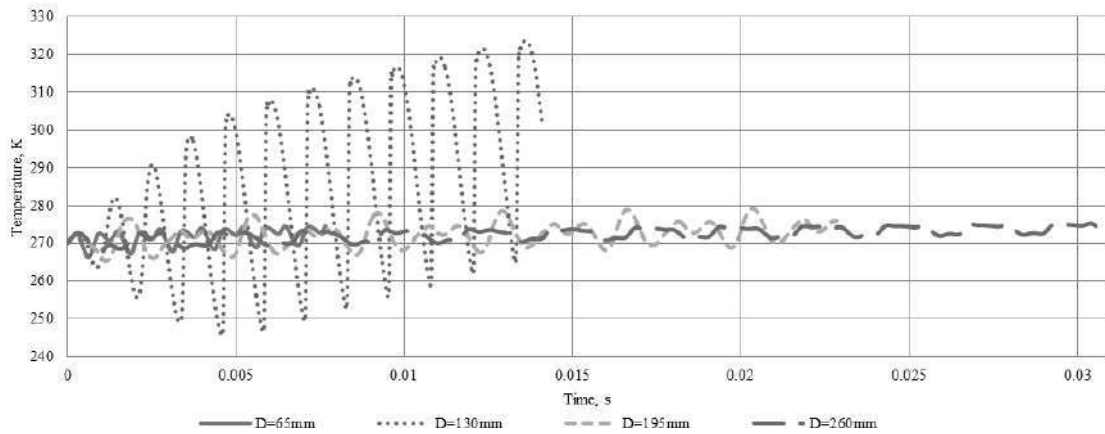


Fig.3. Dependence of the gas-dynamic gap temperature near LS on the time for LS different geometric characteristics.

Thus, when designing the LS it is necessary to take into account the geometric dimensions of the rotor and the gap between the rotor and the LS in order to reduce possible vibrations.

Table 2. Temperature amplitude for LS different geometric characteristics.

Computational experiment	1	2	3	4
Rotor diameter $D$ , mm	65	130	195	260
Temperature amplitude $U_T$ , K	272.4936	323.6043	279.0668	274.4728

### 5.2. Analysis of the influence of kinematic parameters

The dependence of pressure oscillations in the LS gas dynamic cavity on the kinematic parameters of propagation speed of gas oscillations in the circumferential direction varied and equaled 3.5, 7.0 or 10.5 m/s, the working fluid of the gas-dynamic cavity is air, the adiabatic index is 1.4,  $\rho = 1, 29 \text{ kg/m}^3$ .

With speed increase from 3.5 m/s to 10.5 m/s, the pressure oscillations of the gas-dynamic cavity are noted, which have a periodic character (Fig. 4). The pressure amplitude is 100918.3 Pa at  $V = 3.5 \text{ m/s}$ . At  $V = 7.0 \text{ m/s}$ , the pressure amplitude is 103,000 Pa, and at  $V = 10.5 \text{ m/s}$  the pressure amplitude is 104419Pa.

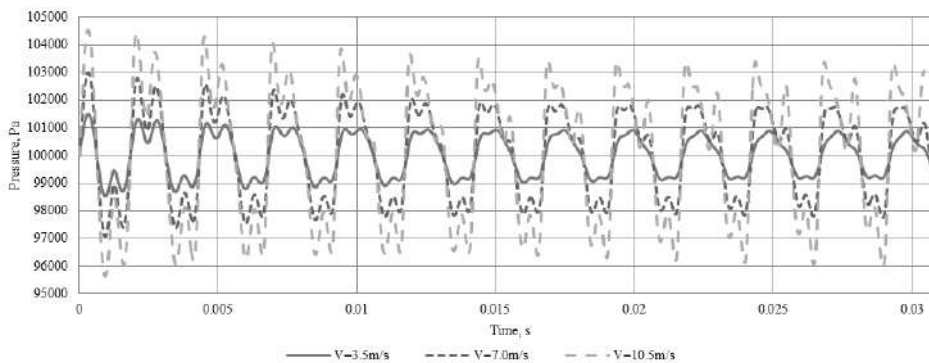


Fig.4. Pressure dependence in the gas-dynamic gap near the LS on time for different elasticity moduli of LS material.

Table 3. Pressure amplitude at LS different kinematic parameters.

Computational experiment	1	2	3
Propagation speed of gas oscillations in the LS circumferential direction, m/s	3.5	7.0	10.5
Maximum pressure amplitude $U_p$ , MPa	0.101	0.103	0.104

When studying the temperature dependence of the gas dynamic cavity in the LS on kinematic parameters, the propagation speed of the gas oscillations in the circumferential direction varied and assumed values of 3.5, 7.0 or 10.5 m/s, the working fluid of the gas-dynamic cavity is air, adiabatic index is 1.4, air density  $\rho = 1, 29 \text{ kg/m}^3$ .

With speed increase from 3.5 m/s to 10.5 m/s oscillations in the temperature of the gas-dynamic cavity are observed, which have a periodic character (Fig. 5). The temperature amplitude is 271.9472K at  $V = 3.5$  m/s. At  $V = 7.0$  m/s the temperature amplitude is 275.1519 K, and at  $V = 10.5$  m/s the temperature amplitude is 279.8211K.

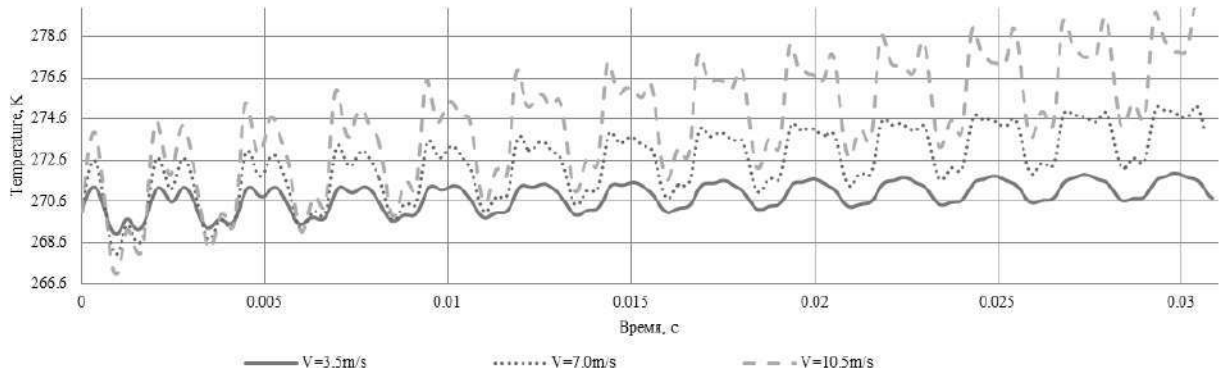


Fig.5. Temperature dependence of the LS gas-dynamic gap on time for various kinematic parameters.

Table 4. Temperature amplitude for LS different kinematic parameters.

Computational experiment	1	2	3
Propagation speed of gas oscillations in the LS circumferential direction, m/s	3.5	7.0	10.5
Temperature amplitude $U_T$ , K	271.9472	275.1519	279.8211

### 5.3. Analysis of the influence of the working fluid characteristics

When studying the dependence of pressure oscillations in the LS gas-dynamic cavity on the working fluid characteristics the adiabatic index varied and was 1.1, 1.25 or 1.4 at density  $\rho = 1,29 \text{ kg/m}^3$ .

With an increase in the adiabatic index from 1.1 to 1.4, the pressure oscillations of the gas-dynamic cavity are observed, which have a periodic character (Fig. 6). The pressure amplitude is 102880.5 Pa for  $k = 1.1$ . For  $k = 1.25$ , the pressure amplitude is 102816.3 Pa, and for  $k = 1.4$  the pressure amplitude is 1003003Pa.

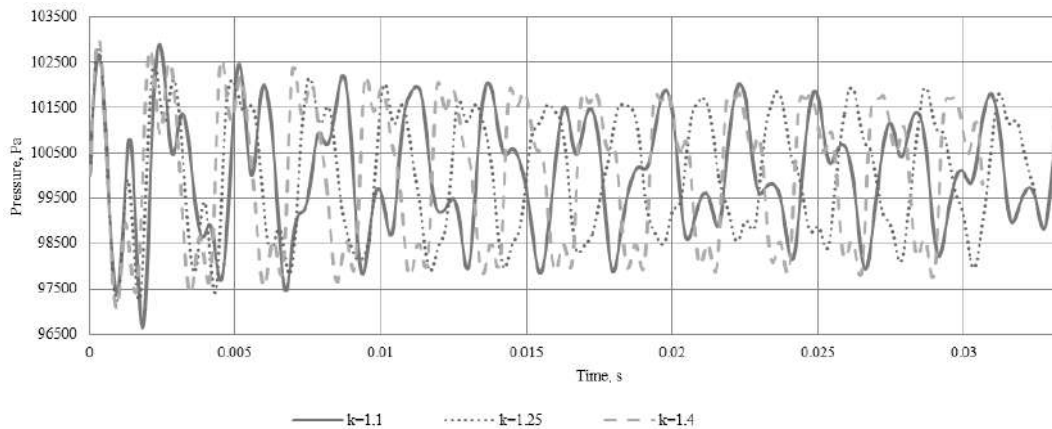


Fig.6. Pressure dependence of the LS gas-dynamic gap on time for different working fluid characteristics.

Table 5. Pressure amplitude for LS different working fluid characteristics.

Computational experiment	1	2	3
Adiabatic index, k	1.1	1.25	1.4
Maximum pressure amplitude $U_P$ , MPa	0.103	0.103	0.1

When studying the dependence of temperature oscillations in the LS gas-dynamic cavity on the working fluid characteristics the adiabatic index varied and was 1.1, 1.25 or 1.4, with density  $\rho = 1,29 \text{ kg/m}^3$ .

With an increase in the adiabatic index from 1.1 to 1.4, the temperature oscillations of the gas-dynamic cavity are observed, which have a periodic character (Fig. 7). The temperature amplitude is 271.8515 K for  $k = 1.1$ . For  $k = 1.25$  the temperature amplitude is 274.0991K, and for  $k = 1.4$  the temperature amplitude is 275.1212K.

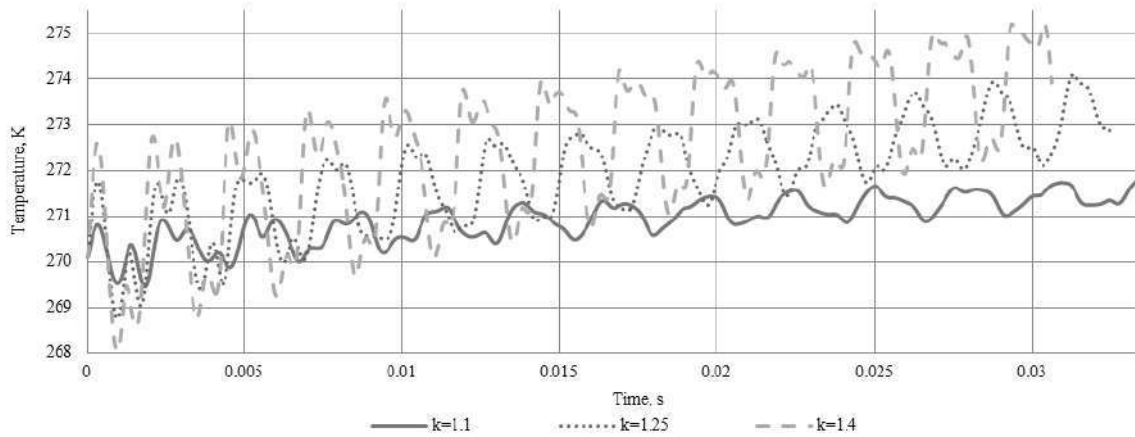


Fig.7. Temperature dependence of the LS gas-dynamic gap on time for different working fluid characteristics.

Table 6. Temperature amplitude for LS different kinematic parameters.

Computational experiment	1	2	3
Adiabatic index, k	1.1	1.25	1.4
Temperature amplitude $U_T$ , K	271.8515	274.0991	275.1212

#### 5.4. Analysis of the influence of physical and mechanical characteristics

In the bi-directional aeroelastic formulation the LS calculation scheme was generated (Fig. 8) and a solution was obtained using equations (1-9) and initial and boundary conditions, which allowed obtaining the following results.

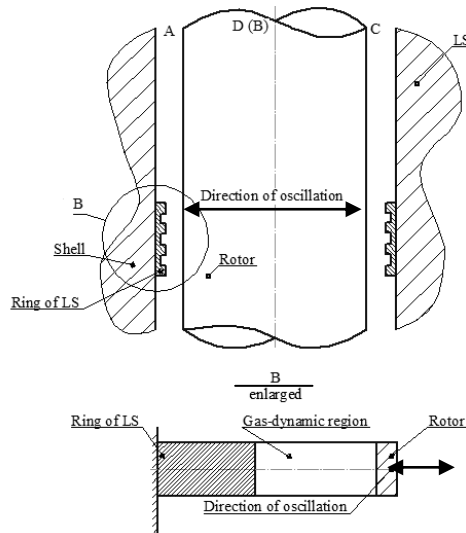


Fig. 8. Formation of the LS calculation scheme (bidirectional FSI-statement)

When studying the dependence of pressure oscillations in the LS gas-dynamic cavity on physico-mechanical characteristics of the LS the material density was set  $\rho = 7800 \text{ kg/m}^3$ , Poisson ratio  $\mu = 0.35$ , the elasticity modulus ranged within 50, 100, 150, 200 GPa. With an increase in elasticity modulus from 50 GPa to 200 GPa, the amplitude of the periodic pressure oscillations of the gas dynamic cavity decreases by 5 times. The pressure amplitude is 1 MPa at  $E = 50 \text{ GPa}$ . At  $E = 100 \text{ GPa}$  the pressure amplitude is 0.5 MPa, and at  $E = 150 \text{ GPa}$  the pressure amplitude is 0.3 MPa, at  $E = 200 \text{ GPa}$  the pressure amplitude is 0.2 MPa (Fig. 9). Fig. 9 shows the dependence of pressure oscillations in the LS gas-dynamic gap on time for LS various physico-mechanical characteristics.

When studying the dependence of displacements in the LS structure on physico-mechanical characteristics the density of the LS material was set  $\rho = 7800 \text{ kg/m}^3$ , poisson ratio  $\mu = 0.35$ , adiabatic index  $k = 1.4$ , air density  $\rho = 1.29 \text{ kg/m}^3$ ,  $P_0 = 0.1 \text{ MPa}$ , the calculated ratio of cells in the structure to the total number of cells calculated  $FL = 0.96$ , the elastic modulus varied within 50, 100, 150 and 200 GPa. Near the gas-dynamic gap with an increase in  $E$  from 50 GPa to 200 GPa oscillations of displacements in the LS structure are observed. The oscillations are periodic in nature and stable in time. The displacement amplitude is  $1 \times 10^{-2}$  microns at  $E = 50 \text{ GPa}$ . When  $E = 100 \text{ GPa}$  the displacement amplitude is  $5 \times 10^{-3} \text{ m}$ , and when  $E = 150 \text{ GPa}$  the displacement amplitude is  $3 \times 10^{-3} \text{ m}$ , for  $E = 200 \text{ GPa}$  the displacement amplitude is  $2.3 \times 10^{-3} \text{ m}$  (Fig.10). Figure 11 shows the relationship of displacements in the LS structure versus time for LS different physico-mechanical characteristics.

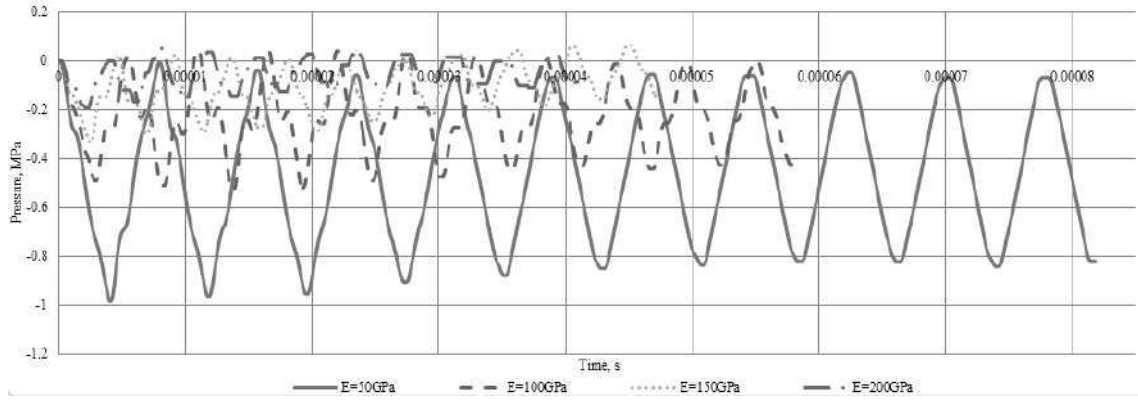


Fig. 9. Dependence of pressure oscillations in the LS gas-dynamic gap on time for LS various physico-mechanical characteristics.

Table 7. Amplitude of pressure oscillations at different values of LS physico-mechanical characteristics.

Computational experiment	1	2	3	4
Elasticity modulus E, GPa	50	100	150	200
Pressure amplitude $U_p$ , MPa	1	0.5	0.3	0.2

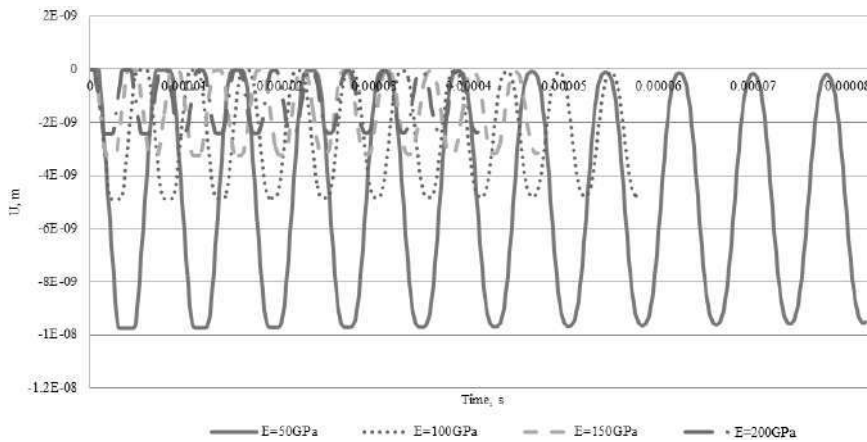


Fig. 10. Dependence of displacements in the LS structure versus time for LS various physico-mechanical characteristics.

Table 8. Displacements in the LS structure for LS different physico-mechanical characteristics.

Computational experiment	1	2	3	4
Elasticity modulus E, GPa	50	100	150	200
Displacement amplitude $U_A$ , m	$1 \times 10^{-2}$	$5 \times 10^{-3}$	$3 \times 10^{-3}$	$2.3 \times 10^{-3}$

## 6. Conclusions

1. With an increase in the compression wave velocity arising from the approach of the rotor to the surface of the LS under vibrations from 3.5 m/s to 10.5 m/s the amplitude of the gas dynamic force increases from 23.6H to 45.5H. The frequency does not change and is equal to 400 Hz. At a natural rotor frequency of 808 Hz, one can expect that with a minimum value of the gas flow velocity in the circumferential direction, weak vibrations may appear. As the speed increases, one can expect an increase in the LS vibrations.

2. With the shaft diameter increase from 65mm to 260mm, the maximum amplitude of the gas dynamic force of 82.7N is observed at a diameter equal to 130 mm, the minimum amplitude of the gas dynamic force is observed at a diameter of 65mm from 7.7H, the frequency is 400Hz. The maximum frequency of the gas dynamic force is 770 Hz with a diameter of 65 mm. The minimum frequency of the gas-dynamic force is 406 Hz with the diameter equal to 260 mm. It can be seen that as the rotor diameter increases, the nominal values of the gas dynamic force increase. This is due to the increase in the rotor area at a constant nominal pressure. In this case the maximum amplitudes of gas-dynamic forces are observed when the oscillation frequency  $f_p$  of the rotor is equal to the first natural frequency of the gas-dynamic pressure fluctuations of the circumferential cavity in the gap. Oscillation amplitude of gas-dynamic forces are lower at the rotor oscillation frequency  $f_p$  equal to the second natural frequency of the gas-dynamic pressure oscillations of the circumferential cavity in the gap. Even lower are the amplitudes of gas-dynamic force oscillations at the rotor oscillation frequency  $f_p$  equal to the fourth natural frequency of the circumferential pressure oscillations of the gas-dynamic cavity in the gap. The oscillation amplitude of the gas-dynamic forces was also low at the natural frequency of the gas-dynamic cavity nonmultiple for the rotor frequency.

3. With an increase in the adiabatic index  $k$  from 1.1 to 1.4 the gas-dynamic force oscillation amplitude increases from 29.63N to 35.15N, and the oscillation frequency of gas-dynamic forces decreases from 392Hz to 388Hz. Thus, we note a weak influence of the working fluid characteristics on the LS vibrations.



4. With increase in elastic modulus from 50GPa to 200GPa the pressure oscillation amplitude decreases from 0.97MPa to 0.19MPa, pressure oscillation frequency rises from 134kHz to 256kHz. Analysis of the influence of physical and mechanical characteristics on the LS vibrations demonstrated that with a hard material the strain rate is lower than with a soft material. The displacements in a softer material under given loads are  $1 \times 10^{-2}$  microns. In a harder material the displacement amplitude is much lower.

5. By changing the rotor diameter, it is possible to reduce vibration. Thus, it is possible to reduce the gaps in the LS and reduce leakage.

## Acknowledgements

The study was performed with a grant from the Russian Science Foundation (project №14-19-00877).

## References

- [1] Butymova LN. Effect of vibration on gas dynamics in modeling labyrinth seals of a gas transmittal unit centrifugal compressor. *Bulletin of Perm National Research Polytechnic University. Aerospace technology* 2016; 47: 243–259.
- [2] Butymova LN, Modorsky VY, Petrov VY. Numerical modeling of the dynamic interaction in system "gas-structure" with harmonic motion of the piston in the variable section pipe. *AIP Conference Proceedings* 2016; 1770: 030103-1–030103-5.
- [3] Butymova LN, Modorskii VY. One-way FSI simulation of the phase and the geometric parameters of the model of compressor blades on the oscillating gas-dynamic processes pipe. *MATEC the Web Conf.* 2016; 75: 4 p.
- [4] Butymova LN, Modorsky VY, Petrov VY. Numerical modeling the kinematic parameter effect on the vibrations of the model compressor vanes in the system "gas-structure". *Scientific and Technical Gazette Volga region* 2015; 5: 157–160.
- [5] Butymova LN, Modorsky VY, Shmakov AF. Experimental estimation of amplitude and phase characteristics of the interaction of gas-dynamic flow and structure. *Scientific and Technical Gazette Volga Region* 2014; 5: 127–129.
- [6] Butymova LN, Modorsky VY. Investigation of gas-dynamic flow and structure in a model experimental setup. *Bulletin of South Ural State University. Series: Computational Mathematics and Computer Science-* 2014; 3(2): 92–100.
- [7] Butymova LN, Modorsky VY. Study of oscillatory processes on resonant modes in the model apparatus. *Scientific and Technical Gazette Volga region* 2013; 6: 193–196.
- [8] Butymova LN, Modorsky VY, Sokolkin YV. Development of experimental apparatus and investigation of the body material influence on the resonance frequencies in the "gas-structure". *Scientific and Technical Gazette Volga Region* 2013; 6: 197–200.
- [9] Mekhonoshina EV, Modorsky VY. Development of numerical simulation techniques of the compressor aeroelastic operation. *Scientific and Technical Gazette Volga Region* 2014; 5: 264–268.
- [10] Mekhonoshina EV, Modorsky VY. Impact of magnetic suspension stiffness on aeroelastic compressor rotor vibrations of gas pumping units. *AIP Conference Proceedings* 2016; 1770: 030113-1–030113-5.
- [11] Makarov AA, Zaytsev NN. Engineering and theoretical problems of labyrinth seal applications in high-speed rotary machines. *Herald PNRPU. Aerospace engineering* 2015; 42: 61–81.
- [12] Brikin BV, Evdokimov IE. Numerical simulation of the experiment on the flow in the labyrinth seal. *Proceedings of the MAI # 61.* URL: [http://mai.ru/science/trudy/published.php\(02/02/2017\)](http://mai.ru/science/trudy/published.php(02/02/2017)).
- [13] Arbuzov IA, Tashkinov AA, Schenyatsky DV, Kirievsky BE, Bulbovich RV, Modorsky VY, Pisarev PV. Analysis of the impact of the entry apparatus in the connecting channel on the vibrational processes in the first stage of two-stage model pump. *Scientific and Technical Gazette Volga* 2012; 6: 108–111.
- [14] Gaynutdinova DF, Modorsky VY, Shevelev NA. Experimental modeling of cavitation occurring at vibration. *AIP Conference Proceedings* 2016; 1770: 030111-1–030111-4.
- [15] Shmakov AF, Modorsky VY. Numerical simulation of gas-dynamic, thermal processes and evaluation of the stress-strain state in the modeling compressor of the gas-distributing unit. *AIP Conference Proceedings* 2016; 1770: 030108-1–030108-5.
- [16] Babushkina AV, Modorsky VY, Sipatov AM, Kolodyazhny DY, Nagorny VS. Modeling technique for the process of liquid film disintegration. *AIP Conference Proceedings* 2016; 1770: 030109-1–030109-7.
- [17] Kalyulin SL, Modorsky VY, Paduchev AP. Numerical design of the rectifying lattices in a small-sized wind tunnel. *AIP Conference Proceedings* 2016; 1770: 030110-1–030110-4.
- [18] Modorsky VY, Shevelev NA. Research of aerohydrodynamic and aeroelastic processes on PNRPU HPC system. *AIP Conference Proceedings* 2016; 1770: 030110-1–030110-4.
- [19] Modorsky VY et al. Numerical study of actual problems of mechanical engineering and mechanics of solid and bulk materials by large particles method. A study of actual problems of mechanics and engineering. Moscow: National Academy of Applied Sciences, International Association of developers and users of the method of large particles 1995; 5: 1658 p.
- [20] Modorsky VY, Shmakov AF, Butymova LN, Gainutdinova DF, Mekhonoshina EV, Kalyulin SL. Parallel calculation of dynamic processes in large-sized supercharger. *Scientific service on the Internet: A variety of supercomputing worlds Proceedings of the International Supercomputer Conference. Russian Academy of Sciences Supercomputing Consortium of Russian Universities* 2014: 258–262.
- [21] Shmakov AF, Modorsky VY. Energy Conservation in Cooling Systems at Metallurgical Plants. *Metallurgist* 2016; 59(9): 882–886.
- [22] Mekhonoshina EV, Modorskii VY. On a phase-shift of waves at the medium interface. *Computer Optics* 2015; 39(3): 385–391. DOI: 10.18287/0134-2452-2015-39-3-385-391.

# High-performance DTW-based signals comparison for the brain electroencephalograms analysis

A.I. Makarova<sup>1</sup>, V.V. Sulimova<sup>1</sup>

<sup>1</sup>Tula State University, Lenin Ave., 92, 300012, Tula, Russia

## Abstract

Automatic analysis of electroencephalograms (EEGs) is one of the promising areas of research, which results can be used, in particular, to build systems of mental control of objects. The Dynamic Time Warping procedure (DTW) is used in this work for comparing signals representing EEG. An important feature of the problem we are considering is the need for multiple comparison of signals at the stage of machine learning, which requires enormous computational costs. We propose a parallel algorithm that was implemented in C++ using the MPI technology and tested using the resources of the supercomputer complex of Moscow State University “Lomonosov”. The results of its testing on real data showed that the proposed method allows achieving an almost linear speedup and reducing the total calculation time from 29 days to 3.5 hours using 128 processes, which opens the possibility of improving the quality of automatic analysis of electroencephalograms.

*Keywords:* high performance computing; comparing fragments of electroencephalograms; Dynamic Time Warping; the MPI technology; the supercomputer complex of MSU “Lomonosov”

## 1. Introduction

Electroencephalography is a method, which consists in reading electrical signals using electrodes located on the scalp. These signals are generated by the brain in the process of brain activity [1,2] and recorded as electroencephalograms for their subsequent analysis.

Automatic analysis of brain electroencephalograms (EEGs) is one of the promising areas of scientific research, which results can be used, in particular, to build systems for mental control of external devices, e.g. a computer (brain-computer interface, BCI) [3].

The idea of such “brain-computer” interface is based on the fact that signals generated by the brain under the influence of certain (target) stimuli, corresponding to a specific task, contain specific components which are absent under the influence of other stimuli. This happens in particular in the recognition of an object of a given type among a number of other objects. It is obvious that the problem of automatic recognition of the target object on the basis of the analysis of electroencephalograms plays one of the key roles in the construction of a system for the mental control of external objects [4,5]. Thus, it is especially important to improve the quality of its solution.

In this paper we, following [5], focus on the experimental research scheme, according to which a person is provided with a series of mammogram images and his task is to find among them mammograms containing pathologies (the so-called target mammograms). EEG signals from 66 electrodes fixed on different parts of one's head are recorded during the process of viewing images. The task in this case is the automatic detection of signals registered in the process of viewing target mammograms. Figure 1a shows the process of registering electroencephalograms, and figures 1b and 1c - examples of mammograms with pathology and without pathology, respectively [5].

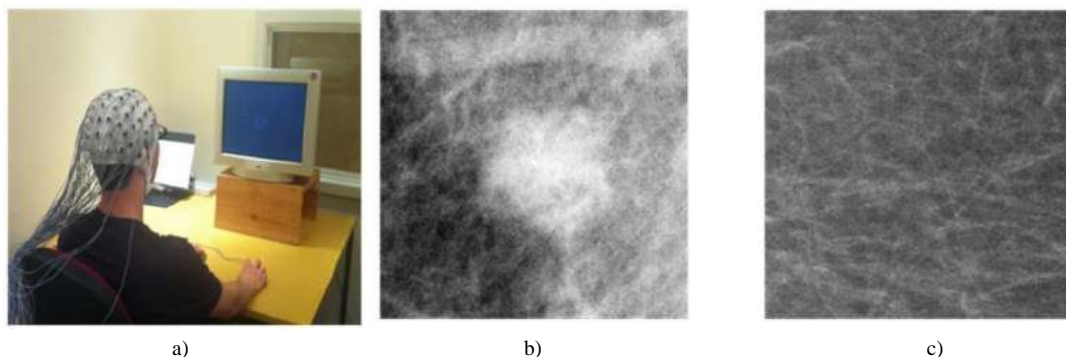


Fig. 1. a) The process of registering electroencephalograms, b) an example of a target object (mammograms with pathology), c) an example of an un-target object (mammograms without pathologies).

Figure 2 shows examples of signals recorded from two of the 66 electrodes during viewing target (solid lines) and non-target (dashed lines) objects. The signals are subjected to preprocessing (filtration, scaling and smoothing by a sliding window), used by us to improve the quality of target objects recognition.

From a mathematical point of view, the electroencephalograms, which need to be analyzed for each of the 66 electrodes, fixed in different parts of the head are single-component discrete signals.

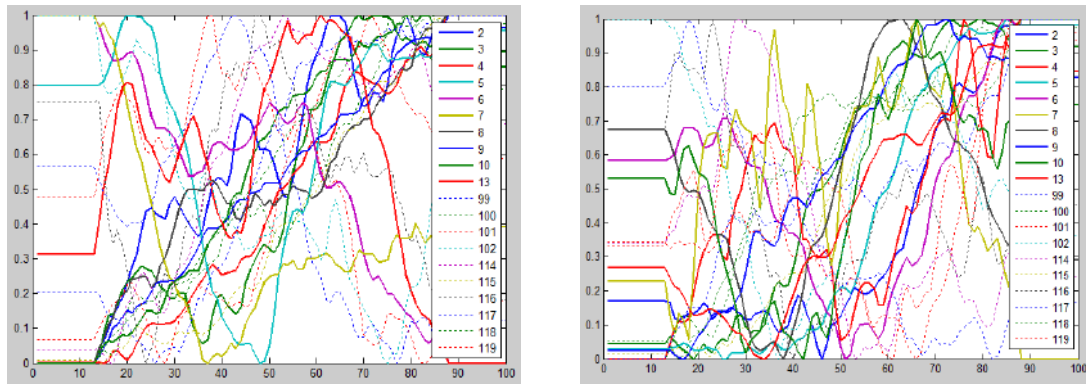


Fig. 2. Examples of signals received from two electrodes. The solid lines show the signals corresponding to the target objects, and the dotted lines show the nontarget objects.

Obviously, the analysis of electroencephalograms must inevitably be based on comparing the signals representing them. Since the form of the response to the stimulus and the time of its onset can vary, in this work, for the comparison of signals an adapted version of Dynamic Time Warping (DTW) procedure is used. This algorithm based on the search for optimal pairwise alignment of compared signals by local compression and stretching of their axes. Initially, the DTW method was proposed to compare speech signals [6], but was later adapted for many other areas [7,8].

The task of analyzing electroencephalograms within the framework of this work is formulated in the form of a two-class pattern recognition problem, which solution is carried out in two stages - training and recognition [9]. At the training stage, a decision rule is formed for assigning new signals to the target or non-target class on the basis of processing a certain finite set of signals with a known class affiliation. At the recognition stage, the constructed decision rule is applied to new signals. Recognition can be carried out very quickly, while the process of constructing a good decision rule, which allows to classify new signals with high accuracy, is very laborious.

Modern methods of machine learning in the construction of decision rules are based on measures of object comparison and allow us to automatically choose the most suitable ones in the training process, improving the quality of the solution of the problem [10]. The great complexity of training is due to the necessity of multiple comparison of long signals in the calculation of a whole series of matrices of their pairwise dissimilarity (for different electrodes, different types of preprocessing, different values of the parameters of the comparison algorithm), choosing the most suitable of which is not possible a priori.

In accordance with the above, an extremely topical task is to increase the productivity of the electroencephalogram comparison.

There are a number of ways to speed up the Dynamic Time Warping procedure, however, they either do not guarantee the finding of the optimal solution [6,11-14], or the performance improvement is provided only in cases, when comparing close or sparse signals [15,16]. But such situations are rare in the analysis of electroencephalograms. The use of modern parallel computing technologies is a fundamentally different direction for increasing productivity of signals comparison. However, the implementation of known parallel versions of DTW procedure (as well as similar tasks with cycles having diagonal dependencies) do not give the desired effect due to the need for frequent synchronization of processes or threads, and in some cases may even lead to an increase in the operating time compared to the serial version because of less efficient work with the cache memory [17-21].

In this paper, we propose a parallel algorithm that significantly improves the performance of calculating the matrix of pairwise comparisons for any signals representing fragments of an electroencephalogram. Increase in productivity is performed without loss of accuracy of calculations due to taking into account the features of the task, allowing to implement parallelization at a higher level.

The proposed algorithm is implemented in C++ programming language using the MPI technology [21] and tested using the resources of the supercomputer complex of Moscow State University "Lomonosov" [23]. The results of the research on real data showed that the proposed method allows to achieve near-linear acceleration, and to reduce the total calculation time from 29 days to 3.5 hours using 128 processes, which opens the possibility of improving the quality of automatic analysis of electroencephalograms.

## 2. Comparison of fragments of electroencephalograms based on DTW

### 2.1. The mathematical formulation of the problem of comparing two signals

Let  $\mathbf{x} = (x_1, x_2, \dots, x_{N_x})$  and  $\mathbf{y} = (y_1, y_2, \dots, y_{N_y})$  - are two single-component discrete signals, which length are  $N_x$  and  $N_y$ , respectively, and which consists of  $x_i, y_j$  OR,  $i = 1, \dots, N_x, j = 1, \dots, N_y$  elements.

Specific pairwise warping

$$\mathbf{w}(\mathbf{x}, \mathbf{y}) = \left\{ \begin{matrix} w_{i,1} \\ w_{i,2} \end{matrix}, i = 1, \dots, N_w \right\}, w_{i,1} \in \{1, \dots, N_x\}, w_{i,2} \in \{1, \dots, N_y\}$$

of two signals  $\mathbf{x}$  and  $\mathbf{y}$  uniquely determines the correspondence of signal's elements and has a length  $N_w$ , equal to the number of such pairwise correspondences. At that first and last elements of signals are certainly corresponded to each other:

$$w_{1,1} = w_{1,2} = 1, w_{N_w,1} = N_x, w_{N_w,2} = N_y.$$

A warping of two signals is considered as optimal one if it ensures the minimum value of the following optimality criterion:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} J(\mathbf{x}, \mathbf{y}, \mathbf{w}), J(\mathbf{x}, \mathbf{y}, \mathbf{w}) = \sqrt{\min_{\mathbf{w}} \sum_{i=1}^{N_x} D(\mathbf{x}, \mathbf{y}, \mathbf{w}, i)}, \quad (1)$$

where

$$D(\mathbf{x}, \mathbf{y}, \mathbf{w}, i) = |x_i - y_i| + \begin{cases} 0, & (w_{i,1} = w_{i-1,1} + 1) \wedge (w_{i,2} = w_{i-1,2} + 1), \\ \beta, & \left\{ \begin{matrix} (w_{i,1} = w_{i-1,1} + 1) \wedge (w_{i,2} = w_{i-1,2}) \\ (w_{i,1} = w_{i-1,1}) \wedge (w_{i,2} = w_{i-1,2} + 1) \end{matrix} \right\} \wedge \wedge \\ \infty, & \text{иначе.} \end{cases}$$

bi 0 - penalty for non-parallel references between elements of signals, corresponded to local warping axes.

The optimal value of the criterion can be considered as a dissimilarity of the signals:

$$r(\mathbf{x}, \mathbf{y}) = \sqrt{J(\mathbf{x}, \mathbf{y}, \hat{\mathbf{w}})}.$$

### 2.2. A sequential algorithm for computing the dissimilarity of two signals

The minimum of the optimality criterion (1) can be found by means of the dynamic programming procedure [24]. It is convenient to represent the algorithm of finding the optimal warping in terms of an oriented graph of pairwise correspondences (Figure 3), in which the nodes correspond to the comparison of the signal's elements, the horizontal and vertical edges correspond to the local warping of the axes (passing through them is penalized with the positive penalty  $\beta$ ), and the diagonal edges - parallel references between signal's elements.

The algorithm of finding the optimal pairwise warping consists in consecutive passing through all the vertices of the graph, beginning with the upper left and ending with the bottom right vertex. An incomplete value of the optimality criterion  $\bar{J}_{i,1}$  is calculated at each vertex on the basis of the initial parts of the signals  $\mathbf{x}_{1..i} = (x_1, \dots, x_i)$ ,  $\mathbf{y}_{1..j} = (y_1, \dots, y_j)$ :

$$\bar{J}_{1,1} = |x_1 - y_1|,$$

$$\bar{J}_{1,j} = \bar{J}_{1,j-1} + |x_1 - y_j| + \beta, \quad j = 1, \dots, N_y,$$

$$\bar{J}_{i,1} = \bar{J}_{i-1,1} + |x_i - y_1| + \beta, \quad i = 1, \dots, N_x,$$

$$\bar{J}_{i,j} = |x_i - y_j| + \min \{ \bar{J}_{i-1,j-1}, \bar{J}_{i-1,j} + \beta, \bar{J}_{i,j-1} + \beta \}, \quad j = 2, \dots, N_y, i = 2, \dots, N_x.$$

The sought value of the dissimilarity of two signals:  $r(\mathbf{x}, \mathbf{y}) = \sqrt{\bar{J}_{N_x, N_y}}$ .

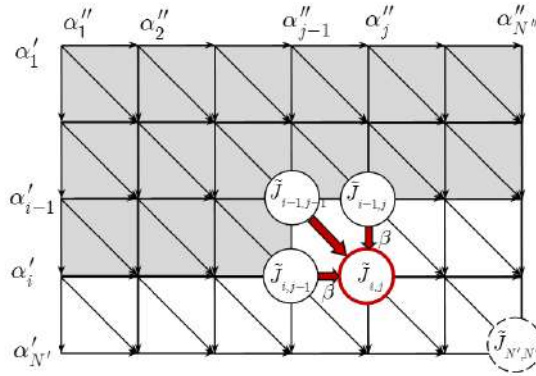


Fig. 3. A graph of pair correspondences of signal samples representing an electroencephalogram.

### 2.3. Calculation of the matrix of values of dissimilarity of electroencephalograms

As already mentioned above, at the stage of learning the computer algorithms for data analysis, it is necessary to calculate the matrices of the values of dissimilarity for all pairs of signals representing fragments of electroencephalograms from a certain training set  $X = \{x_1, x_2, \dots, x_K\}$ .

The matrix of pairwise dissimilarities calculated in accordance with the algorithm given in Section 2.2 is symmetric and contains zero values on the main diagonal, so it is sufficient to calculate only the values belonging to the upper (or lower) triangle (Figure 4), the number of which can be determined according to the expression  $K(K - 1)/2$ .

	$x_1$	$x_2$	$x_3$	...	$x_K$
$x_1$	0	$r_{1,2}$	$r_{1,3}$	...	$r_{1,K}$
$x_2$	$r_{1,2}$	0	$r_{2,3}$	...	$r_{2,K}$
$x_3$	$r_{1,3}$	$r_{2,3}$	0	...	$r_{3,K}$
...	...	...	...	...	...
$x_K$	$r_{1,K}$	$r_{2,K}$	$r_{3,K}$	...	0

Fig. 4. Matrix of values of pairwise dissimilarity of signals.

Since the number of computations has a quadratic dependence on the number of signals compared, even for a small volume of training set. In this work the training set consists of  $K = 755$  fragments of electroencephalograms (for each electrode) and so it is necessary to perform 284635 pairwise comparisons of signals to calculate one matrix. Thus, it is necessary to perform 56357730 pairwise comparisons of signals to calculate such matrices for all 66 electrodes and for three different values of the penalty for warping signal's axes. Since the time of one pairwise comparison requires, on average, 0.045 seconds, the calculation of all the matrices takes about 29 days, which makes it impossible to carry out experiments on real data and requires taking special measures to improve computing performance.

### 3. Parallel comparison of fragments of electroencephalograms

The task of computing several matrices of pairwise dissimilarity of fragments of electroencephalograms has several levels of data parallelism. Independently from each other can be calculated:

- 1) the incomplete values of the criterion, located on one minor diagonal of the graph of pairwise correspondences,
- 2) all elements belonging to the upper (or lower) triangle of the matrix of values of pairwise dissimilarity (Figure 4),
- 3) all matrices of pairwise dissimilarity.

Parallelization at the first level requires frequent interaction of processes or threads and is associated with the costs of synchronization, what can also lead to inefficient use of the cache [17-21] and, accordingly, does not provide the desired acceleration of computations.

Parallelization at the third level obviously makes sense only if the number of matrices that need to be calculated is greater than the number of available calculators. It also seems inappropriate, since in this study we focus on the computational capabilities provided by the supercomputer complex "Lomonosov" of Moscow State University [23], consisting of more than 5000 nodes and more than 12000 processor cores.



So, in the framework of this paper, parallelization is performed at the second level, i.e. the tasks of calculating the elements of a single matrix of pair-wise dissimilarity for a set of signals are identified as parallel tasks. In this case, if it is necessary to calculate several matrices, then they are computed sequentially.

MPI technology [21] has been chosen as a technology for organizing parallel computing, which allows to organize the interaction of processes running on different computing nodes.

In this case, since even for the Lomonosov supercomputer, the number of elements of each calculated matrix turns out to be much larger than the number of processors, then the parallel tasks are aggregated by means of a one-time distribution of the matrix elements between the processes.

In this paper the following scheme is used to distribute  $M = K(K-1)/2$  elements between the  $P$  processes: the first  $M \bmod P$  processes receive the  $(M - M \bmod P)/P + 1$  elements, and the remaining  $M - M \bmod P$  processes receive the  $(M - M \bmod P)/P$  elements, where mod is the modulo operation which gives the remainder after division of one integer by another.

This scheme allows us to distribute the work between processes as evenly as possible. The one-time even distribution of elements between the processes proves to be the most effective in this situation, since all compared signals have the same length and, accordingly, the comparison of any pairs of signals takes approximately the same time. As a result, this scheme allows to ensure the most efficient use of the resources of the computer system.

Graphical representation of the scheme of data distribution by processes is shown in Fig. 5.

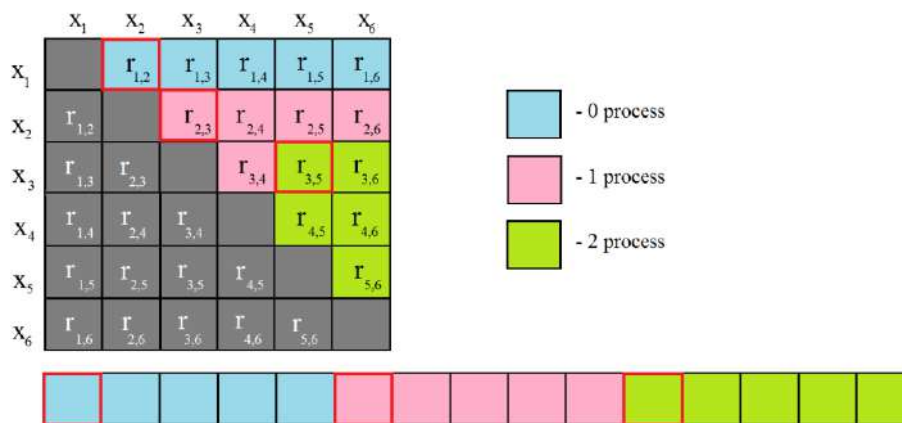


Fig. 5. Scheme of data distribution by processes for three processes.

According to the proposed scheme of parallel computing, each process independently of the others:

- 1) reads the original signals representing fragments of electroencephalograms from the input file,
- 2) determines the number of matrix's elements that it must process and the linear index of the initial element belonging to its range, according to the above-mentioned principle of uniform distribution of elements between processes,
- 3) calculates the row and column number determining the position of the given element in the matrix from the found linear index,
- 4) performs a sequential search of the consecutive elements of the matrix belonging to the upper triangle starting from the element defined in clause 2, at that performs for each element of its range a comparison of the corresponded signals (according to the sequential algorithm described in Section 2.2) and stores the calculated values in their copy of the dissimilarity matrix. Calculations continue until the number of elements obtained in step 2 is processed.

All results are merged on the process with the number 0 using the function `MPI_Reduce`, after each process computes its elements of the dissimilarity matrix.

#### 4. Experimental study

Brain electroencephalograms obtained in the course of the study described in [5] have been used to investigate the proposed algorithm.

In total, it is required to calculate three matrices (for different values of the penalty) of a pairwise dissimilarity of 755 signals for each of the 66 electrodes. I.e. it is required to calculate 198 matrices. However, the calculation of all matrices of dissimilarity takes approximately the same time and in this connection, it is enough to perform testing for one of the matrices.

The performance study was implemented using the resources of the MSU supercomputer complex "Lomonosov" [23]. Testing was conducted for different matrix sizes and different number of processes to identify possible behavioral features of the proposed algorithm. The results of time measurements in the calculation of one matrix of pair-wise dissimilarity of signals are given in Table 1.

The acceleration, which was obtained for each case, was calculated as the ratio of the sequential computation time to the calculation time on  $P$  processes. The results are shown in Table 2.

Table 1. Time of calculating the matrices of pairwise comparison of signals for different matrix sizes and different number of processes

Number of signals	Running time of the algorithm (sec)							
	Number of processes							
	1	2	4	8	16	32	64	128
10	2,14918	1,12535	0,701826	0,48118	0,21083	0,191736	0,205418	0,353755
20	9,04609	4,61894	2,38828	1,15214	0,619012	0,381232	0,297953	0,392734
40	37,1698	18,7533	9,42092	4,82778	2,34974	1,26236	0,822015	0,615929
100	232,34	117,278	58,3369	29,1305	14,696	7,72897	4,09282	2,46599
200	943,653	472,307	238,218	117,96	59,0047	30,2108	15,3994	8,34048
300	2124,3	1083,46	541,713	268,855	133,068	67,1879	34,5807	17,4532
500	5912,42	3010,1	1498,6	736,2	371,859	185,836	93,3412	49,3313
755	13675,7	6864,46	3418,42	1700,52	847,277	422,34	212,281	106,928

Table 2. Acceleration obtained in calculating the matrices of pairwise comparison of signals for different matrix sizes and different number of processes

Number of signals	Acceleration of the algorithm							
	Number of processes							
	1	2	4	8	16	32	64	128
10	1	1,909788	3,062269	4,466478	10,1939	11,20906	10,46247	6,075335
20	1	1,958477	3,787701	7,851554	14,61376	23,72857	30,3608	23,03363
40	1	1,98204	3,945453	7,69915	15,81869	29,44469	45,21791	60,34754
100	1	1,981105	3,982728	7,975833	15,80974	30,06093	56,76771	94,21774
200	1	1,997965	3,9613	7,999771	15,99284	31,23562	61,27856	113,1413
300	1	1,960663	3,921449	7,901285	15,96402	31,6173	61,43022	121,7141
500	1	1,964194	3,945296	8,030997	15,89963	31,81526	63,34202	119,8513
755	1	1,992247	4,000591	8,042069	16,14077	32,38078	64,42263	127,8963

Figure 6 shows graphs illustrating the data from Tables 1 and 2.

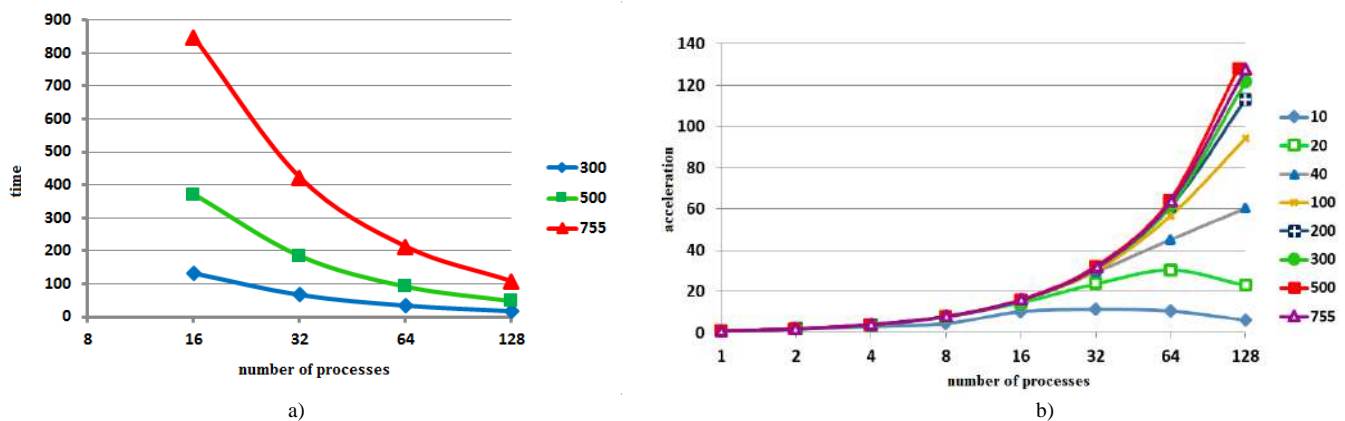


Fig. 6. Graphs of a) time dependence and b) acceleration of the algorithm from the number of processes.

As expected, the acceleration achieved by using the proposed parallel algorithm increases with the size of the calculated matrix of signal dissimilarity. And the acceleration is almost linear in the calculation of the total matrix.

Thus, the proposed approach launched on 128 processes of “Lomonosov” supercomputer complex allowed us to reduce the time of calculating one complete matrix of dissimilarity of fragments of electroencephalograms by 127.89 times. The calculation time for one matrix was reduced from 3.8 hours to 1.78 minutes, and the total calculation time of all 198 matrices was reduced from 29 days to 3.5 hours.

## 5. Conclusion

In this paper, a high-performance algorithm is proposed which calculates the matrix of the dissimilarity of signals representing fragments of brain electroencephalograms. The proposed algorithm was implemented in C++ programming language with using MPI parallel programming technology. It was tested using the resources of “Lomonosov” supercomputer

complex at Moscow State University [23]. Experimental studies implemented on real data have shown that the proposed algorithm allows to achieve an almost linear speedup and to reduce the total calculation time from 29 days to 3.5 hours by using 128 processes. And so, it opens the possibility of improving the quality of automatic analysis of electroencephalograms.

## Acknowledgments

This work is supported by the Russian Fund for Basic Research, grant 15-07-08967.

The authors would like to thank rector of Lomonosov Moscow State University Viktor Sadovnichiy and Moscow State University Supercomputing Center for providing “Lomonosov” supercomputer complex to perform experimental study.

## References

- [1] Zenkov LR, Zenkov KS. Clinical electroencephalography (with elements of epileptology). 3rd ed. Moscow: Publishing house MEDPRESS-INFORM, 2004; 368 p. (in Russian)
- [2] Teplan M. Fundamentals of EEG Measurement. *Measurement Science Review* 2002; 2(2): 1–11.
- [3] Wolpaw J, McFarland DJ, Neat GW, Forneris CA. An R. EEG-based brain-computer interface for cursor control. *Electroencephalography & Clinical Neurophysiology* 1991; 8(3): 252–259.
- [4] Tran L. EEG Features for the Detection of Event-Related Potentials Evoked Using Rapid Serial Visual Presentation. PhD Thesis 2014; 63 p.
- [5] Hope C, Sterr A, Langovan PE, Geades N, Windridge D, Young K, Wells K. High Throughput Screening for Mammography using a Human-Computer Interface with Rapid Serial Visual Presentation (RSVP). *Proc. SPIE 8673, Medical Imaging: Image Perception, Observer Performance, and Technology Assessment* 2013; 867303. DOI:10.1117/12.2007557
- [6] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1978; 26(1): 43–49.
- [7] Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. *Association for the Advancement of Artificial Intelligence, Workshop on Knowledge Discovery in Databases* 1994: 229–248.
- [8] Keogh E, Pazzani M. Scaling up dynamic time warping for datamining applications. *Proceedings of the sixth ACM SIGKDD intern. conf. on Knowledge discovery and data mining*. ACM Press, New York, NY, USA 2000: 285–289.
- [9] Vapnik V. *Statistical Learning Theory*. John-Wiley & Sons, Inc., 1998.
- [10] Tatarchuk A, Sulimova V, Mottl V, Windridge D. Supervised Selective Kernel Fusion for Membrane Protein Prediction. *Lecture Notes in Computer Science* 2014; 8626: 98–109.
- [11] Myers C, Rabiner LR, Rosenberg AE. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 1980; 28(6): 623–635.
- [12] Keogh E, Ratanamahatana C. Exact indexing of dynamic time warping. *Knowledge and Information Systems* 2004; 7(3): 358–386.
- [13] Lemire D. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern Recogn.* 2009; 42(9): 2169–2180.
- [14] Salvador S, Chan P. Toward accurate dynamic time wrapping in linear time and space. *Intelligent Data Analysis* 2007; 11(5): 561–580.
- [15] Al-Nayma G, Chawla S, Taheri J. SparseDTW: A Novel Approach to Speed up Dynamic Time Warping, 2012
- [16] Silva DF. Speeding up all-pairwise dynamic time warping matrix calculation 2016. URL: <http://sites.labic.icmc.usp.br/prunedDTW>
- [17] Lamport L. The parallel execution of DO loops. *Commun. ACM* 1974; 17(2): 83–93.
- [18] Babichev AV, Lebedev VG. Parallelization of program cycles. *Programming* 1983; 5: 52–63. (in Russian)
- [19] Fernandez A, Llberia JM, Valero-Garcia M. Loop Transformation Using Nonunimodal Matrices. *IEEE Transactions on Parallel and Distributed Systems* 1995; 6(8): 832–840.
- [20] Abu Khalil JM, Morylev RI, Shteinberg BI. Parallel Algorithm of Global Alignment with Optimal Memory Usage. *Modern problems of science and education* 2013; 1. URL: <http://www.science-education.ru/107-8139>. (in Russian)
- [21] Steinberg BI. Optimizing the use of the memory cache in computational tasks and optimizing compilation. *The All-Russian Scientific Conference on Informatics Problems SPISOK-2013, mekmat of St. Petersburg University, St. Petersburg, 2013*. (in Russian)
- [22] Antonov AS, Tutorial A. *Parallel Programming Using MPI Technology*. Moscow: MGU, 2004; 71 p. (in Russian)
- [23] Voevodin VIV, Zhumatiy SA, Sobolev SI, Antonov AS, Bryzgalov PA, Nikitenko DA, Stefanov KS, Voevodin VadV. *Practice of ‘Lomonosov’ Supercomputer*. Open Systems. Moscow: “Open Systems” Publishing house 2012; 7: 36–39. (in Russian)
- [24] Bellman R, Kalaba RM. *Dynamic Programming and Modern Control Theory*. Science 1969; 118 p. (in Russian)



# Software for heterogeneous computer systems and structures of data processing systems with increased performance

A.A. Kolpakov<sup>1</sup>, Ju.A. Kropotov<sup>1</sup>

<sup>1</sup>Murom institute (branch) VISU, Orlovskaya st., 23, 602264, Murom, Vladimir region, Russia

## Abstract

The issue of creating high-performance computing systems based on heterogeneous computer systems is topical, as the volumes of processed information, calculations and studies with large data sets are constantly increasing. The aim is to develop software design techniques heterogeneous computer data processing system. As a result, a technique has been developed for combining shaders to improve the performance of heterogeneous computations. The presented solution is supposed to be implemented as a software module for a computer system using CUDA technology.

*Keywords:* parallel computing; heterogeneous computing systems; graphics processors; CUDA; OpenCL

## 1. Introduction

The GPGPU program can be conditionally represented using the following sets:

- set of shaders that perform calculations;
- set of variables that control the computations;
- the set of data over which the calculations are performed and in which their results are recorded;
- a set of instructions that run a particular shader, provide him with input for certain data and output the result to a certain texture.

### 1.1 GPU programming based on vertex and pixel programs

The elementary primitive for visualization, with which the graphics processor works, is a triangle. With each vertex of a triangle, it is possible to associate a limited set of arbitrary data, for example, its color, normal, and other user data. Up to 8 textures can be associated with the primitive itself – one-, two-, and three-dimensional images. The structural scheme of data processing based on vertex and pixel programs is shown in fig. 1.

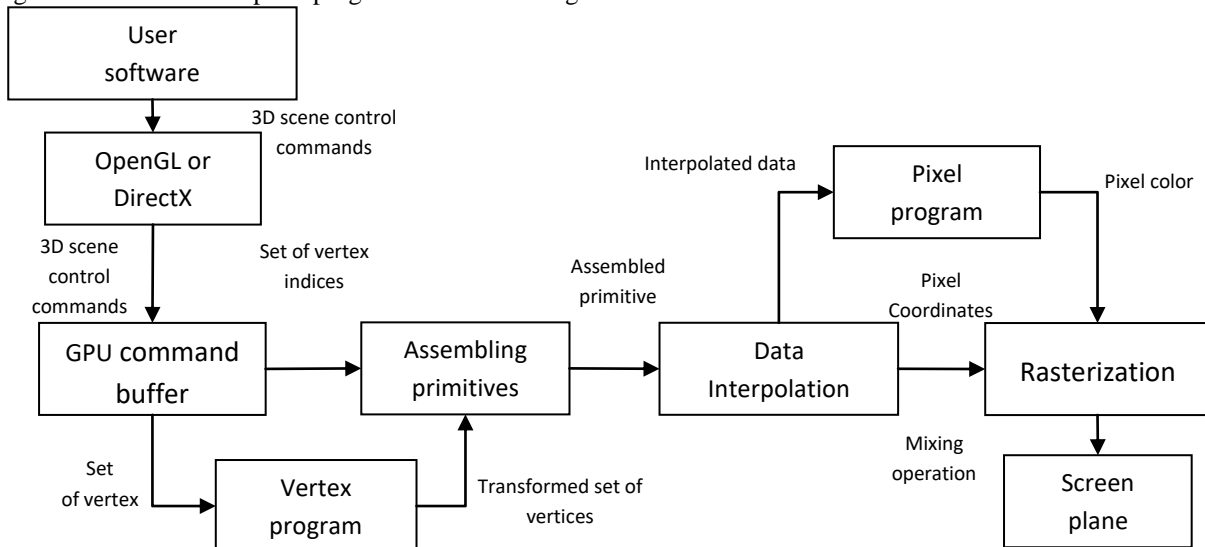


Fig. 1. The structural scheme of data processing based on vertex and pixel programs.

As shown in fig. 1, the user application sends requests for visualization of a 3D-scene to the OpenGL or DirectX low-level programming libraries, which are parts of the operating system. Then this data is transformed with the graphics card driver into direct commands of the graphics processor [1,2,3].

### 1.2 GPU programming based on CUDA library

Despite the fact that programming vertex and pixel programs has proven effective for modeling various physical processes, the use of this method is not convenient for the programmer. The programmer needs to have a sufficiently high qualification to perform computational tasks on the graphics processor, i.e. to use the principles of the GPU in detail, because a small inaccuracy

in the control code of the graphics processor can lead to a significant distortion of the result of the calculations. The structural scheme of the software on the basis of CUDA library is shown in fig. 2.

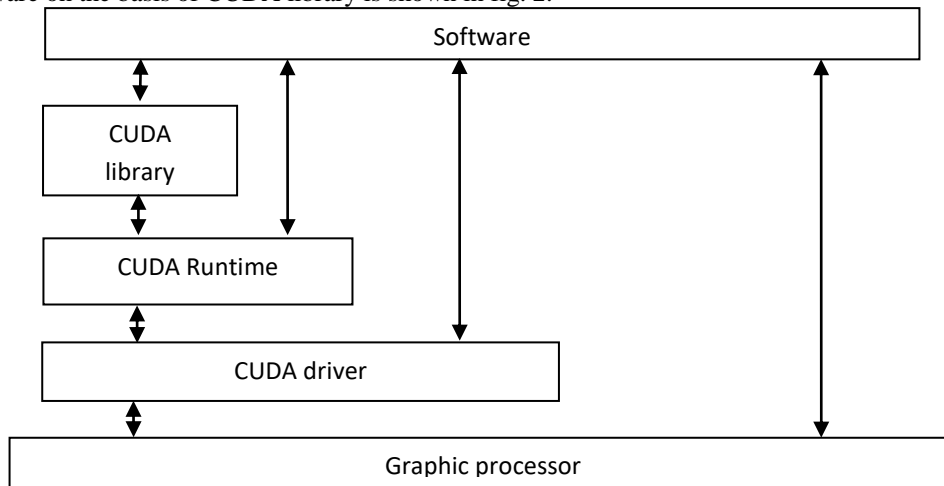


Fig.2. The structural scheme of the software on the basis of CUDA library.

The executable unit of the CUDA program is warp. The size of warp is 32 threads. This is due to the fact that latency is 4 cycles when executing one instruction on the multiprocessor. Only with respect to warp we can talk about the parallel execution of flows, no other assumptions can be made. However, this does not mean that warps are executed sequentially on the multiprocessor. The execution of warps can be parallel, for example, in the event that one warp expects data from global memory, other warps can be executed at this time.

Interaction between threads can be carried out only within the block. Data is exchanged via a shared memory which is common to all streams in the block. The execution of threads can synchronize by calling special synchronization functions.

### 1.3 GPU programming based on OpenCL library

The development of the GPU and its use in tasks unrelated to computer graphics has resulted in the development of a single standard for describing computations on highly parallel systems – OpenCL (Open Computational Library). The generated OpenCL library appeared on the basis of the previously developed Nvidia CUDA technology, which describes the interface of application interaction with the computing resources of the graphics processor. Unlike CUDA technology, OpenCL technology describes a computation model without connection to a specific type of device on which these calculations will be executed. Due to the fact that OpenCL is designed exactly as a standard for computations on highly parallel systems, many specific features of CUDA technology have been excluded from the standard. In general, CUDA technology has more possibilities, in comparison with OpenCL for describing parallel computing, if the Nvidia graphics processor is the computing device.

OpenCL allows to describe the calculations, abstracting from a particular device, on which these calculations will be implemented. In general, OpenCL algorithms can be executed on several CPU cores, on a graphics processor or on IBM Cell / B.E processors. OpenCL implementation uses extensions of the C language to describe the algorithm [3,4].

The OpenCL library is a promising library for use in various scientific research. The advantage of the OpenCL library is support from high-performance clusters. Any application that uses OpenCL can be run without modifications on a cluster that contains, among other things, graphics processors. This application will be available to all existing computing resources in the system. The structural scheme of the software on the basis of the OpenCL library is shown in fig. 3.

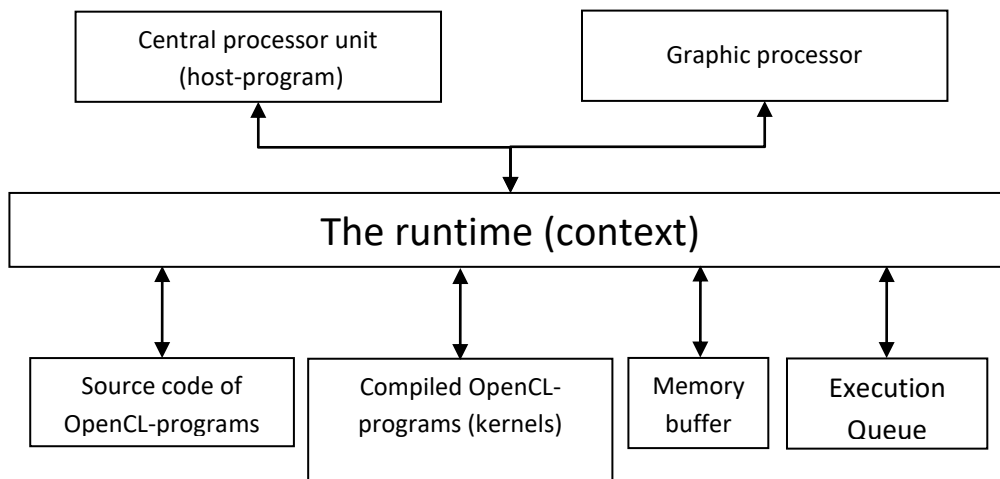


Fig. 3. The structural scheme of the software on the basis of the OpenCL library.

The central element of the OpenCL platform model is the concept of a host, the primary device that manages OpenCL calculations and performs all interactions with the user. The host is always represented in a single instance, while the OpenCL-devices on which the OpenCL-instructions are executed can be represented in the plural. OpenCL-device can be CPU, GPU, DSP or any other processor in the system, supported by the OpenCL-drivers installed in the system.

## 2. Software of a heterogeneous computer data processing system

The block diagram of the developed software of a heterogeneous computer data processing system is shown in fig. 4.

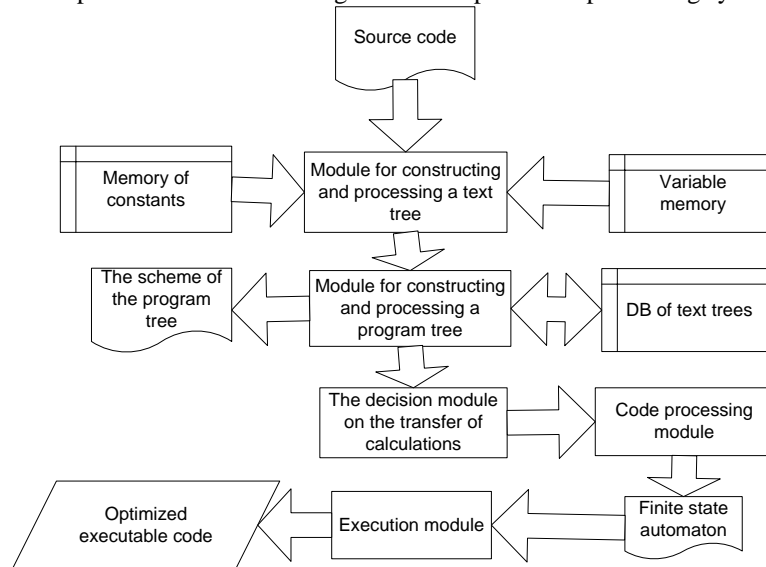


Fig. 4. The block diagram of the developed software of a heterogeneous computer data processing system.

According to fig. 4 the input source for the software is the source code of the program being processed. It enters the module for building and processing a text tree, in which the initial processing of the source code occurs and the construction of a text tree on its basis. Memory constants and variable memory are used to work with a text tree.

The received text tree is transferred for processing to the module for constructing and processing the program tree. The module for building a software tree uses a database of previously processed text trees, which allows to significantly accelerate the construction of a program tree [5,6].

For the processing of the program it is necessary to build a tree that represents a text description in a convenient format. There can be several types of nodes in the program tree:

- root node – contains all the data that must be computed during the execution of the program as descendants;
- node with data – represents an array of data that is required to transmit the input ancestor or into which to write the result descendant;
- node with variable – represents a variable that must be passed to the input of the ancestor shader or that describes the condition of the conditional parent operator;
- node with no operations – transfers the data of the descendant to the ancestor without changing them;
- node with shader – performs a shader with input data received from the children and passes the result to the ancestor;
- node with arithmetic operation – performs arithmetic transformation of input data received from descendants and passes the result to the ancestor;
- node with branching start – describes a conditional statement; the first child is a variable that describes which branch to choose, followed by the various branches of execution;
- node with branching completed – describes the completion of the conditional statement, all its branches are reduced to this node.

Trees are alternately built for each instruction. The parser parses the string representation, defines the output array with the data, optionally expands the abstractions and defines the set of functions that must be performed to obtain the result.

In each subsequent tree, the node with the data that was recorded in one of their previous instructions is replaced by the tree of the last instruction. The program tree is obtained after processing the last instruction. It should be noted that the software tree may not be a tree by definition - some nodes may have more than one ancestor, but most often it represents a tree or a structure close to it.

The software tree is transferred to the module of decision about the transfer of calculations, where the original algorithm is divided into stages and a decision is made to transfer the calculations to the GPU for each stage with further transfer to the code processing module.

### 3. Development of the code processing module

The code processing module makes the necessary changes to the source code program to produce a finite automaton of the program transmitted in the execution unit, generating executable code treated.

The level of the target executor for the family of accelerators takes on the input a graph of functional operations and on the output receives a program for the GPU [7,8]. This program has the form of a performance script, in which operations are available for starting the shader on the GPU with a certain set of parameters, allocating and freeing memory, transferring data from RAM to video RAM and back. The basic stages of broadcasting the program at the level of the target performer:

1. Splitting the original graph into subgraphs corresponding to different passes. Subsequently each subgraph is broadcast separately.

2. Select the display for the data and for the code. In this case, for example, one array of logical data can be mapped to several physical GPU buffers, and the restore operation in the source code is performed in a loop, with each iteration processing a block of 4 elements.

3. Code generation for the GPU. After the mappings are selected, they are generated by the code on the GPU. In particular, the indexes for the GPU buffers are calculated and the repetitive reads are eliminated.

4. Conversion of the generated code. At this stage, the arithmetic expressions are merged into vector commands and the constants are convoluted.

Since the individual expressions calculated on the GPU are relatively small, and the passage requires a large computing power of the shader, the first stage is relatively simple. Most often, the result of his work is the only subgraph that coincides with the original graph. The main criterion for partitioning into passages is the reuse of data. Splitting is only possible if you can not avoid multiple calculations of the same data without it.

The most difficult from the point of view of implementation are the stages 2 and 3. Step 4 is relatively simple and is performed using arithmetic transformations. The possibility of merging arithmetic expressions into vector operations is provided by an efficient choice of mappings in stages 2 and 3.

The functional graph C\$ defines the following types of vertices [9]:

Sheet tops:

1. Constants.
2. Related variables. In fact, are similar to the cycle. They run through a certain range and can be used to calculate the index expressions.

3. Functions. They can be member functions (including standard operations) or arrays. If the function is found expression, it is applied to its arguments, so in the end it will not be presenting.

Internal vertices:

5. Applying a function to arguments (@). This can be an application of a normal function or a reference to an array element.
6. Reduction operation (R). It has 2 arguments, the first of which specifies a reducing function, and the second one is reducible. Reduction is applied to all dimensions of an array that does not contain associated variables, as well as related variables, which allows partial reduction.

Associated variables "spread" from the bottom up the graph. They can end their distribution only at the vertices of reduction or at the root apex. We also assume that an acyclic graph is arranged in such a way that all paths of propagation of a bound variable terminate on the same vertex.

Below is an example of combining two shaders, in which you can see the non-optimal parts of the code. The block diagram of the first shader is shown in fig. 5.

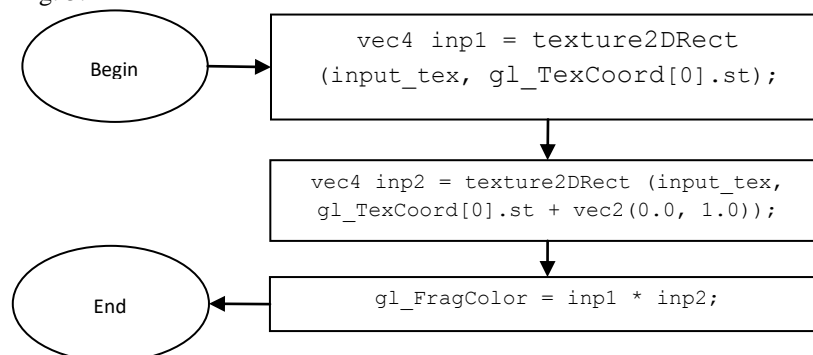


Fig. 5. Block diagram of the first shader.

A block diagram of the second shader is shown in fig. 6.

The block diagram of the combined shader, obtained after merging the sequence of these shaders, is shown in fig. 7.

To perform the shader more efficiently, it is necessary to perform its correction. A block diagram of the final adjusted shader is shown in fig. 8.

After carrying out all the above-described changes, the shader code will be compiled and will be ready for use in calculations.

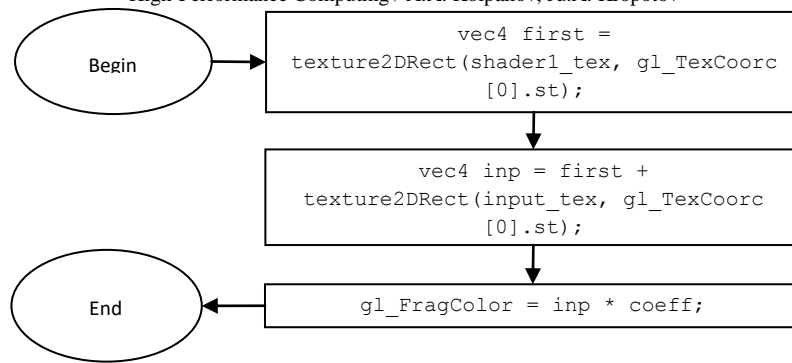


Fig. 6. Block diagram of the second shader.

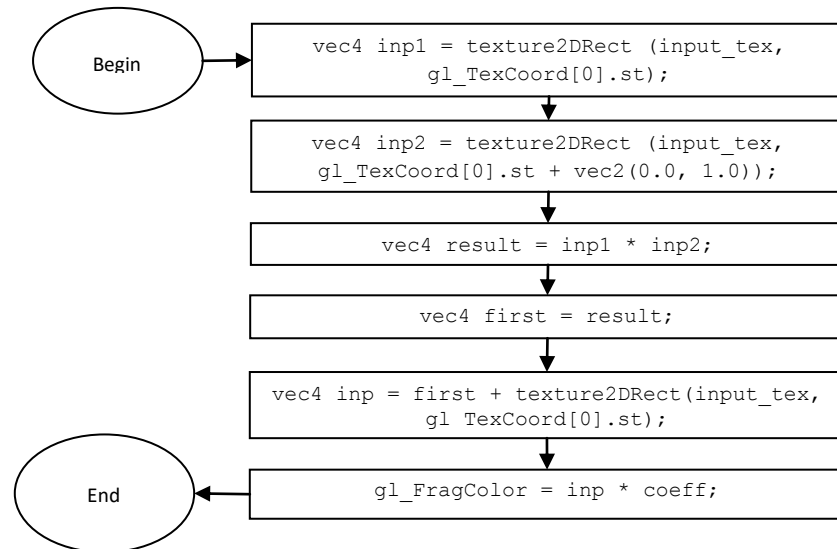


Fig. 7. Block diagram of the combined shader.

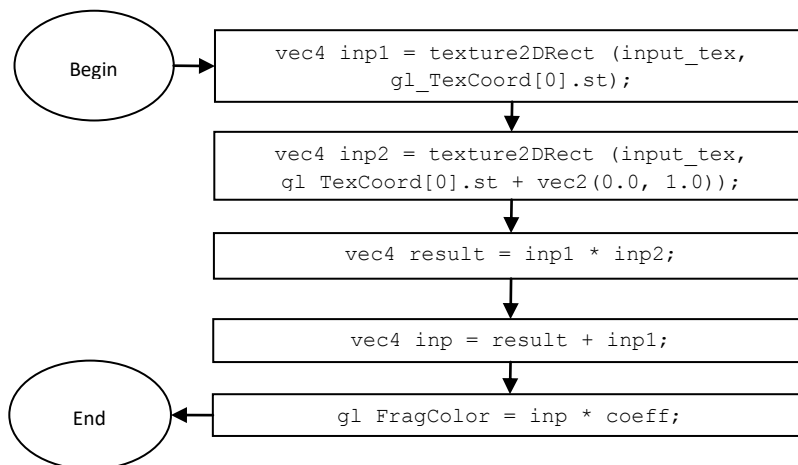


Fig. 8. Block diagram of the resulting shader.

#### 4. Conclusion

The program provides the schemas of the processed program tree and the generated automaton and the received shaders as output. The scheme of the processed program tree is shown in Fig. 9.

As can be seen from fig. 9, the software tree consists of nodes of various types, each of which describes the behavior of the program: the root node; node with data; node with variable; node with no operations; node with a shader; node with arithmetic operation; node with the beginning of branching; node with the completion of branching.

The code for the graphics processor is generated in accordance with the selected display. Each key vertex is associated with a set of output variables, for the calculation of which it responds. The root node, in addition to their calculation, is responsible for writing them to the output buffers. For each sub-graph of operations, a code template is generated. Based on this template, a code is generated for each set of values of the block measurement instances [10].



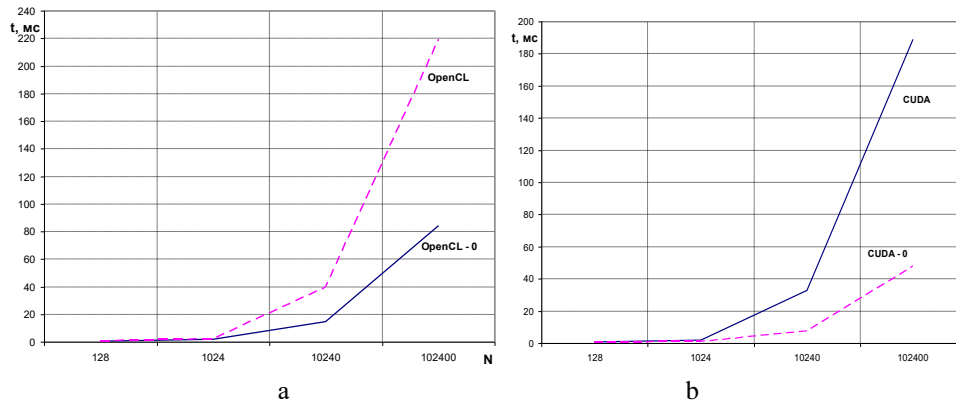


Fig. 10. The average time spent by a parallel system to receive a new generation, with  $M = 10$ , a: OpenCL – basic algorithm, OpenCL-O – developed algorithm, b: CUDA – basic algorithm, CUDA-O – developed algorithm.

## References

- [1] Bahvalov NS, Voevodin VV. Modern Problems of Computational Mathematics and Mathematical Modelling. Computational Mathematics Book. Vol. 1. Moscow, Science, 2005; 342 p. (in Russian)
- [2] Borekov AV. Shaders development and debugging. Sankt-Peterburg: BHV-Peterburg, 2006; 488 p. (in Russian)
- [3] Kolpakov AA. Theoretical evaluation of growth performance computing systems from the use of multiple computing devices. V mire nauchnykh otkrytii 2012; 1: 206–209. (in Russian)
- [4] Kolpakov AA. Optimizing the use of genetic algorithms for computing graphics processors for the problem of zero bit vector. Informatsionnye sistemy i tekhnologii 2013; 2(76): 22–28. (in Russian)
- [5] Barkalov KA, Gergel VP. Parallel global optimization on GPU. Journal of Global Optimization 2016; 66(1): 3–20.
- [6] Strongin RG, Gergel VP. Parallel computing for globally optimal decision making on cluster systems. Future Generation Computer Systems 2005; 21: 673–678.
- [7] Galimov MR, Biryalcev EV. Some technological aspects of GPGPU applications in applied program systems. Vychislitelnye metody i programmirovaniye 2010; 11: 77–93. (in Russian)
- [8] Kropotov JuA, Belov AA, Proskuryakov AJu, Kolpakov AA. Methods of Designing Telecommunication Information and Control Audio Exchange Systems in Difficult Noise Conditions. Sistemy upravleniia, sviazi i bezopasnosti 2015; 2: 165–183. (in Russian)
- [9] Borekov AV. OpenGL extension. Sankt-Peterburg: BHV-Peterburg, 2005; 672 p. (in Russian)
- [10] Ermolaev VA, Kropotov JuA. Algorithms for processing acoustic signals in telecommunication systems by local parametric methods of analysis. Proceedings International Siberian Conference on Control and Communications (SIBCON), 2015. URL: <http://ieeexplore.ieee.org/document/7147109/>.

# Advanced mixing audio streams for heterogeneous computer systems in telecommunications

A.A. Kolpakov<sup>1</sup>, Ju.A. Kropotov<sup>1</sup>

<sup>1</sup>Murom institute (branch) VLSU, Orlovskaya st., 23, 602264, Murom, Vladimir region, Russia

---

## Abstract

This paper presents an algorithm enhanced mixing of audio streams for computation on GPUs, which combines multiple stages of mixing by using two-pass rendering, which significantly reduces the switching time between buffers. Methods of computer experimental comparative studies were carried out evaluating the performance of the developed algorithm. The purpose of the present paper is development of an efficient algorithm for mixing audio streams for processing on GPUs. The method of transfer operations computing on graphics processors with the use of Shader programs was developed. Novel features presented solutions is using a two-pass rendering. The results showed that the application of the developed algorithm leads to an increase in computational performance up to 6 times. Presented solution can be implemented as software in the telecommunications multiprocessor systems.

*Keywords:* two-pass rendering; algorithm for increasing productivity; parallel computing; heterogeneous computing systems; graphics processors; mixing of audio data

---

## 1. Introduction

The development of information and telecommunication systems of digital in-house operational, command, loudspeaker and telephone communication on their basis becomes especially urgent with the development of modern computer networks. It is justified to use panel or tablet computers based on ready-made solutions to ensure optimum performance when designing such systems. Since voice communication plays an important role, the voice conferences mode with several participants is the main one in such systems.

The simplest scenario of organizing voice communication is that each sound source sends its audio stream to each receiver independently. This method is simple and convenient, but it requires high network bandwidth, which is not always possible. Therefore, the best method is audio mixing, which means combining the audio streams from each source into one. Based on the possibility of sound waves superimposed on each other, this method can provide an acceptable sound quality, while ensuring a decrease in network congestion. However, the application of this method can significantly increase the load on the server CPU, which can negatively affect the overall system performance. The way out of this situation may be the use of GPUs to solve this problem [1].

## 2. The problems with using graphics processors for audio mixing

Although the graphics processors are quite efficient, there are several problems in using them for audio mixing, which are related to the architecture and the limited functionality [2].

The first problem is the bus bandwidth between the GPU and the main memory, which is smaller than between the main processor and the main memory. For example, the Intel 975X chipset provides theoretical bandwidth for a CPU of 10.7 GB/s, and for the GPU only 8 GB/s. Practice shows that the lack of support for asynchronous I/O requires a lot of time for additional operations, such as locking / unlocking the buffer. Since the general calculations on the GPU are based on 3D rendering, the write speed is usually higher than the read speed. This asymmetry makes the procedure for reading the result sufficiently long [3, 4].

Secondly, the general calculations on the GPU are based on 3D models, different tasks require different GPU settings, such as 3D models, transforming matrices and shader programs. During the loading of the settings, the computational flows of the GPU are not involved. Worst of all, the GPU does not tell the CPU when the task is completed, so the CPU must periodically check the status of the GPU. This is quite a time-consuming operation, as it breaks the parallelism between the GPU and the CPU.

Third, the disadvantage of the GPU is its performance in logical operations. As you know, CPU tracks branches, GPU works differently: each branch is first executed, and then the desired result is selected. This makes parallelization easier, but requires more resources.

Finally, the GPU instruction set is incompatible with the CPU. In addition, execution time and code length are limited. All this makes it difficult to transfer existing algorithms to graphics processors [5].

## 3. The structure of the developed algorithm

The basic algorithm of audio mixing consists of five steps. The first step is to summarize the audio samples from different sources, which can be represented by the following formula [6,7]:



$$\vec{M}_t = \sum_{k=0}^n \vec{u}_{k,t}, \quad (1)$$

где  $\vec{u}_{k,t}$  – the sample vector; i.e. the vector of samples obtained by microphone  $k$  in time  $t$ ;

$\vec{M}_t$  – resulting mixing vector.

The second stage is echo cancellation, which in its basic form consists in excluding the sample of the  $i$ -th device from the final vector [8]. This stage is represented in the form

$$\vec{M}_{i,t} = \vec{M}_t - \vec{u}_{i,t}, \quad (2)$$

где  $\vec{M}_{i,t}$  – resulting mixing vector for the  $i$ -th device;

$\vec{u}_{i,t}$  – sample of the  $i$ -th device.

For the correctness of the resulting vector  $\vec{M}_{i,t}$ , it is necessary that its dimension be equal to the dimension of the incoming vectors. However, after stages 1 and 2, the vector  $\vec{M}_{i,t}$  may be overcrowded, leading to unwanted noise [9]. In order to use the original vector of large dimensions  $\vec{M}_{i,t}$ , it is necessary to compress it for further use. This is done in the third stage by the formula

$$\vec{M}_{i,t} = \vec{M}_{i,t} \times frac_{i,t}, \quad (3)$$

where  $frac_{i,t}$  – attenuation factor for the  $i$ -th device. This coefficient must be calculated automatically, starting from the maximum compressed sample. The search for the maximum among the compressed samples occurs in the fourth stage, and the correction of the attenuation coefficient is made on the fifth stage.

A block diagram of the basic mixing algorithm is shown in fig. 1.

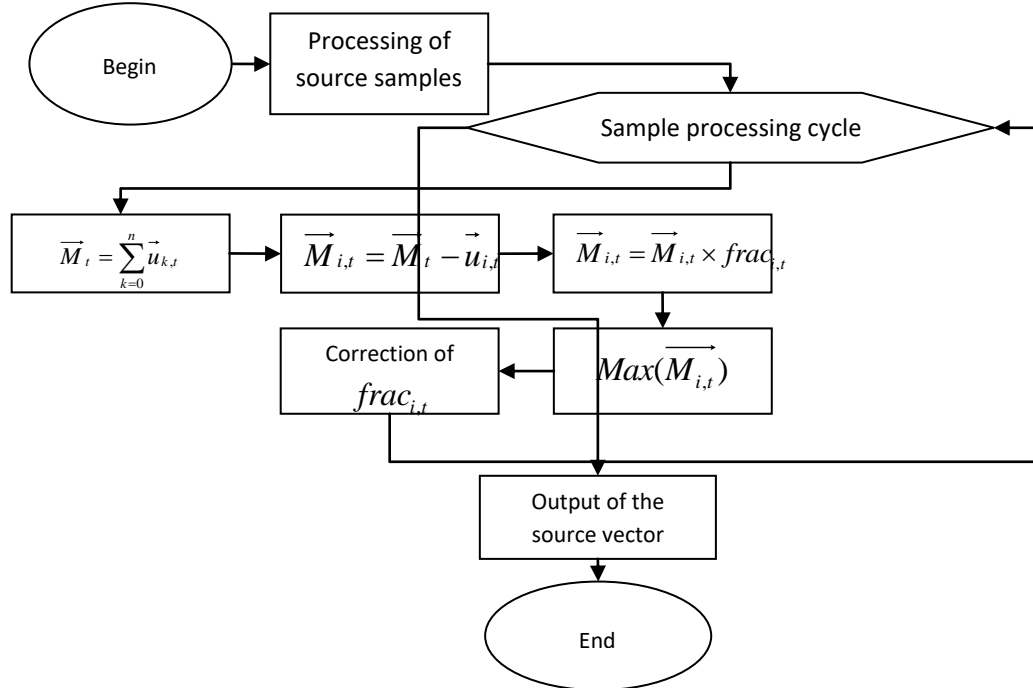


Fig. 1. A block diagram of the basic mixing algorithm.

For a better description of the algorithm, the GPU capabilities are represented as an  $f$ -function of the  $X$  texture and pixel coordinates represented by the formula

$$f(X, P) = \{y_i \mid \forall p_i \in P_i y_i = f(X, p_i)\}, \quad (4)$$

где  $y_i$  – projection of pixel  $p_i$  on the texture  $X$ ,

$P$  – 3D model projection.

For each pixel in the 3D model projection  $P$ , the function (4) will be calculated and then the result will be written to the render buffer.

From the formula (4) the sample of calculations on GPU can be presented as  $A=(X,P,f)$ .

Suppose that  $n$  is the total number of sound sources in the session, and  $L$  – the length of one audio sample. Each of the first three mixing steps (sample accumulation, echo cancellation and compression) yields  $n$  sequences of  $L$  bytes. At the same time, the last two steps (searching for the maximum sample and adapting the attenuation coefficient) yield only  $n$  integers. Since the first three steps can be performed within one projection for computing on GPU, they can be combined into one step, as shown below

$$m_i = \left( \sum_{j=0}^{n-1} u_j - u_i \right) \times frac_i. \quad (5)$$

Since the fourth step uses a different measurement than the first three, it cannot be combined within the framework of formula (5). A block diagram of the extended mixing algorithm is shown in fig. 2.

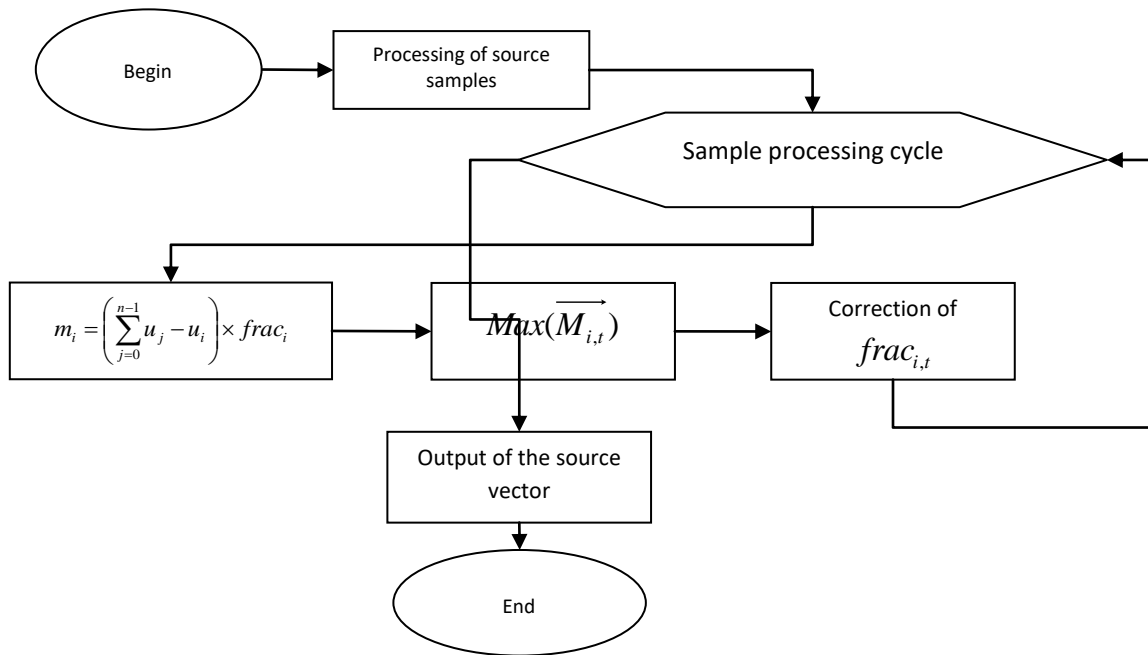


Fig. 2. A block diagram of the extended mixing algorithm.

Usually, calculations can be combined according to general criteria when their projections do not coincide and they do not have intersections. Therefore, if these calculations are denoted as  $A1=(X1, P1, f1)$  and  $A2=(X2, P2, f2)$ , then they can be combined as shown below.

$$X = \langle X_1, X_2 \rangle, P = P_1 \cup P_2, f(\langle X_1, X_2 \rangle, p) = \begin{cases} f_1(X_1, p), p \in P_1, \\ f_2(X_2, p), p \in P_2. \end{cases} \quad (6)$$

As seen from (6), the main algorithm is checking the coordinates of each pixel to select the module's execution function. Since the GPU executes all branches before selecting the desired one, each branch will be executed for each pixel, which will take a long time.

The developed algorithm presents an alternative method for executing a large number of functions, which outputs data of different lengths to a single render buffer using multithreaded rendering. Here the main rule is to move the projection by modifying the projection matrix, so that the computational area of each function is limited to the required area, rather than the entire buffer.

Depending on the pixel, it can have a different amount of information without loss of versatility.  $w$  denote the total number of pixels needed to store  $L$  bytes. Thus, the render buffer can be represented as  $(w+1)*n$ . 3D model of the developed algorithm is a rectangle, which lies in the plane  $Z$ . It has the dimensions:  $2w/(w+1)$  units in width and 2 units in height. Coordinates of vertices  $-(-1,-1,0), (1-2/(w+1),-1,0), (-1,1,0), (1-2/(w+1),1,0)$ .

At the first rendering pass, a unit matrix is chosen as the projection matrix. This ensures that the projection matches the 3D model. After converting the visible area, in this pass the formula (5) is applied to perform the first three steps of the mixing. Transformations of the 3D model, produced in the first three steps, are shown in Fig. 3.

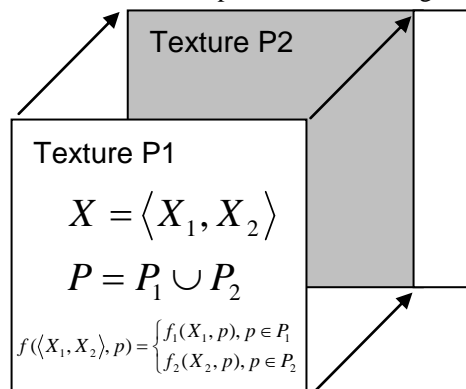


Fig. 3. The first pass of the algorithm.

In the second rendering pass, the projection is shifted to the left by  $L$  pixels. At the same time, the shader program switches to the search mode of the maximum sample. In this pass, only one column can be written, since most parts of the projection lie outside the render buffer and will be automatically ignored by the GPU. Since the clipping of the projection was performed at the beginning of the render, this method calls the function only for the correct pixels, and not for the entire buffer. The actions performed on the second pass of the algorithm are shown in fig. 4.

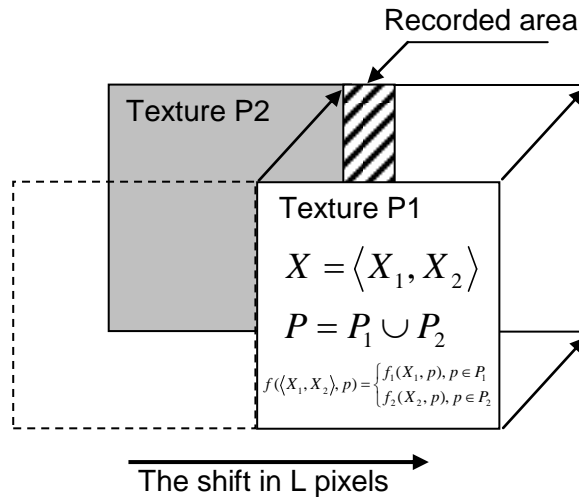


Fig. 4. The second pass of the algorithm.

#### 4. Using a single-texture algorithm

As noted above, both pass algorithms requires as input data samples every  $n$  sequences. Each sequence must have its own unique texture. Total needed  $n$  texture dimension  $L$ . This multi-textural technology is not suitable for audio mixing. First, each sound source requires its own texture. Therefore, additional passes may be required. Secondly, loading a lot of small textures is much slower than loading a single large one.

In the developed algorithm, it is proposed to load a single texture. As an example, the first pass of the algorithm is presented. The input contains  $n$  samples of sequences of size  $L$  bytes and  $n$  attenuation coefficients. Two textured RGBA format buffers are used. Texture T1 has the dimension  $[L/4]*n$  and stores all sample sequences in a line. Texture T2 has a dimension of  $1*n$  and stores the attenuation coefficients. The coordinates of all the textures are given in Table 1.

Table 1. Coordinates of textures used in the developed algorithm.

Vertexes	Coordinates of texture T1	Coordinates of texture T2
(-1, -1, 0)	(1/2w, 1)	(0.5, 1)
(1-2/(w+1), -1, 0)	(1, 1)	(0.5, 1)
(-1, 1, 0)	(1/2w, 0)	(0.5, 0)
(1-2/(w+1), 1, 0)	(1, 0)	(0.5, 0)

Audio mixing is performed independently for each pixel. The pixel's texture coordinates  $(x, y)$  are calculated by interpolating the vertex texture coordinates. Based on Table 1, the texture coordinate for T1 or pt.t0 should be  $(X, Y)$ , and for T2 or pt.t1 –  $(0.5, Y)$ . Pt.t0 refers to the sample for which the current mixing transformations are made, this sample is called the "aiming point". To exclude the appearance of an echo, the other texture coordinate of ptCur is restricted by accessing T1 instead of pt.t0. The  $x$  component of the ptCur texture is identical to the pt.t0 texture, and the  $y$  component is calculated from a cyclic variable that designates each sound source. In the loop, the position register  $v$  is used to skip the "aiming point". Finally, the attenuation coefficients are read from the texture pt.t1. Since the coefficient is stored in the first byte, the sample is produced only by the blue component.

#### 5. Experimental study of the developed algorithm

Test input data for mixing are sequences of 320 samples. All samples are generated randomly. Test bench configuration: CPU Intel Core i3-4130, 4 GB RAM, graphics card NVIDIA GeForce GT730. The number of  $M$  sequences was varied. The results for the basic and developed algorithms at the output are identical. The results of an experimental study of the dependence of the mixing time  $t$  on the number of sequences  $M$  are shown in fig. 5.

As can be seen from the test results shown in Fig. 5, application of the advanced algorithm of audio mixing allows to increase productivity of computer system in 5-6 times [9].

Table 2 shows the results of measuring the average running time of the algorithm  $t$  for 8 random sequences of 320 samples. During the measurement, 1000 processing cycles were carried out, for each of which new random sequences of samples were

generated. For the results obtained, the dispersion of the output sequences is calculated, the values of which are also given in Table 2.

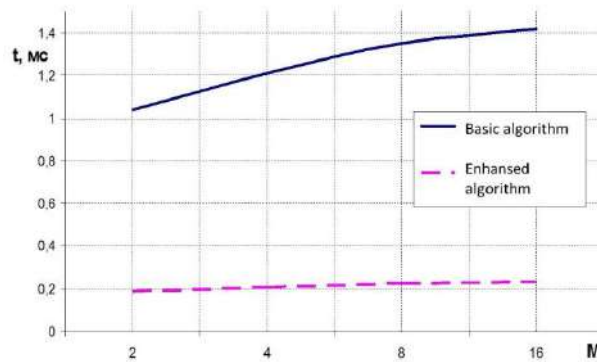


Fig. 5. The results of an experimental study of the dependence of the mixing time  $t$  on the number of sequences  $M$ .

Table 2. Results of the study of the algorithm for extended audio mixing for 8 sample sequences.

The used algorithm	Average time of the algorithm, ms	Dispersion of output Sequences
Basic algorithm	1,351	3,757
Developed algorithm	$2,226 \times 10^{-1}$	$1,426 \times 10^{-3}$

As can be seen from the test results presented in Table 2, the variance of the output sequences for the basic algorithm is substantially higher than for the developed one [12,13]. This is due to the fact that 32-bit numbers are used for the calculations in the GPU, whereas for the calculations on the central processor – 64-bit. Application of the developed algorithm in a heterogeneous computer system reduces the processing time to  $0.22226 \times 10^{-3}$  s instead of  $1.351 \times 10^{-3}$  s – the time of data processing by the basic algorithm.

## 6. Conclusion

This paper presents an algorithm for the extended downmix audio streams for computing on GPU. Its main advantage is the combination of multiple stages of mixing by using a two-pass rendering, which significantly reduces the switching time between buffers. The use of one texture for calculations increases the efficiency of I/O operations. Although I/O operations take approximately half the computation time, experimental studies of the developed algorithm showed an increase in performance up to 6 times.

## References

- [1] Lindholm E, Nickolls J, Oberman S, Montrym J. NVIDIA Tesla: A unified graphics and computing architecture. *IEEE Micro* 2008; 28(2): 39–55.
- [2] Luebke D, Harris M, Kruger J, Purcell T, Govindaraju N, Buck I, Woolley C, Lefohn A. GPGPU: general purpose computation on graphics hardware. *ACM SIG-GRAPH*. New York, USA: Course Notes, 2004; 33 p. DOI: 10.1145/1103900.110393.
- [3] Kolpakov AA. Theoretical evaluation of growth performance computing systems from the use of multiple computing devices. *V mire nauchnykh otkrytii* 2012; 1: 206–209. (in Russian)
- [4] Borekov AV. Shaders development and debugging. Sankt-Peterburg: BHV-Peterburg, 2006; 488 p. (in Russian)
- [5] Nekrasov KA, Potashnikov SI, Boyarchenkov AS, Kupryazhkin AY. Parallel computing for general purpose graphics processors: text book. Ministry of Education and Science of the Russian Federation, Ural Federal University. Ekaterinburg : Publishing house of the Ural University, 2016; 104 p. (in Russian)
- [6] Kropotov JuA. Experimental study of the law of distribution of probability of amplitudes of signals of systems of transmission of voice information. *Proektirovanie i tekhnologiiia elektronnykh sredstv* 2006; 4: 37–42. (in Russian)
- [7] Belozyorov AS, Korobicyn VV. Implementation of computations on a graphics processor using Nvidia CUDA platform. *Programmnye produkty i sistemy* 2010; 1: 62–64. (in Russian)
- [8] Kropotov JuA, Belov AA, Proskuryakov AJu, Kolpakov AA. Methods of Designing Telecommunication Information and Control Audio Exchange Systems in Difficult Noise Conditions. *Sistemy upravleniia, sviazi i bezopasnosti* 2015; 2: 165–183. (in Russian)
- [9] Ermolaev VA, Kropotov JuA. Algorithms for processing acoustic signals in telecommunication systems by local parametric methods of analysis. *Proceedings International Siberian Conference on Control and Communications (SIBCON)*, 2015. URL: <http://ieeexplore.ieee.org/document/7147109/>.

# The algorithm for a video panorama construction and its software implementation using CUDA technology

I.A. Kudinov<sup>1</sup>, O.V. Pavlov<sup>1</sup>, I.S. Kholopov<sup>1,2</sup>, M.Yu. Khramov<sup>1</sup>

<sup>1</sup>Ryazan State Instrument-making Enterprise, Seminarskaya str. 32, 390000, Ryazan, Russia

<sup>2</sup>Ryazan State Radio Engineering University, Gagarina str. 59/1, 390005, Ryazan, Russia

---

## Abstract

A video panorama constructing algorithm based on information from five different types pre-calibrated cameras with partially overlapping fields of view was developed and implemented using the CUDA C language. Distortion compensation, image stitching on the virtual unit sphere surface, and blending procedures are performed for the operator-controlled 1024×768 pixels region of interest with 50 fps.

**Keywords:** video panorama; camera calibration; distortion compensation; spherical panorama; region of interest; inclinometer; blending; CUDA technology

---

## 1. Introduction

The automatic generation of high-resolution video panoramas from information of cameras with partially overlapping fields of view (FoV) is one of the modern trends in the vision systems development. Generally, panorama navigation implies the presence of a user-controlled region of interest (RoI) with the predefined angular FoV dimensions and resolution [1]. In avionics, for example, the advantages of panorama systems in comparison with traditional electro-optical systems are [2, 3], at first, the possibility of simultaneous use of a panorama field by several independent operators, and, secondly, the absence of mechanical parts.

## 2. Problem statement

A panoramic image from  $N$  frames with overlapping (or pairwise overlapping) FoVs is formed by finding a correspondence between the pixel coordinates of each frame. This correspondence for the camera frames with the numbers  $i$  and  $j$  is determined by the 3×3 dimension homography matrix  $\mathbf{H}_{ij}$  [4]:

$$\mathbf{x}_i = \mathbf{H}_{ij}\mathbf{x}_j, \quad (1)$$

where the matrix transformation (1) performs the recalculation of the  $j$ -th camera image homogeneous pixel coordinates into the  $i$ -th camera coordinate system,  $\mathbf{x}_j = [u_j, v_j, 1]^T$  and  $\mathbf{x}_i = [u_i, v_i, 1]^T$  are homogeneous pixel coordinates of  $i$ -th and  $j$ -th images, and  $(u, v)$  are coordinates of pixel which is located at the intersection of  $u$ -th row and  $v$ -th column.

There are several approaches to the panorama construction. In the absence of a priori information, the estimation of the homography matrix  $\mathbf{H}_{ij}$  is based on  $m$  interest points (IP) detection, building descriptors of the neighborhood for each IP and their automatic matching. In this case homography matrix  $\mathbf{H} = [[h_1, h_2, h_3]^T, [h_4, h_5, h_6]^T, [h_7, h_8, 1]^T]$  estimation is performed by pseudosolution (in terms of minimum mean square error) of overdetermined system of equations for  $m \geq 4$  inliers:

$$\mathbf{A}_{(2m \times 9)}\mathbf{h}_{(9 \times 1)} = \mathbf{0}_{(2m \times 1)},$$
$$\mathbf{A}_{(2 \times 9)} = \begin{bmatrix} -u_j & -v_j & -1 & 0 & 0 & 0 & u_i u_j & u_i v_j & u_i \\ 0 & 0 & 0 & -u_j & -v_j & -1 & v_i u_j & v_i v_j & v_i \end{bmatrix}. \quad (2)$$

The pseudosolution of (2) is the last column-vector of matrix  $\mathbf{V}$  (which is the result of the matrix  $\mathbf{A}_{(2m \times 9)}$  singular value decomposition) corresponding to the minimal singular value  $\Sigma_{\min}$ :

$$\mathbf{A}_{(2m \times 9)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad \mathbf{h} = \mathbf{V}^{<\Sigma_{\min}>}.$$

The pay for the universality of this approach to panorama construction is a low efficiency on homogeneous surfaces (grass, arable land, forest, water surface, sky), and under low contrast conditions, and when the overlapping of cameras FoVs is small.

With priori information about the mutual position of the camera's coordinate systems obtained during the preliminary calibration, the homography matrix can be estimated by the formula:

$$\mathbf{H}_{ij} = \mathbf{K}_i[\mathbf{R}_{ij} - \mathbf{t}_{ij}\mathbf{n}^T/d]\mathbf{K}_j^{-1}, \quad (3)$$

where  $\mathbf{K}_i$  and  $\mathbf{K}_j$  are intrinsic matrices,  $i, j = 1..N$ ,  $\mathbf{R}_{ij}$  and  $\mathbf{t}_{ij}$  are respectively a rotation matrix and translation vector for transition from  $j$ -th camera coordinate system to  $i$ -th camera coordinate system,  $d$  – the perpendicular length to the shooting

plane with the normal  $\mathbf{n}$  in the reference ( $i$ -th) camera coordinate system. If the distance to the observed objects is large ( $\|\mathbf{t}_{ij}\| \ll d$ ), then we have the following approximate equality from (3):

$$\mathbf{H}_{ij} \approx \mathbf{K}_i \mathbf{R}_{ij} \mathbf{K}_j^{-1}. \quad (4)$$

While combining information from several cameras, it is advisable to form the resultant panoramic image not in the plane in accordance with (1), but on a virtual uniform curvature surface: usually a unit sphere or a cylinder with a unit radius [5]. It allows us to work with normalized homogenous pixel and spatial coordinates. The geometric problem statement for combining  $N = 3$  camera frames with intersecting FoVs on the unit sphere surface (the coordinate system of camera with number 0 is taken as reference) is shown at Fig. 1. It is expected that by analogy with (4) all nodal points of the camera lenses are located in the unit sphere center  $O$ .

In the Fig. 1 the symbol  $\mathbf{M}$  denotes the point of the object spatial homogeneous coordinates, which image is projected onto the point on the unit sphere surface with spatial coordinates  $\|\mathbf{M}_{\text{sph}}\| = 1$  in the coordinate system of the sphere  $O X_{\text{sph}} Y_{\text{sph}} Z_{\text{sph}}$ , and the symbols  $\mathbf{x}_0$  and  $\mathbf{x}_2$  are the homogeneous pixel coordinates of its image on the frames of cameras with numbers 0 and 2 respectively.

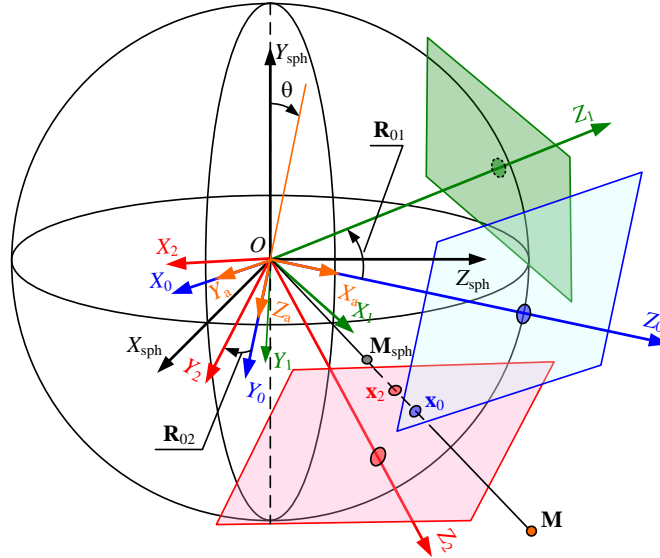


Fig. 1. Unit virtual sphere for panorama forming.

The correct spherical panoramas construction implies the availability of information about the angular position of the optical axes of the cameras in relation to the horizon plane. This problem is solved in our work by mounting of inclinometer (based on the triaxial MEMS accelerometer) on the reference camera case. An accelerometer coordinate system is designated as  $O X_a Y_a Z_a$  (Fig. 1). The roll  $\varphi$  and pitch  $\theta$  angles are estimated by the formulas [6]:

$$\varphi = \text{atan2}(a_y, a_z), \quad \theta = \text{atan2}[-a_x, (a_y^2 + a_z^2)^{0.5}], \quad (5)$$

where  $[a_x, a_y, a_z]^T$  – a vector of accelerometer measurements (taking into account its calibration [7]).

The choice of pixels in areas where the radius-vector  $\mathbf{OM}_{\text{sph}}$  crosses several camera planes can be fulfilled by several criteria: for example, the maximum distance from the pixel to the intersection line of the camera planes or the minimum length of the normalized pixel coordinates vector. In order to minimize computations we choose a criterion of minimum angle between a vector to a pixel and the  $i$ -th camera principal axis, as such pixels, usually, have the minimal distortion correction error:

$$\min_i \left[ (\mathbf{R}_{0i} \mathbf{R}_{\varphi\theta})^{\langle 3 \rangle} \cdot \mathbf{M}_{\text{sph}} \right], \quad (6)$$

where

$$\mathbf{R}_{\varphi\theta} = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} - \quad (7)$$

is a rotation matrix of reference camera coordinate system relative to the horizon plane,  $\mathbf{R}_{0i}$  – estimated during calibration rotation matrix of  $i$ -th camera relative to the reference camera,  $\langle k \rangle$  –  $k$ -th column of matrix, and symbol  $\langle \cdot \rangle$  is a dot product.

To realize the sliding RoI function with size  $W \times H$  pixels with angular dimensions in the horizontal and vertical directions  $\Delta\varphi_w$  and  $\Delta\varphi_h$  respectively we introduce a quaternion  $\mathbf{q}_{vis}$  [8] that is defined by the current line of sight azimuth  $\alpha$  and the elevation angle  $\beta$ :

$$\mathbf{q}_{vis} = [\cos(\alpha/2)\cos(\beta/2), \cos(\alpha/2)\sin(\beta/2), \sin(\alpha/2)\cos(\beta/2), \sin(\alpha/2)\sin(\beta/2)]^T. \quad (8)$$

To each pixel of the RoI with coordinates  $(u, v)$  corresponds the point  $\mathbf{M}_{uv} = [x_{uv}, y_{uv}, z_{uv}]^T / \|[x_{uv}, y_{uv}, z_{uv}]^T\|$  on the unit sphere (Fig.2), determined by the radius-vector with the corresponding quaternion

$$\mathbf{q}_{uv} = \mathbf{q}_{vis}\mathbf{q}_{uv0}, \quad (9)$$

where from geometric constructions of Fig. 2 the initial position ( $\alpha = \beta = 0, z_{uv} = 1$ ) of the RoI pixels  $(u, v)$  corresponds to quaternion

$$\mathbf{q}_{uv0} = [\cos(\alpha_u/2)\cos(\beta_v/2), \cos(\alpha_u/2)\sin(\beta_v/2), \sin(\alpha_u/2)\cos(\beta_v/2), \sin(\alpha_u/2)\sin(\beta_v/2)]^T, \quad (10)$$

$$\alpha_u = \arctg(x_{uv}), \quad \beta_v = \arcsin[y_{uv}/(x_{uv}^2 + y_{uv}^2 + 1)^{0.5}],$$

$$x_{uv} = (2u/W - 1)\text{tg}(\Delta\varphi_w/2), \quad y_{uv} = -(2v/H - 1)\text{tg}(\Delta\varphi_h/2).$$

Quaternion  $\mathbf{q}_{uv}$  allows us to determine coordinates of the point  $\mathbf{M}_{uv}$  in the unit sphere coordinate system [8]:

$$\mathbf{M}_{uv} = [2(q_x q_z + q_w q_y), 2(q_y q_z - q_w q_x), q_w^2 + q_z^2 - (q_x^2 + q_y^2)]^T, \quad (11)$$

where  $q_w$  and  $[q_x, q_y, q_z]^T$  are scalar and vector parts of quaternion  $\mathbf{q}_{uv}$ .

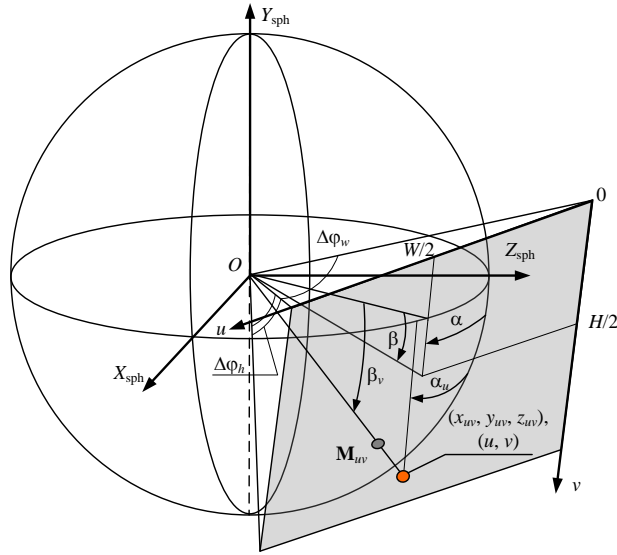


Fig. 2. The relationship between the angular and spatial coordinates of the point  $\mathbf{M}_{uv}$  on the virtual unit sphere surface.

On the each  $i$ -th camera principal plane the homogenous pixel coordinates  $\mathbf{x}_{uvi}$  corresponding to the point  $\mathbf{M}_{uv}$  are determined (taking into account the orientation relative to the horizon plane) through the projection matrix  $\mathbf{P}_i$  [4]:

$$\mathbf{x}_{uvi} = \mathbf{P}_i \mathbf{M}_{uv}, \quad (12)$$

where  $\mathbf{P}_i = \mathbf{K}_i[\mathbf{R}_{0i}\mathbf{R}_{\varphi\theta} | \mathbf{0}]$ ,  $\mathbf{0} = [0, 0, 0]^T$ , and the selection of the pixel transferred from the  $i$ -th camera plane to the RoI is performed by the criterion (6).

For the cameras with wide-angle lenses it is necessary to do the distortion compensation: for the normalized coordinates with distortion

$$\mathbf{x}_{hi} = \mathbf{K}_i^{-1} \mathbf{x}_{uvi} \quad (13)$$

coordinates without distortion are estimated according to the Brown-Conrady model [9] (in order to reduce the amount of calculations in our work only the first two coefficients of radial distortion are used). Corrected pixel coordinates without distortion are calculated by multiplying on the intrinsic matrices  $\mathbf{K}_i$ :

$$\mathbf{x}_{\text{corr}i} = \mathbf{K}_i \mathbf{x}_{hi} \quad (14)$$

and since it is fractional so the brightness value is interpolated (in our work we used a bilinear interpolation [10]).

Since the scenes of each camera are different, then in the automatic exposure time mode the average brightness of the composing panorama frames is different too. Therefore, after panorama filling (or the RoI filling) it is additionally necessary to perform a smoothing procedure for brightness differences – blending. We used a simplified approach to blending, similar to the work [11].

### 3. The algorithm for a video panorama construction and its software implementation

The algorithm for forming the RoI image contains the following steps.

1. Initialization: calculating quaternions  $\mathbf{q}_{uv0}$  by formula (10).

Main operation cycle:

2. Estimation of current angular position of reference camera by (5) and its rotation matrix  $\mathbf{R}_{\varphi_0}$  by (7).
3. Quaternion  $\mathbf{q}_{\text{vis}}$  calculation for the current line of sight angular position by (8) and quaternions  $\mathbf{q}_{uv}$  by (9).
4. Filling the RoI (with distortion compensation) according to (11)-(14) by criterion (6).
5. Blending procedure performing (optional).

As the processing for each RoI pixel is homogeneous this allows us to parallelize computations. In the layout based on a PC for implementing the algorithm steps 1, 4, and 5 we used the resources of the NVIDIA GPU: using CUDA technology and CUDA C language the whole amount of computations was distributed to 3072 parallel blocks (64 horizontally and 48 vertically) with 256 threads in each block (16 horizontal and vertical threads). In this case, according to [12], the pixel indices are represented as:

$$u = \text{blockIdx.y} * \text{blockDim.y} + \text{threadIdx.y}; \quad v = \text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x}; \quad (15)$$

In (15) the following standard CUDA notation is used: blockIdx is a block identifier (number), blockDim – a block dimension, threadIdx – an identifier (number) of a parallel thread, and attributes "x" and "y" indicate the axis of the coordinate system in a two-dimensional Euclidean space.

As the copying from CPU memory to GPU memory and back is rather slow [12], so by the implementing of the algorithm the number of such operations is minimized: by the initialization GPU memory arrays are recorded with the data on the camera parameters (intrinsic matrices and distortion coefficients) and initial quaternions  $\mathbf{q}_{uv0}$  (10) that correspond to RoI pixels. In the main operation cycle frames from each camera, the line of sight angular coordinates and the angular coordinates of reference camera, estimated from MEMS signals, are copied into the GPU memory, then the steps 3-5 of the algorithm are performed and the result (array of brightness values in the RoI) is copied back to the CPU memory for displaying on the PC monitor.

### 4. Results and Discussion

The layout of the video panorama system is implemented on a PC with Intel Core-i5 processor and NVIDIA GeForce GTX 560 Ti GPU (with 384 cores) and consists of five digital GigE interface video cameras, mounted on a special rigid polyamide bracket (Fig. 3): two Basler acA2000 cameras with 2048×1088 frame resolution and 5 mm megapixel Cowa lenses (from below) and three IDS 5240 RE cameras (from above) with 1280×1024 frame resolution and 5 mm megapixel Computar lenses (two outer cameras) and 8 mm Navitar lens (central reference camera), an evaluation board with a InvenSense MPU 9250 MEMS sensor mounted on the reference camera, two motorized rotation positioners Standa 8MR190-2 for moving the bracket with cameras in horizontal and vertical planes, and USB joystick Defender Cobra M5 for the RoI navigation.



Fig. 3. General view of the video panorama forming layout.

The layout allows to create a panorama with a 3600×2400 pixels resolution and 180°×120° FoV or display RoI with user-defined resolution and angular FoV dimensions.



Preliminary camera calibration was performed using OpenCV libraries [13]: on 40 test "chessboard" pattern frames, containing  $12 \times 11$  cells with a side of 3 cm, the intrinsic matrices and distortion coefficients of their lenses were estimated; then on 40 test "chessboard" pattern frames, containing  $9 \times 6$  cells with a side of 3 cm, the rotation matrices of cameras relative to the reference camera were estimated.

The influence of the angular orientation estimation error of the reference camera relative to the horizon plane on the geometry of the panorama is shown in Fig. 4. As can be seen from the figure, the pitch estimation error leads to an incorrect horizon display (the position of the spherical panorama equator is shown by the orange line).

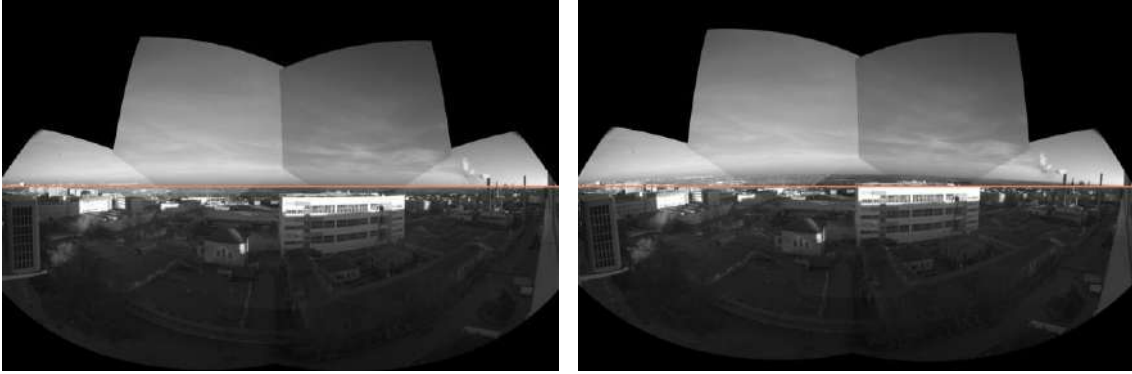


Fig. 4. The influence of the pitch estimation error on the spherical panorama geometry (without blending): left – the pitch of the reference camera coincides with the true,  $\theta = 15^\circ$ , right – the pitch of the reference camera doesn't coincide with the true,  $\theta = 12^\circ$ .

The results of the  $1024 \times 768$  pixels RoI forming with a  $40^\circ \times 30^\circ$  FoV and the line of sight angular coordinates  $\alpha = 15^\circ$  and  $\beta = -5^\circ$  with blending and without it are shown in Fig. 5 and 6 respectively. The times required to create one frame on the CPU and GPU resources are summarized in Table 1 (the time for copying the data from the CPU memory to the GPU memory and back is 3.5 ms), and the times for different RoI sizes are summarized in Table 2.

Table 1. Time required for the  $1024 \times 768$  pixels RoI forming, ms.

	CPU	GPU with CUDA	Gain
Without blending	114	6.5	17.5
With blending	272	17.2	15.8

Table 2. Time required for the RoI forming on GPU with CUDA for different resolution, ms.

RoI resolution, pixels	800×600	1024×768	1280×1024
Without blending	5.1	6.5	9.7
With blending	12.9	17.2	23.4
CPU ↔ GPU copy time	2.9	3.5	3.7

In Fig. 7 and 8 are shown the results of intermediate computations for blending implementation: the boundaries of the camera frames in the RoI (Fig. 7) and the 100-pixel width binary mask, over which the brightness is smoothed (Fig. 8).



Fig. 5. RoI without blending.



Fig. 6. RoI with blending.

Increasing of the RoI forming time with blending more than 2 times is explained by the need to perform auxiliary procedures: searching the intersection lines of camera frames in the RoI, determining areas for brightness fusion and smoothing images to estimate the low-frequency component of brightness in the each camera frame. To reduce the amount of calculation, we apply a smoothing procedure with an  $8 \times 8$  window and an accumulator. Its computational complexity is  $O(n^2)$ .

For  $\|t_{0i}\| < 15$  cm and distance to objects  $d > 70$  m the hypothesis about camera lenses nodal points superimposition provides the error of stitching on the camera frame boundaries not more than 5 pixels.

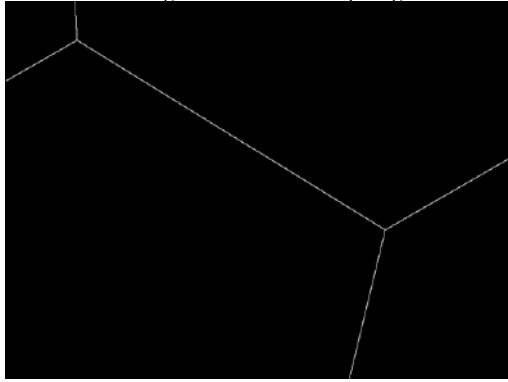


Fig. 7. The boundaries of the camera frames in the RoI (binary image).



Fig. 8. Blending binary mask obtained from Fig. 7.

The results of the experiment showed that the use of GPU resources makes it possible to reduce the RoI forming time up to 16 times in comparison with the CPU (when only one CPU core is used).

## 5. Conclusion

The algorithm for a video panorama construction, designed and implemented with the CUDA technology and parallelization of computations on a GPU, for a one megapixel resolution region of interest provides panorama navigation with 50 fps on the whole 8.2 megapixels panoramic video frame with  $180^\circ \times 120^\circ$  field of view.

## References

- [1] Banta B, Donaldson G. Apparatus and method for panoramic video hosting. Patent US US9516225, 2016.
- [2] AN/AAQ-37 Distributed Aperture System (DAS) for the F-35. URL: <http://www.northropgrumman.com/Capabilities/ANAAQ37F35/Pages/default.aspx> (01.10.2016).
- [3] IronVision™. URL: <http://elbitsystems.com/media/IronVision.pdf> (10.09.2016).
- [4] Hartley R, Zisserman A. Multiple view geometry in computer vision. 2<sup>nd</sup> edition. Cambridge: Cambridge University Press, 2003; 656 p.
- [5] Szeliski R. Image alignment and stitching: a tutorial. Foundations and trends in computer graphics and vision 2006; 2(1): 1–104.
- [6] Tilt sensing using a three-axis accelerometer. URL: [http://www.freescale.com/files/sensors/doc/app\\_note/AN3461.pdf](http://www.freescale.com/files/sensors/doc/app_note/AN3461.pdf) (11.04.2016).
- [7] Wang L, Wang F. Intelligent calibration method of low cost MEMS inertial measurement unit for an FPGA-based navigation system. International J. of Intelligent Engineering and Systems 2011; 4(2): 32–41.
- [8] Kuipers JB. Quaternions and rotation sequences. New Jersey: Princeton University, 1998; 400 p.
- [9] Brown DC. Close-range camera calibration. Photogrammetric Engineering 1971; 37(8): 855–866.
- [10] Parker JA, Kenyon RV, Troxel DE. Comparison of interpolating methods for image resampling. IEEE Trans. on Medical Imaging 1983; 2(1): 31–39.
- [11] Xiong Y, Pulli K. Fast image stitching and editing for panorama painting on mobile phones. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 13-18 June 2010: 47–52.
- [12] Sanders J, Kandrot E. CUDA by example. New York: Addison-Wesley, 2010; 290 p.
- [13] Camera Calibration and 3D reconstruction. URL: [http://docs.opencv.org/2.4/modules/calib3d/doc/camera\\_calibration\\_and\\_3d\\_reconstruction.html](http://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html) (14.03.2015).

# Various algorithms, calculating distances of DNA sequences, and some computational recommendations for use such algorithms

B.F. Melnikov<sup>1</sup>, S.V. Pivneva<sup>2</sup>, M.A. Trifonov<sup>2</sup>

<sup>1</sup>Center of Information Technologies and Systems for Executive Power Authorities, 19, str. 1, Presnenskiy Val, 123557, Moscow, Russia  
<sup>2</sup>Togliatti State University, 14, Belorusskia st., 445020, Togliatti, Russia

---

## Abstract

In recent years, various approaches have been described to determine the similarity of the DNA sequences; each of which defines a metric for the set of DNA sequences. In this paper, we propose a new approach to solving this problem; moreover, algorithms for its implementation are based on multiheuristic approach to the discrete optimization problems previously developed by us. However, the main focus of this article is to describe our original approach to compare the quality of defined metrics on the set of DNA sequences. The last approach is based on the fact, that the triples of distances between genomes should ideally form isosceles acute triangles. On the basis of this assumption, we proposed value of the norm, gives in practice acceptable results; the validity of this approach is also discussed in the article. In the course of work on the implementation of algorithms have been carried out computational experiments with 100 DNA of “distant” species, as well as with representatives of several genomes of great apes and humans. Several possible standards defined comparative quality algorithms describing metric distances on DNA sequences.

Thus, the main focus of this article is to describe our original approach to compare the quality of defined metrics on the set of DNA sequences. The approach is based on the fact, that the triples of distances between genomes should ideally form isosceles acute triangles. On the basis of this assumption, we proposed value of the norm, gives in practice acceptable results. In the course of work on the implementation of algorithms have been carried out computational experiments with 100 DNA of “distant” species, as well as with representatives of several genomes of great apes and humans.

*Keywords:* metric evaluation; algorithms; multiheuristic approach; original approach to compare the quality of defined metrics on the set of DNA

---

## 1. Introduction

The problem of determining the similarity of DNA is a special case of non-exact matching sequences[1]. The “non-exacting” (“mistaking”) is that when comparing lines it is possible to identify similar sequences, despite the errors and distortions in them, for example, changing, deleting, or inserting some characters. The amount of such distortion sets metric on the set of rows, which is determined by the minimum number of edit operations that provide a single line of another. This problem occurs in many areas. For example, a comparison of the genes and chromosomes of proteins is a major challenge and one of the main tools of molecular biology and bioinformatics [1,2,3,4,5,6,7]. The exact comparisons of nucleotide chains (and also computing distances using such comparisons) are unacceptable because of errors in the data, and due to possible mutations. Inaccurate mapping is carried out like the text processing. One of the metrics obtained by comparing the words (Levenshtein distance) is used to correct errors, to enhance recognition of scanned documents, to search in the information systems and databases [1]. To find an approximate solution, there are different algorithms in different subject areas, for example, to search a database of genetic information is widely used algorithm BLAST ([2]etc.), approximating Needleman-Wunsch algorithm.

Thus, in Section 2 of this paper we describe the application to the defining similarity of DNA sequences so called multiheuristic approach [8,9], which is in fact the big extension of the branch and bound method. Note that previously, before our works, branch and bound method, apparently, was not used to solve this problem.

Thus, the calculation of the distance (metric) between the rows of DNA of different species of organisms is one of the most important tasks of modern bioinformatics. As already noted, today there are many algorithms allowing to make an approximate calculation of the polynomial time ([4,5,6,7,10] and many others). The obvious disadvantage when calculating the distance between the one and the few lines of DNA is to provide different results when using different algorithms to calculate metrics. However, the authors do not know the work, which compared to a variety of algorithms for solving this problem. In this regard, one of the tasks that are discussed in this article was to develop a method for the comparative evaluation of such algorithms; moreover, this problem seems to be the most important our consideration. As a result, we have proposed a method of evaluation using the properties of an isosceles triangle in a metric space (Section 3, so called “triangular norm”, we calculate using it so called badness, related to some metric for several species).

We are also looking at options to improve some existing metrics. In this case, none of the methods we are considering the construction of the distance between the strands of DNA is not a disadvantage to use it to evaluate the distances between the neighbors as species (pairs “human – chimpanzee” and “human – bonobo” etc.), and between more distant species (pairs “human – crocodile” and “chimpanzee – crocodile” etc.). This is because we consider first of all corners of the triangles in the Euclidean space. However, we have made some computational experiments connected with the application for conversion metrics of continuous monotone functions (Section 4).

Brief results of computational experiments over 100 genomes are considered in Section 5. Among these results, it is worth noting the following. Firstly, for the “distant” species badness is very small, which indicates that the right choice of our approaches and relevant specific algorithms; in this case the fact is true for a number of different metrics. Secondly, as for the “distant” species we proposed approach to the definition of the metric gives the best results (all considered “triangular”

standards), among 5 considered metrics [4,5,6,7,10]. For the “near” species (human and apes), the results are somewhat worse (value of badness is increased, and, besides, our version of the metric gives 2nd for the quality of the result). Third, it is unlikely any of these metrics are appropriate for determining the distance between the subspecies: so that the application of these algorithms to the human race sometimes arises violation of the triangle inequality. The accurate explanations of recent facts, apparently, should lead biologists, but we also try to explain them, from our point of view.

Some possible areas for further work already ongoing by our group at the moment are summarized in Conclusion (Section 6).

## 2. Algorithm for determining the distance between nucleotide sequences based on the multiheuristic approach

As we said before, the multiheuristic approach to the problems of discrete optimization was considered [8,9] and in many other following papers. In this section, we describe its version for determining the distance between nucleotide sequences. For this problem, it was used as follows<sup>1</sup>. Let  $x, y$  be corresponding strings,  $i, j$  be indexes of considered symbols of strings  $x$  and  $y$  correspondently,  $r$  be the value of metric to be found. By shifting the line, we mean increasing by 1 of the corresponding index. The general scheme of the algorithm can be described as follows.

```

Input:          strings x and y.
Step 1:         i := 0, j := 0, r := 0;
Step 2:         if x[i] = y[j] then begin
                  shift both lines;
                  r := the cost of matching of symbols x[i] and y[j];
                end
                else begin
                  apply heuristics for generating
                  possible "trajectories" of the shift
                  in the position of i' and j',
                  such that x[i'] = y[j'];
                  evaluate them with other heuristics;
                  average these estimates using risk functions;
                  make shifting
                  (value can be changed);
                end;
Step 3:         repeating the second step until
                it reaches the end of one of the lines.

```

The cost of matching two symbols in a simplest case equals to 1; for DNA, it can be defined using some table of amino acid replacement costs, e.g. BLOSUM, see [1, 2, 11].

For this algorithms, the following heuristics were used.

1. We select such trajectories that the value  $(i' - i) + (j' - j)$  is minimum, or close to minimum. E.g. we first lookup all the trajectories with one string shifted by one symbol; next with one string shifted by two symbols or both strings shifted by one symbol, etc.
2. We shift a string, which current symbol found less frequent in the other string. For this heuristics it's preferable to know probabilities of appearance of a given symbol in each of the strings. If those probabilities are not known a priori, we consider them being equal. While following the algorithm we can adjust those probabilities or use aging algorithm [12], such that probability of a given symbol will be defined by some fragment of a string instead of a whole string. If probabilities for both strings are equal, we shift a string in which more symbols are left.
3. Combination of previous heuristics (1 and 2); to calculate the position using second heuristics we sum probabilities of finding other string for all symbols that will be passed by a shift.
4. Use of an algorithm of a longest common subsequence search for  $x[i..i+k]$  and  $y[j..j+k]$ , where  $k \sim 15$ . For shift we use  $i', j'$ , at which the longest common subsequence ends. If no common subsequence found, the search range is increased. When using this heuristics the result is close to the longest common subsequence value.
5. Combination of 3 and 4; the position  $(i', j')$  given by forth heuristics is a ratio of length of the longest common subsequence of strings  $x[i..i']$  and  $y[j..j']$  to an average shift length from  $(i, j)$  to  $(i', j')$ .
6. We use algorithm [13, 14] for strings  $x[i..i+k]$  and  $y[j..j+k]$ , where  $k \sim 15$ , then shift to  $(i', j')$ , having the greatest value in Needleman-Wunsch table.

Combination of 3 and 6; the position  $(i', j')$  given by sixth heuristics is a ratio of a value in Needleman-Wunsch table, corresponded to that position, to average shift length from  $(i, j)$  to  $(i', j')$ .

## 3. The decision-making using some different greedy heuristics simultaneously

Thus, we shall to use not only the heuristics, which forms branch-and-bounds method (BBM) for the considered discrete optimization problem (DOP), but also the other one. Namely, it is heuristics for selecting element, which separate the considered problem into right and left sub-problems. However, we can often replace a heuristics separating algorithm for some other. Moreover, solving a DOP using BBM, it is often desirable to choose one of some separating algorithms, depend upon the solved sub-problem. The various separating algorithms can be selected depend upon dimension of the solved problem, its bound, and also many other descriptions, which are based on the solved DOP.

<sup>1</sup> We have changed here the description of the algorithm given in [10]. The authors are willing to send me the source code when prompted by email.

In classical examples of using BBM for TSP, some good separating algorithms were used (it means, that they give the reasonable results comparing other ones). However, long before, for the BBM-branching various other heuristics were used. Let us mention, for example, the following heuristics for the reduced TSP-matrix: total number of zeroes, sum of minimums for all the rows and columns, sum of some minimum values of considered row and columns multiplied by special “damnation constants”; all these values are computed by the TSP-matrix after reducing and selecting separating element (i.e., separating edge for branching). Probably, the author mentioned here less than 10% of the heuristics used before.

Thus, how can we use the fact, that in various situations (i.e., in various sub-problems of the same considered DOP) different heuristics relatively better are used? (This question can be putted for both exact and unfinished algorithms.) We need to make a decision for selecting the separating element for branching. We have information of various experts, i.e., of various special heuristics, so called predictors (or estimators). The predictors often give discrepant information, and we have to average it in a special way. Unlike all the algorithms published before, the authors, like programming nondeterministic games, use dynamic risk functions for this thing.

Since various heuristics give values of various units, we have to normalize them for computing the final result. Using the special set of normalizing coefficients (it is adjusted, for example, for genetic algorithms, we shall consider them below) is, probably, a possible method, but it is not the main one for this paper. The authors used only one algorithm in various DOP; that was a special modification of “voting method”, where each of heuristics gives the considered variants of selecting (e.g., for traveling salesman problem, those are zeroes of the reduced matrix, corresponding some edges, which are the candidates for branching). After that, we use special dynamic risk functions for the results of voting.

Thus, selecting edge for branching is constructed for BBM by using dynamic risk functions; see previous sections for details. It is important to note, that the dynamic selecting of the particular risk function is similar to selecting it in programming nondeterministic games, considered in. Since we consider here DOP (not programming nondeterministic games), we have to add here a new heuristics, i.e., one for selecting “current position estimation”, i.e., evaluation of the situation, which is obtained by the solving some DOP using BBM.

Thus, let us have some various heuristics for selecting the element for the next step of BBM (or, generally speaking, for selecting the strategy of solution some DOP). Let each of possible strategies have some various expert evaluations of availability (i.e., let us have some independent expert sub-algorithms, i.e., predictors). Then the concluding strategy could be chosen by maximum of average values. However, let us consider the following example [24]; this example is connected with backgammon programming, because it uses 36 predictors).

Let expert evaluations of availability have values in the segment  $[0,1]$ . Let for the 1<sup>st</sup> strategy, the 1<sup>st</sup> expert has the evaluations of availability being equal to 1, and other 35 experts have the evaluation 0.055. And for the 2<sup>nd</sup> strategy, 2 experts have the evaluation 0.95, and other 34 experts have the evaluation 0. Very likely, each user (expert-human) on basis of these values chooses the 2<sup>nd</sup> strategy.

However, averaging-out by the simplest algorithm (i.e., simply the simple average of expert evaluations) gives 0.081 for the 1<sup>st</sup> case and 0.053 for the 2<sup>nd</sup> one; i.e., do we have to choose the 1<sup>st</sup> strategy?

Let us make computations like to our previous papers, i.e., having the same algorithms of dynamical risk functions (DRF) constructing. For the 1<sup>st</sup> strategy, we obtain the following risk function:

$$-0.685 \times x_2 + 1.300 \times x + 0.386,$$

and for the 2<sup>nd</sup> strategy:

$$-0.694 \times x_2 + 1.374 \times x + 0.321.$$

The final values of averaging-out of expert evaluations using these risk functions are 0.111 for the 1<sup>st</sup> strategy and 0.147 for the 2<sup>nd</sup> strategy. Therefore, using such algorithms of DRF for special averaging-out of expert evaluations gives “natural” answers.

Remark that twice repeated using of averaging-out (i.e., averaging-out using preliminary values of the first step of DRF-using) chooses the 1<sup>st</sup> strategy again. However, in the limit we have “natural” answers again. Let us describe these results by the following table; the names of columns are equal to the number of step of averaging-out using DRF (i.e., the number of using algorithm constructing DRF). Then the column 0 is the simple average of expert evaluations), and the column ¥ is the limit value.

	0	1	2	3	4	5	...	¥
1 <sup>st</sup> strategy	0.081	0.111	0.104	0.106	0.105	0.105	...	0.105
2 <sup>nd</sup> strategy	0.053	0.147	0.094	0.118	0.106	0.112	...	0.110

Let us remark, that this example is really possible in solving real problems: in the real computations for the mentioned DOP, the situations, when the difference between values of maximum and minimum expert values is more than 0.5 (i.e., more than 50% of the segment of expert values) are very often; for example, for accidental traveling salesman problem having dimension 75 and some of predictors mentioned before, they contain, by statistics of the author, about 10%.

#### 4. Some versions of “triangular” norm of quality definitions for distance metric

So, there are various algorithms to determine the distances between genomes; they can be called algorithms definition of the metric on the set of genomes<sup>2</sup>. However, this raises not only the usual questions about the adequacy of the corresponding mathematical models (from the point of view of the authors, they are usually solved in this domain by experts in biology, [15] etc.), but also on the comparative evaluation of these models. The most important matter in this case appears the following one: can we talk about the effectiveness of such algorithms and the adequacy of these models based on only one analysis matrices

<sup>2</sup> Mathematical aspects of the correctness of using the concept “metric” in this situation, we are expecting to discuss in a future article.

proximity (distance) between the genomes, without the involvement of biologists? The authors of this paper believe that this question should be answered in the affirmative.

For several different algorithms [4,5,6,7,10], we consider the matrices of distances between the genomes; in our computational experiments (see. below), we used five different algorithms<sup>3</sup> and made corresponding distance matrices, in which the number of genomes reached 100.

In this case, we used the following natural philosophy (we have not found analogues in the literature); we give it for the example of human (H), chimpanzee (C) and bonobo (B). According to biologists, Sand B dispersed (had a common ancestor), according to various estimates, about 2–2.5 million years ago (no wonder; the alternative name of B is “pygmy C”, [16]), and H dispersed with both other species 5.5–7 million years ago<sup>4</sup>. In this connection, the following question arises: why H should be closer to B comparing S? or vice versa: why it should be closer to C comparing B? Obviously, the answer to both these questions is negative, i.e., by other words, the explanation of the greater intimacy cannot exist. Therefore, in the matrix of distances between the genomes of all received triangles *should ideally be acute isosceles* ones.

To compare the quality of algorithms for constructing the distance we have offered several versions of “waste” (so called badness) of such “longisosceles” triangles. Apparently, when calculating the badness of the whole matrix for each option, we should always appropriate to summarize all the badness of all possible triangles of the matrices; we make this thing in our work.

So, *in simple cases*<sup>5</sup>, we will assume badness (norm) of the entire sum of the distance matrix, and for the badness of each triangle will apply one of the following 4 options. (We assume everywhere, that the considered triangle has sides a, b and c, moreover  $a \geq b \geq c$ ; the angles are  $\alpha$ ,  $\beta$  and  $\gamma$ , moreover  $\alpha \geq \beta \geq \gamma$ .)

1.  $(\alpha - \beta) / \pi$ .
2.  $(\alpha - \beta) / \alpha$ .
3.  $(a - b) / a$ .
4. For the final norm, we consider *separately* “violation of an isosceles” and “violation of an acute-angled”:
  - (A)  $1 - \min(b/a, c/b)$  ;
  - (B)  $\max(a - \pi/3, 0) / (2\pi/3)$  ;
 and the general answer is  $(A+B) / 2$  .

The maximum value of badness (in each of these four cases) to a triangle may be equal to 1. At the same bad case of algorithms for constructing metrics (i.e., when violation of the triangle inequality occurs) we believe this value from 1 to 2 (also depending on the quantitative characteristics of the violation).

As we noted above, some results of calculations are given below.

## 5. The preprocessor computing as a special normalization of date

Results and discussion may be presented in separate sections or combined into a single section, whichever format conveys the results in the most lucid fashion. The authors should discuss the significance of observations, measurements, or computations and should also point out how these contribute to the aims indicated in the Introduction. Tables, Figures, and Figure Captions should be embedded within the Section.

In this section we consider another heuristic, which can be considered optional for all heuristics of “violation of an isosceles and of an acute-angled” considered before. For this thing, we consider a function of the type  $f(x) = x^\alpha$ , where the value  $\alpha$  (usually,  $0 < \alpha < 1$ ) is chosen for each matrix of distances (see below some more about selecting  $\alpha$ ). Where each of the  $x$  of distance matrix is replaced by  $f(x)$ .

To select specific values of  $\alpha$ , *improving, from our point of view, the quality of the choice of metrics*, we applied the following considerations. Below, considering the description of the results of computational experiments, we will show, that various heuristics select metrics are relatively different priorities for genomes for “distant” and “close” species; and it is worth noting that such a priority varies little with his study at different rates described above. Attempts to improve the value of these norms (badness) applying some functions of the type  $f(x) = x^\alpha$  are unsuccessful: solutions of the corresponding minimization problems given either the maximum or minimum value  $\alpha$  (among all the possible ones); it is easy to understand, that in this case, we obtain the matrix of distances between genomes triangles “are closest to the acute-angled isosceles”. There fore, if we really try to improve quality metrics, it is necessary to use a fundamentally different heuristics. For this thing, we were trying to find a function of the above type in which the set of values of the distance matrix, viewed as a distribution of a random variable, is obtained as close as possible to a uniform distribution<sup>6</sup>; in advance, we note *that for different tasks* (i.e., for different concrete matrices of distances), the values of  $\alpha$ , obtained by pseudo-optimal real-time algorithms (which are realized by algorithms of [8,9,14] etc.) *are different*.

In this case, we have chosen the goal function on the basis of the entropy maximization ([17] and many others). Specific outcomes associated with the use of this heuristic are given below.

<sup>3</sup> Especially note again, that among these algorithms is one of our, the original one.

<sup>4</sup> It is important to note that the exact values of time such models are not important!

<sup>5</sup> We note in advance that we will consider a somewhat more complex option.

<sup>6</sup> Informally that can explain, for example, as follows. We already know, that in our model, genomes of human (H), bonobo (B) and crocodile (C) form a “stretched” acute-angled triangle, which is close to an isosceles one. In this case, the exact values of the lengths H–C and B–C unlikely to be of interest; it is only important, that they are approximately equal. Also unlikely, that the value ratio of the length of H–B to the length of H–C is to be of interest.

**6. Some local results of calculations**

Table 1. Panin's algorithm.

	Bison bison	Bos taurus	Canis lupus	Drosophila simulans	Felis catus	Gadus morhua	Gallus gallus	His1 virus	Homo sapiens
Bison bison	100	89,32	70,96	40,04	74,2	58,51	57,94	36,64	69,43
Bos taurus	89,32	100	73,62	40,02	71,4	60,78	55,84	36,41	66,27
Canis lupus	70,96	73,62	100	39,32	71,1	59,63	53,61	35,86	63,53
Drosophila simulans	40,04	40,02	39,32	100	43,53	41,51	40,82	39,34	41,42
Felis catus	74,2	71,4	71,1	43,53	100	55,74	58,55	35,55	67,26
Gadus morhua	58,51	60,78	59,63	41,51	55,74	100	53,18	35,08	56,03
Gallus gallus	57,94	55,84	53,61	40,82	58,55	53,18	100	34,19	57,7
His1 virus	36,64	36,41	35,86	39,34	35,55	35,08	34,19	100	41,07
Homo sapiens	69,43	66,27	63,53	41,42	67,26	56,03	57,7	41,07	100

Table 2. Winkler's algorithm.

	Bison bison	Bos taurus	Canis lupus	Drosophila simulans	Felis catus	Gadus morhua	Gallus gallus	His1 virus	Homo sapiens
Bison bison	1	0,8763	0,8614	0,8076	0,8652	0,8376	0,85	0,8032	0,8432
Bos taurus	0,8763	1	0,905	0,7834	0,8791	0,8556	0,8334	0,8442	0,8587
Canis lupus	0,8614	0,905	1	0,7851	0,8765	0,8687	0,8304	0,8438	0,853
Drosophila simulans	0,8076	0,7834	0,7851	1	0,7835	0,7647	0,8176	0,7683	0,7493
Felis catus	0,8652	0,8791	0,8765	0,7835	1	0,857	0,8317	0,8207	0,8555
Gadus morhua	0,8376	0,8556	0,8687	0,7647	0,857	1	0,8151	0,8409	0,8385
Gallus gallus	0,85	0,8334	0,8304	0,8176	0,8317	0,8151	1	0,7838	0,8682
His1 virus	0,8032	0,8442	0,8438	0,7683	0,8207	0,8409	0,7838	1	0,8092
Homo sapiens	0,8432	0,8587	0,853	0,7493	0,8555	0,8385	0,8682	0,8092	1

Table 3. Third algorithm.

	Bison bison	Bos taurus	Canis lupus	Drosophila simulans	Felis catus	Gadus morhua	Gallus gallus	His1 virus	Homo sapiens
Bison bison	1	0,846	0,604220	0,3181985	0,625668764	0,469214183	0,468422307	0,322671569	0,57058362
Bos taurus	0,846	1	0,61892522	0,317847962	0,602974896	0,485864878	0,450071497	0,322071245	0,54511437
Canis lupus	0,604220216	0,61892522	1	0,329786598	0,606326063	0,475700879	0,42999285	0,331161456	0,520353877
Drosophila simulans	0,318198529	0,317847962	0,329786598	1	0,324592863	0,308996167	0,280326501	0,292271663	0,28758525
Felis catus	0,625668764	0,602974896	0,606326063	0,324592863	1	0,451408078	0,475042624	0,335116703	0,563642777
Gadus morhua	0,469214183	0,485864878	0,475700879	0,308996167	0,451408078	1	0,409675882	0,322712027	0,432498802
Gallus gallus	0,468422307	0,450071497	0,42999285	0,280326501	0,475042624	0,409675882	1	0,306720686	0,475869876
His1 virus	0,322671569	0,322071245	0,331161456	0,292271663	0,335116703	0,322712027	0,306720686	1	0,307683023
Homo sapiens	0,57058362	0,54511437	0,520353877	0,28758525	0,563642777	0,432498802	0,475869876	0,307683023	1

Table 4. Winkler's algorithm.

	Bison bison	Bos taurus	Canis lupus	Drosophila simulans	Felis catus	Gadus morhua	Gallus gallus	His1 virus	
Bison bison	1	0,9041	0,7588	0,5315	0,7771	0,6719	0,6723	0,5195	0,735
Bos taurus	0,9041	1	0,7679	0,5307	0,762	0,6824	0,6603	0,5186	0,7186
Canis lupus	0,758	0,7679	1	0,5275	0,758	0,6711	0,644	0,5139	0,6952
Drosophila simulans	0,5185	0,5183	0,5275	1	0,5883	0,5758	0,556	0,5412	0,5604
Felis catus	0,8486	0,8331	0,8486	0,5883	1	0,6477	0,6651	0,5093	0,7142
Gadus morhua	0,6447	0,6554	0,6601	0,5059	0,6477	1	0,6315	0,5095	0,6413
Gallus gallus	0,6572	0,6462	0,6453	0,4977	0,6775	0,6315	1	0,4972	0,6658
His1 virus	0,5051	0,5048	0,5122	0,4819	0,5161	0,5068	0,4972	1	0,5768
Homo sapiens	0,8294	0,8118	0,8043	0,5791	0,84	0,7404	0,7727	0,5768	1

## 7. Some final results of calculations

Below, we shall call:

- our original algorithm for constructing a metric between genomes by the *first* one (below No. 1, see [10]);
- one of algorithms of M. van der Loo etc. (below No. 2, see [5], used function is `jarowinkler()`) by the *second* one;
- another algorithm of M. van der Loo etc. (below No. 3, see again [5], used function is `stringdist()`) by the *third* one;
- one of algorithms of H. Pages etc. (below No. 4, see [6], used function is `stringDist()`) by the *fourth* one;
- another algorithm of H. Pages etc. (below No. 5, see [6], used function is `pairwiseAlignment()`) by the *fifth* one.

Let us denote, that algorithms No. 4 and No. 5 are “non-symmetrical” ones, and, when filling in the distance matrix, we used half-sums of the two obtained values. Also let us note that the violations of the triangle inequality were recorded only as a result of the algorithms No. 4 and No. 5; however, in the case of “distant” species, we had a few such results: approximately, 1 case per 2000 examined potential triangles.

For further counting, we firstly have randomly chosen genomes of 100 representatives of the species, given in [18] (the case of considering “distant species”<sup>7</sup>). Some results of computations (the table 100x100, i.e.,  $100 \cdot 99 / 2 = 4950$  values, making  $(100 \cdot 99 \cdot 98) / (2 \cdot 3) = 161700$  triangles) are given below in Table 1, where:

- the rows are number of the algorithms (as we write before);
- the columns: approximate time of the creation of the distance matrix (for making all the 4950 values, CPU clock speed is approximately 2 GHz); number of violations of the triangle inequality (in average 1000 launches for triangles); middle badness, counted for all the algorithms 1–4 counting badness for each triangle, see Section 3.

All values badness we give up to 3 decimal places (the time of algorithms for constructing matrices was recorded some less accurately). In all tables, we celebrated the best metric for the considered norm (it is singled twice) and also the 2nd place (it is singled once).

Table 5. “Distant” species.

No.	time (h)	violations	badness-1, $(\alpha - \beta) / \pi$	badness-2, $(\alpha - \beta) / \alpha$	badness-3, $(a - b) / a$	badness-4, $(A + B) / 2$
1	27	0	<b>0,0372</b>	<b>0,0822</b>	<b>0,0416</b>	<b>0,196</b>
2	2.1	0	0,0954	0,197	0,0926	0,252
3	2.3	0	0,345	0,476	0,163	0,468
4	28	0.37	<b>0,0416</b>	<b>0,0907</b>	<b>0,0469</b>	<b>0,176</b>
5	28	0.38	0,0549	0,116	0,0556	0,214

As we can see, the algorithm implemented by our group, the majority of rules is optimal. And there is very important to remark (it follows from the above), that *heuristics that were used to create this algorithm had absolutely no connection with the heuristics used to describe the “triangle rules” defined below.*

*Secondly* (the case of considering “near” species), we also randomly have chosen genomes of human and 5 apes (bonobo, chimpanzee, gorilla, orangutan, gibbon), which are also given in [18]. In this case, each species taking 4-5 representatives (of 28 genomes); for the human, we took the genomes of different races. Some results of computational experiments are given in Table 2, where, unlike Table 1, we failed build time. Furthermore, due to the small total number of triangles (less than 5000), we have brought the number of violations of the triangle inequality (rather than relative values of this quantity).

<sup>7</sup> All lists of specific species corresponding genomes taken primarily from the site [18]. The authors are ready to send the obtained values of distance matrices, as well as the source code, by e-mail (at your request). We are also ready to send the detailed results of calculations of the badness, including not only the averaged, but all produced in the process value.



Table 6. “Near” species.

No.	violations	badness-1, $(\alpha-\beta) / \pi$	badness-2, $(\alpha-\beta) / \alpha$	badness-3, $(a-b) / a$	badness-4, $(A+B) / 2$
1	0	0,0757	0,152	0,0645	0,364
2	1	<b>0,0333</b>	<b>0,0687</b>	<b>0,0302</b>	<b>0,215</b>
3	1	0,514	0,622	0,170	0,582
4	32	<b>0,0595</b>	<b>0,122</b>	<b>0,0496</b>	<b>0,341</b>
5	39	0,0741	0,151	0,0615	0,350

As we can see, the relative number of violations of the triangle inequality significantly increases. In addition, our original distance metric between genomes is now not optimal.

Thirdly, we used “preprocessor” algorithms as previously described. It should be noted that the application of these auxiliary algorithms decreased value of badness almost all cells; however, it was *not the goal* of this algorithm. Besides, “leaders little changed”, i.e., our algorithm for constructing the metric (string 1) shows better results (comparing the absence of these auxiliary algorithms); however, the latter fact is just and can be explained by “tuning” the algorithm 1 for its use for a greater range of values. The results of computational experiments are given below in Table 3.

Table 7. “Near” species. (after pre-application of “preprocessor” algorithm).

No.	violations	badness-1, $(\alpha-\beta) / \pi$	badness-2, $(\alpha-\beta) / \alpha$	badness-3, $(a-b) / a$	badness-4, $(A+B) / 2$
1	0	<b>0,0522</b>	<b>0,121</b>	0,0527	0,351
2	0	<b>0,0314</b>	<b>0,0692</b>	<b>0,0290</b>	<b>0,205</b>
3	0	0,501	0,600	0,154	0,580
4	12	0,0527	0,122	<b>0,0482</b>	<b>0,323</b>
5	14	0,0732	0,150	0,0608	0,320

And fourthly, we applied the same algorithm to the genomes of human races (white man, yellow man, black man, bushman, Australian man). In this case, each race took 3–4 representatives (total 18 genomes). Some results of calculations are given in Table 4, where the columns mean the same as in Table 2.

Table 8. Races of human.

No.	violations	badness-1, $(\alpha-\beta) / \pi$	badness-2, $(\alpha-\beta) / \alpha$	badness-3, $(a-b) / a$	badness-4, $(A+B) / 2$
1	17	0,140	0,243	0,0924	0,325
2	29	<b>0,119</b>	<b>0,173</b>	<b>0,0359</b>	0,342
3	30	0,420	0,527	0,187	0,493
4	30	<b>0,119</b>	<b>0,218</b>	<b>0,0880</b>	<b>0,313</b>
5	26	0,129	0,229	0,0881	<b>0,306</b>

We could make a lot of comments to the values listed in this table; let us consider only the main one. A relatively large number of violations of the triangle inequality (and consequently, significantly larger values of badness, at its counting on any of the rules), is apparently due to the large number of people crossing concrete already after the separation of races. I.e., apparently, these algorithms should not be used to the genomes of subspecies (without their further modifications).

However, despite this fact, we are going to publish some further improvement of our original algorithm for constructing metrics, as well as our approach to the description of the badness. Besides, different algorithms for constructing metrics may be more appropriate with respect to different situations.

## 8. Conclusion. Some future ways for research

In this section, we look briefly at some algorithms that are going to be published in subsequent papers.

As a possible connection between the two approaches to solving problems biocybernetic and the traveling salesman problem (at first, so called its pseudo-geometrical version, see [8,19] etc.) we may call not only the above-mentioned multi-heuristic approach to the problems of discrete optimization, but also so called algorithms for pseudo-placing dots in  $k$ -dimensional Euclidean space [19]. These auxiliary algorithms improve the performance of other our algorithms. Moreover, algorithms similar to ones used by us in [13,14] could also be considered as such auxiliary algorithms; they some improve algorithms described in this article. Described in these articles use the risk functions of this direction is also, and we apply the special applications of the well-known “3 sigma rule”.

Besides, one of the most frequently discussed in biocybernetics problems is that the recovery of the distance matrix, when we know a part of the completed element only [11, 20]. Using the same “triangular norm”, we propose an original algorithm for such recovery; we are going to write about it in the near future.

## Acknowledgements

The reported study was funded by RFBR according to the research project № 16-47-630829.

## References

- [1] Gusfield D. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge, 1997; 631 p.
- [2] Toppi J, De Vico Fallani F, Petti M, Vecchiato G, Maglione AG, Cincotti F, Salinari S, Mattia D, Babiloni F, Astolfi L. A new statistical approach for the extraction of adjacency matrix from effective connectivity networks. IEEE Engineering in Medicine and Biology Society (EMBC) 2013; 3-7: 2932–2935.
- [3] Torshin IYu. Bioinformatics in the Post-Genomic Era: The Role of Biophysics. Nova Biomedical Books, NY, 2006.
- [4] Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Survey

- [5] Van der Loo MPJ. The Stringdist Package for Approximate String Matching. *The R Journal* 2014; 6: 111–122.
- [6] Pages H, Aboyoum P, Gentleman R, DebRaoy S. Biostrings: String Objects Representing Biological Sequences and Matching Algorithms. R package version 2.10.1, 2009.
- [7] Morgan M, Lawrence M. ShortRead: Base classes and methods for high-throughput short-read sequencing data. R package version 1.0.6, 2009.
- [8] Melnikov BF. Multiheuristic approach to discrete optimization problems. *Cybernetics and Systems Analysis* 2006; 3: 335–341.
- [9] Melnikov BF. Discrete optimization problems some new heuristic approaches. *Proceedings – Eighth International Conference on High-Performance Computing in Asia-Pacific Region, HPC Asia 2005 8th International Conference on High-Performance Computing in Asia-Pacific Region, China Computer Federation, Beijing, 2005: 73–80.*
- [10] Makarkin S, Melnikov B, Panin A. On the metaheuristics approach to the problem of genetic sequence comparison and its parallel implementation. *Applied Mathematics (Scientific Research Publishing)* 2013; 4(10): 35–39.
- [11] Eckes B, Nischt R, Krieg T. Cell-matrix interactions in dermal repair and scarring. *Fibrogenesis Tissue Repair* 2010; 3-4. DOI: 10.1186/1755-1536-3-4.
- [12] Carr RW, Hennessy J L. WSCLOCK – a simple and effective algorithm for virtual memory management. *SOSP '81 Proceedings of the eighth ACM symposium on Operating systems principles, 1981: 87–95.*
- [13] Melnikov BF. Heuristics in programming of nondeterministic games. *Programming and Computer Software* 2001; 5: 277–288.
- [14] Melnikov B, Radionov A, Moseev A, Melnikova E. Some specific heuristics for situation clustering problems. *ICSOFIT, Technologies, Proceedings 1st International Conference on Software and Data Technologies 2006: 272–279.*
- [15] Foley J. Fossil Hominids: mitochondrial DNA 2011; URL: <http://www.talkorigins.org/faqs/homs/mtDNA.html>
- [16] Frans BM. *Bonobo: The Forgotten Ape.* University of California Press, 1998, 224 p.
- [17] Popkov YS. Substantiation of the entropy maximization method for problems of image restoration from projections. *Automation and Remote Control, 1995; 56(1): 77–82.*
- [18] NCBI: nucleotide database, 2014, URL: <http://www.ncbi.nlm.nih.gov/nucleotide>.
- [19] Makarkin SB, Melnikov BF. Stochastic Optimization in Informatics. Geometric methods for solving the traveling salesman problem pseudo-version 2013: 54–72. (in Russian)
- [20] Midwood KS, Williams LV, Schwarzbauer JE. Tissue repair and the dynamics of the extracellular matrix. *The International Journal of Biochemistry & Cell Biology* 2004; 36(6): 1031–1037.
- [21] Shao M, Lin Y, Moret B. An Exact Algorithm to Compute the DCJ Distance for Genomes with Duplicate Genes. *Research in Computational Molecular Biology, Lecture Notes in Computer Science Volume 2014; 8394: 280–292.*
- [22] Melnikov B, Radionov A. On the decision of strategy in non-deterministic antagonistic games. *Programming and Computer Software* 1998; 24(5): 247–252.

# Development of parallel implementation of the informative areas generation method in the spatial spectrum domain

N. Kravtsova<sup>1</sup>, R. Paringer<sup>1,2</sup>, A. Kupriyanov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

This paper proposes parallel implementation of the image informative segments extraction method. The images are segmented in the spatial spectrum domain. Median energy in each selected segment is viewed as an area. For time saving purpose parallel implementation was developed for the areas calculation phase. The developed software implementation was tested on the high performance multicore computing system.

*Keywords:* diagnostic crystallogram; spatial spectrum; discriminant analysis; k-NN classification; parallel implementation

---

## 1. Introduction

Currently computer processing of medical diagnostic images is one of the vital research tools and a way to improve efficiency of early detection of various diseases. Change of the body fluids composition is one of the information-bearing health condition areas. Metabolic change that occurs due to pathological conditions affects the fluid composition; there are numerous changes in the molecular composition of tissues and body fluids. Converting the fluid from one phase state to another is one of the ways to detect such changes. Crystallization is one of the most convenient methods to change the fluid phase. Crystal properties modification is caused by changed physical and chemical properties of a body fluid. The investigation of these properties is the crucial problem of crystal analysis [1]. In medicine, studied crystallograms are the structures formed by salt crystallization caused by body fluid drying.

In clinical practice the crystallogram analysis is based on their images. It is not always possible to visually identify changes in such key crystallogram parameters as predominant bar direction, bar density etc., which contribute to major pathologic signs. Quantified analysis and objectivity are among computer analysis advantages.

The information contained in the image is structurally excessive. It is known that if the parallel bars of certain direction were predominant on the original image, the bars of the same direction would dominate the Fourier transform of an original image. This property can be used to analyze crystallograms [2, 3] and other images of branched structure.

The developed method based on discriminatory analysis algorithm is applied to generate the informative areas set, which is used in this paper to identify the characteristics of the initial crystallogram images.

In the article, we propose a parallel implementation of the method to speed up computations.

## 2. Informative areas generation method

### 2.1. Description of the areas used

In this work, the areas are derived from calculation of the total energy on the selected spectrum image ranges. Most part of the spectrum does not contain the information suitable to identify the characteristics of an original image.

If the image function and its Fourier transform  $F(u, v)$  are considered in a spatial domain, then the magnitude  $|F(u, v)|^2$  defines an energy spectrum of the image. The energy spectrum of the image can be directly analyzed as a whole or partially.

In this work, we analyzed features derived by calculating the total energy of a selected domain of the spectrum image. The spectrum image in the domain of interest was segmented using a formula:

$$C_{r_1 r_2 \theta_1 \theta_2} = \sum_{r=r_1}^{r_2} \sum_{\theta=\theta_1}^{\theta_2} |F(r, \theta)|^2,$$

where,  $\theta_1$  и  $\theta_2$  are the bounding angles of the sector.

Since the spectral image is symmetric around the center, only half of the image will be used to form up the areas [4, 5].

### 2.2. The technique of building an efficient set of areas for image discrimination

This paper describes the method to extract the informative segments from the spectrum images (in Figure 1 is shown as a stage of smart area analysis process). The segment informative value was estimated using the criteria of separability of discriminatory analysis algorithm.

The methods based on the discriminative analysis algorithm proved to be a good solution to form up new problem-specified areas [6]. These methods permit to improve the reliability of data classification [7, 8].

The discriminative analysis is used to eliminate the correlation between the areas and consequently to reduce the size of the areas set. The usage of this algorithm allows, on the one hand, to maintain the informativeness of the feature set for classification and, on the other hand, to reduce the number of areas to apply less complicated classification methods and to reduce the classification error value.

An individual separability criterion was calculated for each area based on equation:  $J = \text{tr}((\mathbf{T})^{-1}\mathbf{B})$ , where  $\mathbf{T} = \mathbf{B} + \mathbf{W}$ ,  $\mathbf{B}$  is between-group scattering matrix,  $\mathbf{W}$  is intragroup scattering matrix.

The informative areas set is further generated in the following manner:

- The areas are ranged in the order of decreasing of individual separability criteria values.
- The initial areas set consists of a feature with the largest criteria. Classification is carried out.
- Then the area with the next value of separability criterion is added to the set. Classification based on the new set of areas is carried out.
- Repeat item 3 until all areas are included into the set.

The informative set of areas is the one with classification results yielding the minimum error value. The classification error that defines the number of cases with the classifier acquiring incorrect value is calculated from the equation:  $\varepsilon = (m/n) \cdot 100\%$ , – where  $m$  is the number of classification errors,  $n$  is the total number of images tested.

### 3. Parallel implementation of the method to generate informative areas

The entire algorithm of informative areas generation may be presented as a diagram shown in Figure 1. The analysis of the algorithm structure showed that the first stage of areas computing has the maximum computational complexity.

As mentioned above most of the algorithm run time is spent on computing the areas. This stage includes pre-processing of the learning sample, generation of spatial spectrum for each learning image and calculation of the area values. The number of calculations can be reduced if one take into account that the spatial spectrum image is symmetrical relative to its center, and for this reason the areas calculation may and must be performed only on the half of the spatial spectrum image. In order to speed up the algorithm run time this paper will apply task division. This paper will not use an MPI technology but will apply task distribution by threads, that is why the way of splitting the image into tasks is not important. In case of sample pre-processing and spatial spectrum image generation a single image is sent to each thread. Then at the areas calculation stage each thread receives a separate element - an image segment calculated on the basis of the proposed segmentation. The next task is sent to the thread as soon as it completes handling of the element.

Application of this segmentation pattern allows substantial run-time saving at the first stage of the algorithm. During this resource-intensive stage such method of tasks segmentation by threads has permitted to achieve the three-fold acceleration when four threads were used. The curve in Figure 2 shows the relation between acceleration and the number of areas.

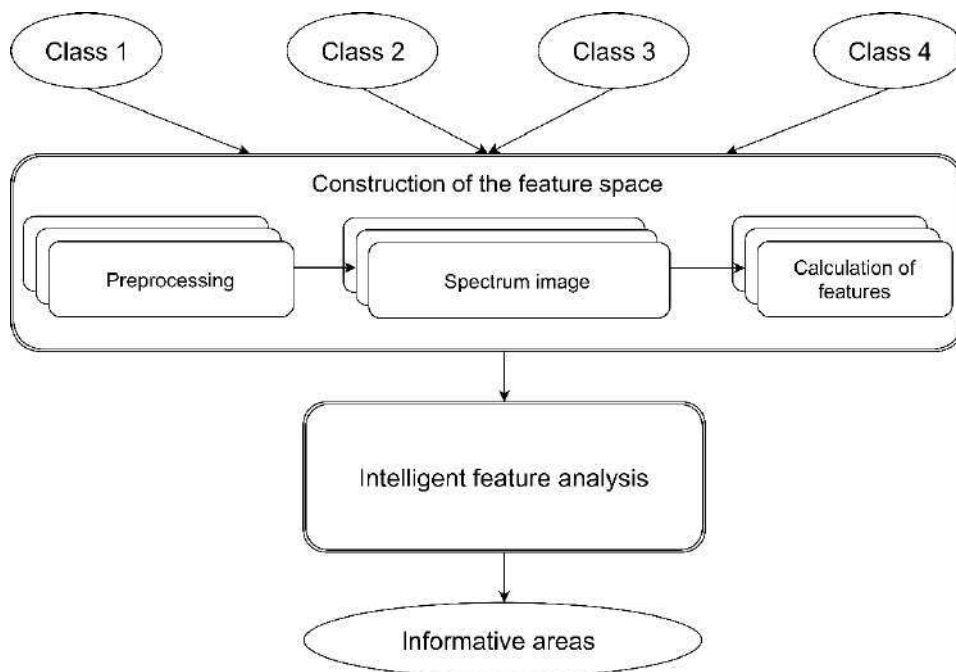


Fig. 1. Informative areas formation algorithm.

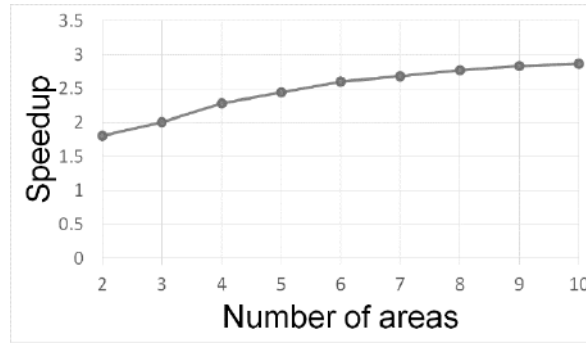


Fig. 2. Speedup graph.

#### 4. Classification results

After classification for all possible segmentations, areas were selected using the developed area selection algorithm. Figure 3 shows the relation between classification error and the number of areas taken in the descending order of their separability criterion in case of classification with 2-nearest neighbors methods to split into 4 sectors and 8 rings

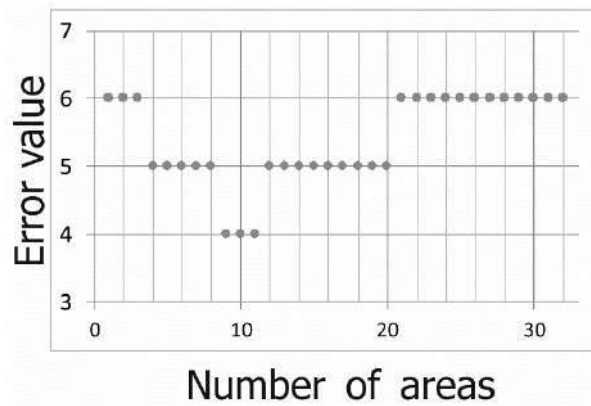


Fig. 3. Relation between classification error and the number of areas.

Tables 1 show classification error value after selecting the informative area set. The error value selection to be included into the final table is shown above.

Table 1. Classification error following area selection, %.

		Number of rings							
		1	2	3	4	5	6	7	8
Number of sectors	1	9	9	9	9	9	9	10	10
	2	8	8	7	7	7	7	7	7
	3	8	7	6	6	6	6	6	7
	4	7	7	6	5	5	5	4	4
	5	7	7	6	6	6	6	7	7
	6	7	7	7	7	7	7	7	7
	7	7	7	7	7	7	7	7	7
	8	7	7	7	7	7	7	7	7

#### 5. Conclusion

This paper presented the method to generate a set of informative local areas of a spatial spectrum in order to classify medical crystallogram images, and an option of parallel implementation of such method. In addition, the research included experimental testing of the developed software implementation, which demonstrated that parallel algorithm implementation provided almost three-fold acceleration at the areas calculation stage. The next step is to implement parallel computing at the smart area analysis stage in order to improve the informative area set calculation speed.

#### Acknowledgements

This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world’s leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29-03823, # 15-29-07077, # 16-41-

## References

- [1] Shirokanev AS, Kirsh DV, Kupriyanov AV. Researching of a crystal lattice parameter identification algorithm based on the gradient steepest descent method. *Computer Optics* 2017; 41( 3): 453–460. DOI: 10.18287/2412-6179-2017-41-3-453-460.
- [2] Paringer RA, Kupriyanov AV. The Method for Effective Clustering the Dendrite Crystallogram Images. 9th Open German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW 2014). Electronic on-site Proceedings, University of Koblenz-Landau, 2014.
- [3] Paringer RA, Kupriyanov AV. Research methods for classification of the crystallogram images. Proceedings of the 12th international conference PRIP'2014. Minsk, Belarus, 2014; 1: 231–234.
- [4] Kravtsova N, Paringer R, Kupriyanov A. Development of methods for crystallogram images classification based on technique of detection informative areas in the spectral space. *CEUR Workshop Proceedings* 2016; 1638: 357–363 DOI: 10.18287/1613-0073-2016-1638-357-36.
- [5] Gaidel AV, Krashennnikov VR. Feature selection for diagnosing the osteoporosis by femoral neck X-ray images. *Computer Optics* 2016; 40( 6): 939–946. DOI: 10.18287/2412-6179-2016-40-6-939-946.
- [6] Fukunaga K. Introduction to statistical pattern recognition. San Diego: Academic Press, 1990; 592 p.
- [7] Ilyasova NYu, Kupriyanov AV, Paringer RA. Formation features for improving the quality of medical diagnosis based on the discriminant analysis methods. *Computer Optics* 2014; 38( 4): 851–855.
- [8] Biryukova E, Paringer R, Kupriyanov A. Development of the effective set of features construction technology for texture image classes discrimination. *CEUR Workshop Proceedings* 2016; 1638: 263–269. DOI: 10.18287/1613-0073-2016-1638-357-363.

# Numerical simulation of motion of dust particles in an accelerator path

A.V. Piyakov<sup>1</sup>, D.V. Rodin<sup>1</sup>, M.A. Rodina<sup>1</sup>, A.M. Telegin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The model of micron charged particles motion in an electrostatic accelerator path is presented. This article describes software that provides formation of a particle packet with given statistical characteristics and the results of modeling, obtained by the supercomputer Sergey Korolev. Performance of software implementation for the supercomputer, parallel implementation for PC and implementation for GPU is being considered. Comparison with experimental data was carried out; convergence of full-scale and numerical experiments was shown.

*Keywords:* trajectory calculation; electrostatic accelerator; supercomputer; GPU; CUDA

---

## 1. Introduction

Space debris remains one of the main causes of degradation of spacecraft structural elements. At the same time tendency to increase concentration of technogenic dust particles in near-earth orbit remains [1]. Considering the current trend for increasing the duration of spacecraft operation, as well as to reduce their weight and dimensions, it is necessary to have ground-based experimental facilities that allow research of new materials in conditions of interaction with high-speed dust particles.

For such research accelerators of various types have been developed, in particular, the Van de Graaf accelerator, electrostatic and electromagnetic accelerators. Electromagnetic accelerators working on the principle of a railgun are most widely used to accelerate large particles (from 1 mm and above) [2]. Electrostatic accelerators, containing drift tubes to which an accelerating voltage is applied, are used to research the degradation of materials in the flow of high-speed micron particles. The in-phase switching of the accelerating voltages with increasing particle velocity can be provided in two ways: 1) increasing the length of the drift tubes; 2) increasing the frequency of the accelerating voltage. The second method is more preferable, because it provides greater flexibility in tuning such an accelerator [3].

The main characteristics of such an experimental facility are the range of output velocities and the particle flux density at the exit from the accelerator path. The main task in developing such accelerators is a maximization of these parameters. The solution of this problem is complicated by the fact that the full-scale modeling of constructions is largely hampered by their large geometric dimensions, use of high voltages and the need to create a vacuum system for each design. The only way to design such units is through mathematical modeling. To evaluate the characteristics of particles at the exit of the accelerator path with given parameters (such as the number of drift tubes, accelerating voltage and geometric parameters of tubes.) the authors developed software that solves the assigned task.

## 2. Preparation of initial data

In the laboratory facility, the injector generates a stream of charged particles with a probabilistic distribution of velocities, as well as probabilistic radial and angular distribution of particles in a packet. Then particles come on the input of electrostatic accelerator.

For the purpose of mathematical simulation of particle trajectories in the accelerator path, it is necessary to form model packets of the particles with probabilistic characteristics matching the characteristics of real packets. In order to do that, the software module that generates random variables with required distribution law was written.

To generate a model packet of particles with a distribution similar to the real one, the Box-Muller transformation with subsequent summation of the velocity vector components and normalization for the most probable energy corresponding to 450 m/s was used.

The algorithm of the generator of a model packet of particles can be represented in the following form:

- 1) formation of two random values  $U_1$  and  $U_2$  on the interval (0; 1] with the help of a generator of uniformly distributed random variables;
- 2) obtaining two random variables distributed normally, in accordance with expression:

$$x = \sqrt{-2\ln U_2} \cos(2\pi U_1);$$

$$y = \sqrt{-2\ln U_2} \sin(2\pi U_1);$$

- 1) the repetition of points 1 and 2 makes it possible to obtain three normally distributed numbers whose geometric sum is a velocity vector in three-dimensional space:

$$V = V_x^2 + V_y^2 + V_z^2;$$

2) normalization of the total velocity vector with the help of the scale factor obtained from the most probable speed.

The distribution of velocities of charged particles at the output of the injector is shown in the Fig. 1. The distribution of the model packet of particles by velocities is shown in the same figure.

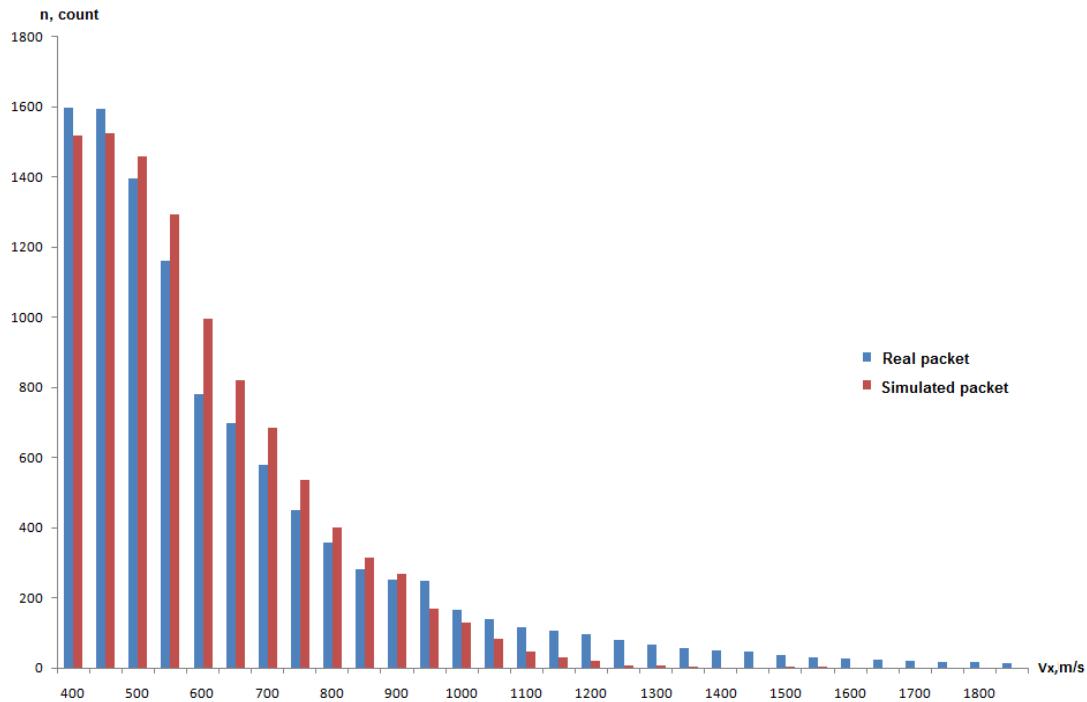


Fig. 1. Distribution of real and model particle packets on axial speeds.

The generation of the radial coordinates of the model packet of particles is also based on the Box-Muller transformation. The distribution of a model packet at radial coordinate is shown in Fig. 2.

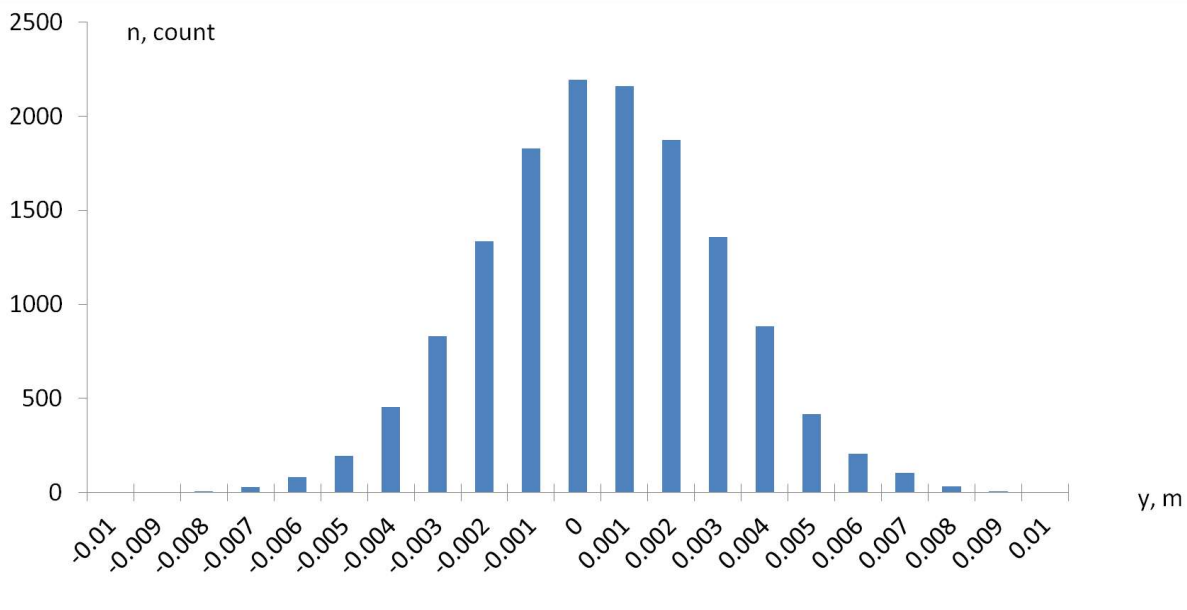


Fig. 2. The distribution of a model packet of particles at radial coordinates.

Due to the ratio of the size of the hole at the output of the injector and the distance to the charging needle, charged particles enter the input of the electrostatic accelerator at a solid angle equal to  $2^\circ$ . When forming the distribution of a model packet of particles at radial velocities (Fig. 3), particles with an unsuitable ratio of the radial and axial velocity components are excluded from consideration.

The acceleration received by the particle inside the path is largely determined by the ratio of mass to a charge. Therefore, the generator of the model packets also formed the probability distribution of the mass-to-charge ratio. The initial form of a distribution for a full-scale experiment was obtained indirectly, starting from a priori knowledge of the accelerating potential, as well as the velocities of the particles at the input and the output of the path. The real and model distributions are shown in Fig. 4.



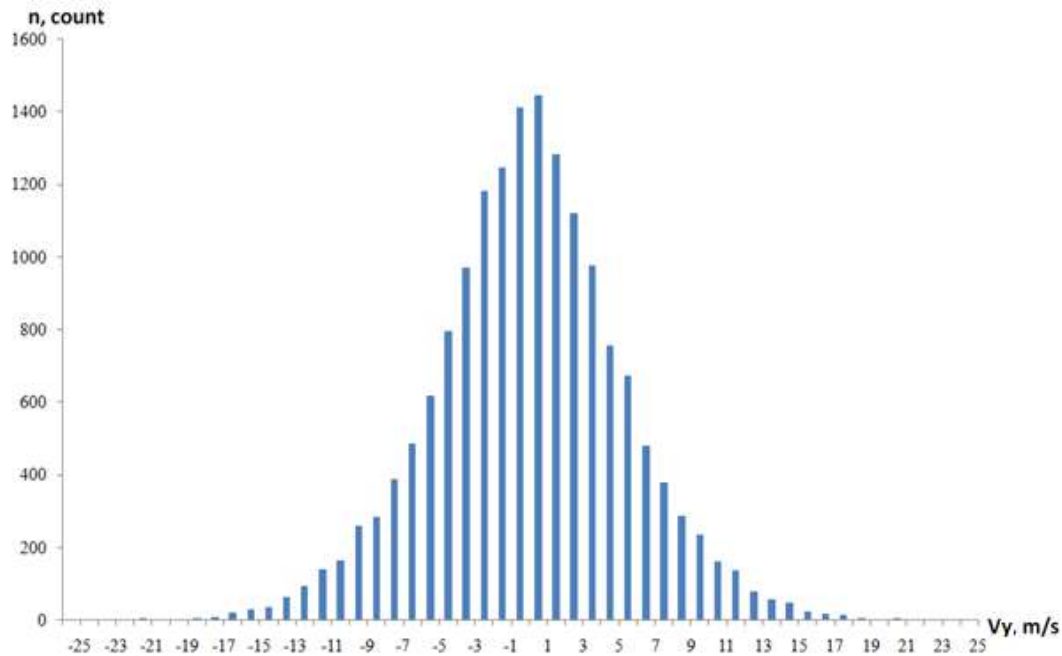


Fig. 3. The distribution of a model packet of particles on radial velocities.

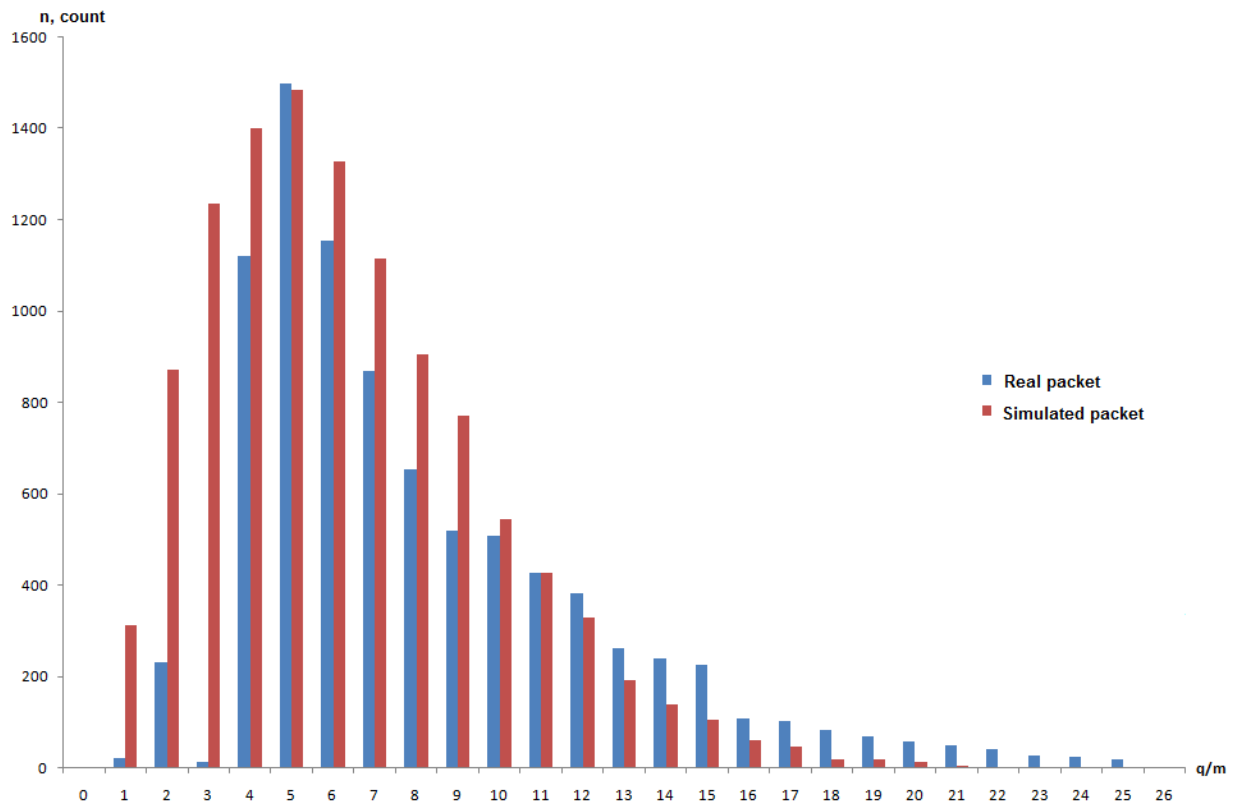


Fig. 4. Distribution of real and model particle packets on mass to particle charge in the package.

The motion of charged particles was simulated on a two-dimensional rectangular grid because of the axial-symmetry of the task. The grid nodes contain the values of electrostatic potential. Using the numerical calculation of the potential distribution for five accelerating tubes with a 10 kV potential difference between adjacent tubes, which corresponds to the real parameters of the accelerator, the field values at the nodes were obtained. The grid of the cell has a step of  $9,775 \cdot 10^{-5}$  m which corresponds to splitting the 10 cm section into 1023 intervals or 1024 nodes.

### 3. Calculation of the trajectory of particles in a linear accelerator path using a personal computer

The program for calculating the trajectories of charged particles in the path of a linear electrostatic accelerator using a personal computer was written in the C # programming language. The package of model particles contains 16384 pieces and

was formed according to the algorithms described in paragraph 2. All particles start from the coordinate  $x = 0$  m, the radial coordinate is subject to the normal distribution law and lies in the range from  $-0.01$  to  $0.01$  m. The components of the particle velocity vector correspond to the parameters of the real distribution of charged particles after the injector. To preserve the state of particles, the Particle class was used, which has the necessary fields for storing two coordinates, two components of the velocity vector, a time quantum and the total time of flight.

Interpolation of the field was carried out on the assumption that the particles have only a positive coordinate at the radial axis. For this the operation of taking the module from the radial coordinate of the particle is added to the interpolator function. Interpolation occurred for a field section of  $1 \text{ cm} \times 10 \text{ cm}$ , respectively, the  $x$  coordinate must always lie in the range  $0 \div 0.1$  m.

After a interpolation operation for particles having a negative radial coordinate, the radial field component must also be inverted, using the negative radial coordinate flag.

Particles were traced by the fourth-order Runge-Kutta method. The calculation for each of the particles was completed after the onset of one of two events: the particle exceeded the limit  $-0.01$  m or  $0.01$  m along the radial coordinate, which corresponds to the precipitation of the particle on the accelerating tube, or the particle exceeded the limit of  $1$  m along the longitudinal coordinate, which corresponds to the passing of a particle of the entire accelerator section.

Every  $0.025$  m the slice of all the characteristics of particles was built. After each iteration, the value of the time quantum for each of the particles was adjusted from the condition of the minimum number of steps per one grid cell.

All the initial parameters, such as a set of particles and field values, were saved for later use on the supercomputer Sergey Korolev (SK) and for calculations using the GPU accelerator.

The results of modeling the flight of a particle packet through the accelerator path are shown in Fig. 5 – 7.

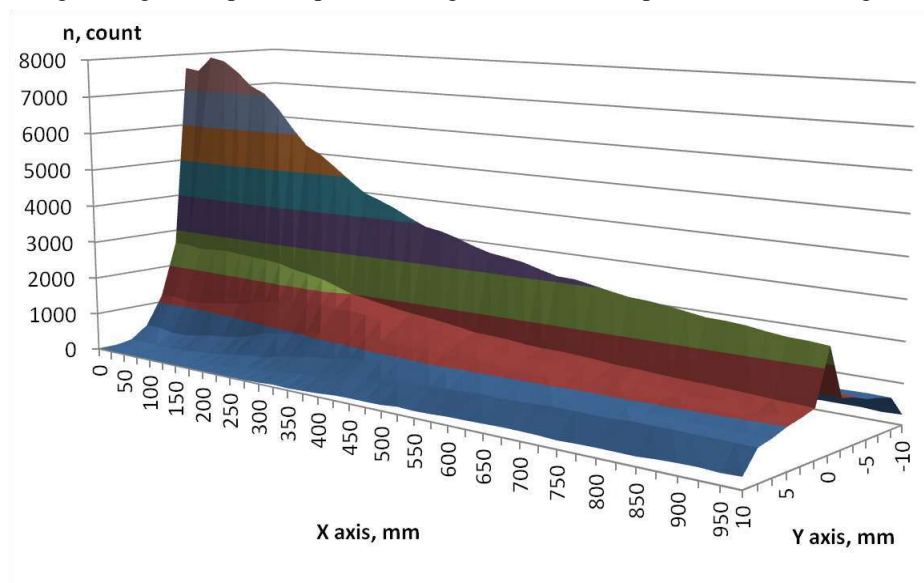


Fig. 5. The distribution of particles in the accelerator path in sections at the radial and longitudinal axes.

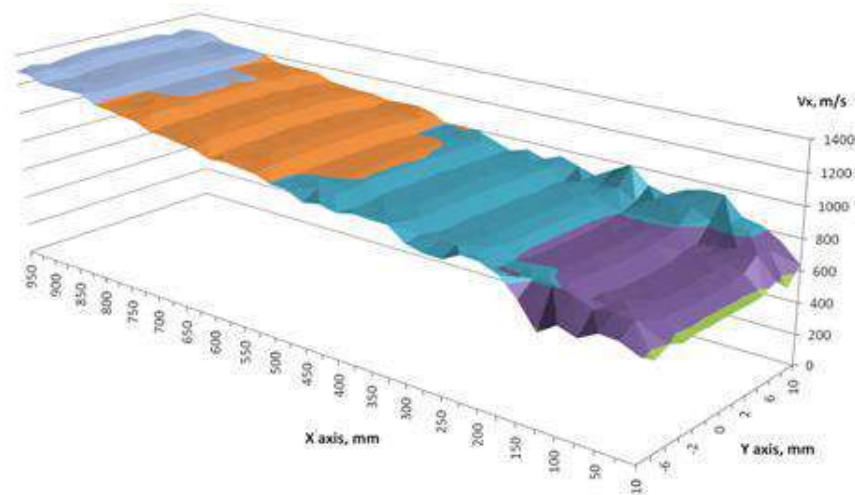


Fig. 6. Distribution of the average longitudinal velocity in the accelerator path in sections at the radial and longitudinal axes.

It can be seen that at the output from the accelerator path the average particle velocity increases more than twice, and the number of particles on the axis of the path decreases by approximately 2.5 to 3 times, however, the number of particles at the periphery increases.

The graph of distribution of mean radial velocity allows for the conclusion about the effect of focusing in the radial plane, because most of the time of flight of the accelerator path, the particle velocity along the radial coordinate has a sign opposite to the sign of their radial coordinate, i.e. particles tend to the axis of the device.

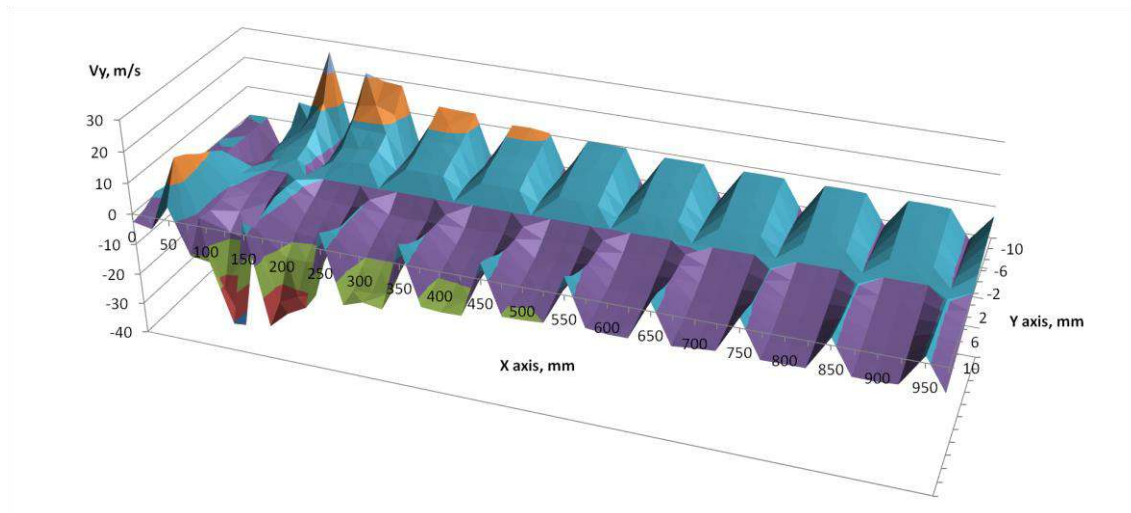


Fig. 7. Distribution of the mean radial velocity in the accelerator path in sections at the radial and longitudinal axes.

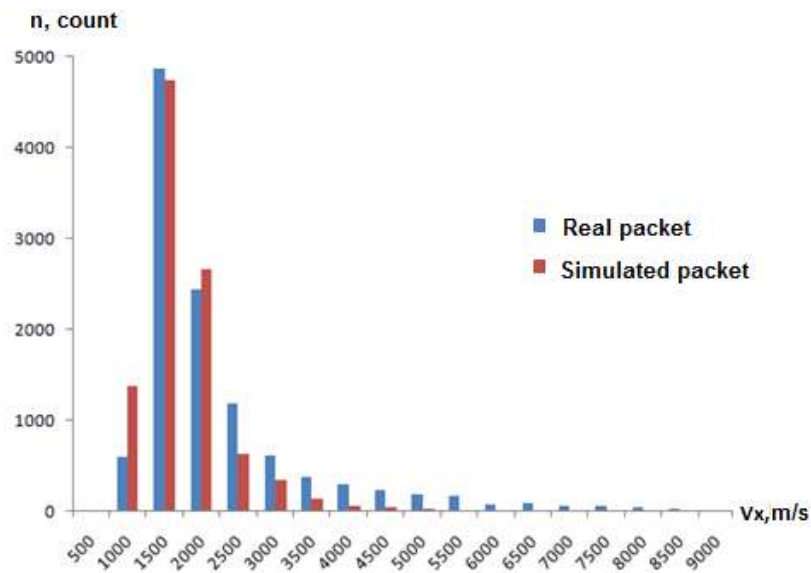


Fig. 8. Distribution of real and model particle packets on longitudinal velocities.

The main characteristic of an electrostatic accelerator is the range of speeds obtained at the output of its path. Data obtained by numerical simulation and by a real accelerator are shown in Fig. 8.

The above graphs show a good convergence of the results of numerical and full-scale experiments. In particular, the most probable particle velocity at the output from the path is approximately  $1.5 \div 2$  km/s and the number of particles reaching the output of the path for a real accelerator is 69.9 %, and for the model – 71.3 %.

#### 4. Calculation of the trajectory of particles in the linear accelerator path using the supercomputer SK

To calculate the trajectories using the supercomputer SK, an implementation of the program was developed to run on 16 processors. The implementation of the program is written in the programming language C, for storing particle parameters, a structure similar to the Particle class from paragraph 3 was used.

The matrix of the electrostatic field, as well as an array with particle parameters were connected as external header files with two-dimensional arrays declared and assigned inside. The implementation of multithreading was provided by connecting the MPI library. Another difference in the implementation of the program for the supercomputer SK was that two separate methods were used to interpolate the field, since the field arrays were connected as two separate header files.

Similarly to the program implementation for a personal computer, the calculation for each of the particles was carried out by fourth-order Runge-Kutta method before the particle leaves the path. Slices of the state of the particle packet were also constructed every 0.025 m by outputting data to the output stream.

After the end of the program, the supercomputer SK software automatically collected data from each of 16 nodes into one file on the head node. The result of execution is a set of states of particles on slices that needed to be sorted by the particle index.

The results of modeling the flight of a particle packet through the accelerator path using the supercomputer SK almost completely coincide with the graphs in Fig. 5 – 7, built on the data obtained on a personal computer.

## 5. Calculation of the trajectory of particles in the linear accelerator path using the GPU accelerator

This implementation of the program for calculating the trajectories of charged particles in the path of a linear electrostatic accelerator was written in C # using CUDA libraries. The difference from the other two implementations is the calculation of particle trajectories on the GPU accelerator, which requires special data preparation.

The initial data for the calculation was loaded from the files saved by the program from paragraph 3. Further, all the variables of the double data type were stored in arrays of the same length as the number of particles. In the host memory, data arrays of vector type int2 were created for storing 64-bit variables, then the data was converted and data of the vector type was copied to the GPU memory where the calculation was performed. The values of the field at the grid nodes were stored in the texture memory, the values of the particle characteristics – in the surface memory.

Calculation of the trajectories was carried out by fourth-order Runge-Kutta method, data upload from the GPU was performed during the flight of the next 0.025 m, or when particle leaves the path. Similar to the other two implementations, distributions of the particle characteristics at the path sections were constructed.

The results of modeling the flight of a particle packet through an accelerator path using a GPU accelerator coincide with the results obtained on a personal computer and a supercomputer (Fig. 5 – 7) with small differences obtained for the average radial velocity. These differences will be explained in paragraph 6.

## 6. Comparison of calculation results for three software implementations

To compare the results of the calculations obtained by three software implementations, a trajectory of a single particle, not exceeding the limits of the accelerator path, was constructed. Then the dependencies of the relative error of the radial coordinate, the relative error of the longitudinal velocity and the relative error of the radial velocity on the longitudinal coordinate were plotted, they are shown in Fig. 9 – 11.

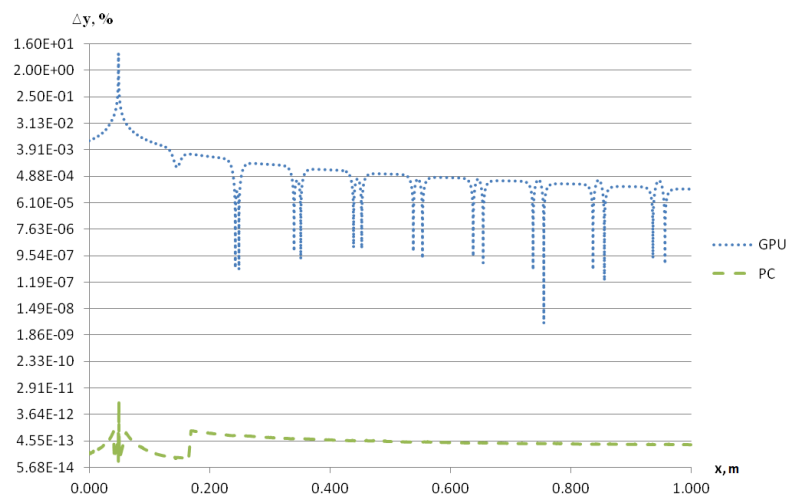


Fig. 9. Dependencies of the relative error of the radial coordinate from the longitudinal for a single particle, calculated using a GPU accelerator (GPU) and a uniprocessor personal computer (PC).

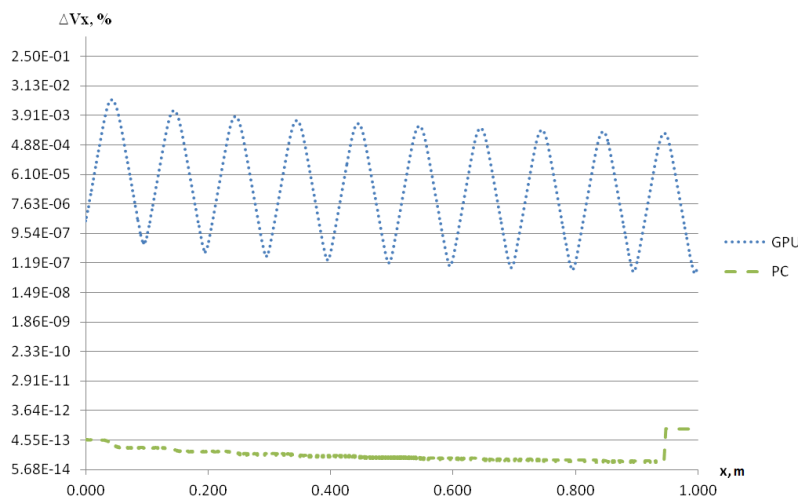


Fig. 10. Dependencies of the relative error of the longitudinal velocity at the coordinate  $x$  for a single particle, calculated using the GPU accelerator (GPU) and a uniprocessor personal computer (PC).

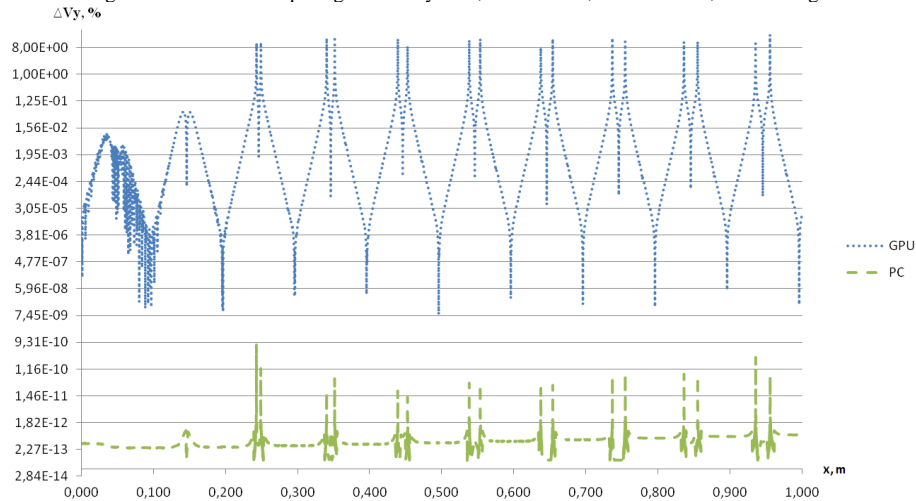


Fig. 11. Dependencies of the relative error of the radial velocity at the coordinate  $x$  for a single particle, calculated using the GPU accelerator (GPU) and a uniprocessor personal computer (PC).

To calculate the relative error, the data obtained from the supercomputer SK were used as reference values.

Relative errors for data obtained when calculating on a personal computer do not exceed  $10^{-10}$  % and, in fact, are a rounding error. The data obtained on the GPU accelerator has a greater relative error, which is due to the aspects of the representation of double precision numbers in the GPU accelerator memory. This is especially noticeable at the transition points of the radial coordinate and radial velocity curves through zero. At these points, the relative computational error is maximal.

## 7. Conclusion

The shown accuracy for all three methods is sufficient to adequately simulate the behavior of the particle flow.

The advantages of the software implementation for a supercomputer and a personal computer include greater accuracy. However, the execution time of the programs in this case strongly depends on the number of available processor cores. So, on 16 nodes of the supercomputer SK, a package of 16384 particles was calculated in 4 min 17 s, and on a single-processor machine in single-threaded mode the calculation took 46 min 34 s.

The software implementation for the GPU accelerator (2,880 cores at 1020 MHz) has a bit lower calculation accuracy. However, this disadvantage is not significant, since the greatest relative error arises for near-zero velocities for a very small number of iterations and, in the final analysis, practically does not affect on the result. For example, for the above particle, for 250,000 iterations, the difference occurs only in the 9-significant digit. In addition, this disadvantage is largely offset by the speed of this implementation of the software: the same package on the GPU accelerator was calculated in 40.64 s.

Since the algorithm for solving this task is parallel in input parameters due to the fact that the trajectories of the particles do not depend on each other, the computational speed is scaled in proportion to the number of particles and the number of processors involved. Thus, the most optimal is the approach in which the primary analysis of alternate designs occurs using the GPU accelerator, with the final verification of the selected design carried out by the supercomputer.

## References

- [1] Semkin ND, Kalaev MP, Telegin AM, Piyakov AV, Rodin DV. Multilayer film structures under the influence of micrometeoroids and space debris particles. *Applied Physics* 2012; 2: 104–115.
- [2] Suhachev KI. Rail electromagnetic accelerator with an external magnetic field. *Bulletin of the Samara State Aerospace University* 2015; 14(1): 177–189. DOI: 10.18287/1998-6629-2015-14-1-177-189.
- [3] Semkin ND, Piyakov AV, Voronov KE, Bogoyavlenskij NL, Goryunov DV. Linear accelerator for simulation of micrometeorites. *Devices and equipment of experiment* 2007; 1: 1–8.

# Performance Analysis of a Simple Runtime System for Actor Programming in C++

S.V. Vostokin<sup>1</sup>, E.G. Skoryupina<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

In this paper, we propose the Templet – a runtime system for actor programming of high performance computing in C++. We provide a compact source code of the runtime system, which uses standard library of C++ 11 only. We demonstrate how it differs from the classic implementations of the actor model. The practical significance of the Templet was examined by comparative study on the performance of three applications: the reference code in C++, managed by the OpenMP; the actor code in C++, managed by the Templet; the actor code in Java, managed by the Akka. As a test problem we used a numerical algorithm for solving the heat equation.

*Keywords:* performance analysis; message-oriented middleware; actor framework; high performance computing; C++ language

---

## 1. Introduction

Actor model proposed by Hewitt in 1973 [1] isn't out of date; on the contrary, it attracts more and more attention to the developers. This is due to the modern trend of widespread hardware solutions for massively parallel computing applications. One of the main features of the actor model is the ability to describe an unbounded natural parallelism. Therefore, actively developing technologies such as the infrastructure software, Internet of Things and high performance computing, which uses massively parallel computations, fit well into the concept of actors [2].

In the area of infrastructure software and the Internet of Things there is a popular framework for interpreted Scala and Java languages called Akka [3]. In high-performance computing, where compiled languages dominated, actors have been of little use. In our opinion, this is due to the prevailing stereotype of the implementation complexity of the actor model for compiled languages. New features of the latest versions of the standard, starting with C++ 11, led to the development of effective and portable implementations of actor models for compiled C++ [4].

The aim of the work is (1) to present a scalable, build-in implementation of the actor model in C++11, (2) demonstrate the effectiveness of the implementation by using high-performance computing test.

The remainder of this paper is organized as follows. First, we discuss the test problem for the performance evaluation in high-performance computing. Then, we describe the implementation of the Templet actor runtime system. Further, we propose three parallel implementations of the test problem: the first one based on OpenMP, the second one using the Templet system and the last using the Akka Framework. After that, we describe the conditions of the computational experiment and compare the results, and make a conclusion based on the results.

## 2. The Method of Efficiency Evaluation: Heat Equation Test

For comparative analysis we used an algorithm describing the solution to the heat equation (see Listing 1, 2). The algorithm was chosen due to the fact that it describes the sample implementation of frequently used finite-difference method and trivial one-dimensional decomposition of the data area for parallelization.

The constants  $W$  and  $H$  keep the width and height of the field grid area. The constant  $T$  is the number of time samples. An elementary operation `op` (Listing 1) shows the use of a differential stencil for calculating field values at the next time step.

```
1 void op(int i)
2 {
3   for (int j=1; j<W-1; j++)
4     field[i][j]=(field[i][j-1]+field[i][j+1]+
5     field[i-1][j]+field[i+1][j])*0.25;
6 }
```

Listing 1. Elementary operation of the heat equation benchmark.

By changing the code of the elementary operation `op` (see Listing 1), we can derive algorithms for solving other problems. For example, it is easy to adapt the algorithm for three-dimensional field domain without changing the overall calculation structure (see Listing 2).

Another feature of the test is re-using the values of the temperature field in the calculation of the time layer by Seidel method. From an algorithmic point of view, this builds an association between iterations for  $i$  (Listing 2), which results in non-trivial issues in creating a parallel solution based on OpenMP.

```

1 double seq_alg()
2 {
3   for (int t=1;t<=T;t++)
4     for (int i=1;i<H-1;i++) op(i);
5 }

```

Listing 2. Sequential algorithm for the heat equation benchmark.

### 3. The Templet Runtime System

The Templet actor computing system consists of three main parts: two primitive operations (send and access) and a function for worker threads that process messages (tfunc). In the following listings we illustrate the mechanism for parallel execution of actors in the shared memory using the C++ standard library threads.

A message sending operation source code is shown in Listing 3. To send a message is to place the message in a shared queue and notify a thread, which may expect the message in the empty queue (line 6). The queue is protected by a mutex. It is captured in line 5. The message contains a reference to the actor-destination and the sign of being sent. They are initialized in line 4. Line 3 presents a guard of re-sending the message. Resending the message is an emergency situation, indicating that an error occurred in the application code.

```

1 inline void send(engine*e, message*m, actor*a)
2 {
3   if (m->sending) return;
4   m->sending = true; m->a = a;
5   std::unique_lock<std::mutex> lck(e->mtx);
6   e->ready.push(m); e->cv.notify_one();
7 }

```

Listing 3. Primitive operation 'send'.

The access to the message object during the actor processing procedure is allowed if the function access (see Listing 4) returns "true". The access is granted if (1) the message refers to the actor, which calls the function; (2) the message is not on delivery (line 3). This condition is an invariant during the processing of a message in the context of a particular actor, otherwise the message will not be sent by the send operation. Sending messages is allowed only if the actor has access to the message (see Listing 4).

```

1 inline bool access(message*m, actor*a)
2 {
3   return m->a == a && !m->sending;
4 }

```

Listing 4. Primitive operation 'access'.

The source code of the worker thread is shown in Listing 5. It implements a thread pool pattern [5]. The task in terms of the pattern is a message which is in the state of delivery. The sending field value of the message is true.

The worker thread polls the task from the queue (line 16) and starts the actor's, message processing procedure (recv). The recv procedure is prepared by several steps: (1) determining the message's destination actor (line 19); (2) setting the lock on the actor (line 21); (3) changing the delivery sign of message to sending = false (line 22); activating the recv procedure to process the message (line 23).

```

1 void tfunc(engine*e)
2 {
3   message*m; actor*a;
4
5   for (;;) {
6     {
7       std::unique_lock<std::mutex> lck(e->mtx);
8       while (e->ready.empty()) {
9         e->active--;
10        if (!e->active) {
11          e->cv.notify_one(); return;
12        }
13        e->cv.wait(lck);
14        e->active++;
15      }
16      m = e->ready.front();

```

```

17     e->ready.pop();
18     }
19     a = m->a;
20     {
21     std::unique_lock<std::mutex> lck(a->mtx);
22     m->sending = false;
23     a->recv(m, a);
24     }
25     }
26     }

```

Listing 5. Worker thread's function and 'recv' callback invocation.

Note that the captured locks are released implicitly when the thread leaves the syntactic scope of the object lock `lck`. The actor system computations are stopped when there are no active working threads.

#### 4. Parallel Algorithms for the Heat Equation Test

We implemented three parallel versions of the code in Listings 1, 2. All these versions are driven by the following rules of parallelization:  $t$  is allowed to start iteration  $t$  along the time axis and  $i$  along the space axis ( $t, i$ ), if (1) the iteration ( $t-1, i+1$ ) and ( $t, i-1$ ) have been completed; or (2), if  $t = 1$  and iteration ( $t, i-1$ ) has been completed. We assume that if an iteration has no  $i+1$  or  $i-1$  neighboring iterations, the neighboring iteration is completed. The first iteration of the calculation (1,1) is performed disregarding these conditions. The algorithm stops when  $T$  iterations are performed for each coordinate. The considered computing algorithm can be implemented on the basis of OpenMP, as shown in Listing 6. The idea of parallelization is as follows: either even or odd iterations  $i$  can be calculated simultaneously on each count  $t$ . A strict compliance with the rules of calculation is guaranteed by the additional check in lines 5, 10 and 15 in Listing 6.

```

1 void par_omp()
2 {
3 #pragma omp parallel shared(H,T)
4 {
5 for (int t = 1; t <= (2 * T - 1) + (H - 3); t++){
6
7     if (t % 2 == 1){
8 #pragma omp for schedule(dynamic,1)
9     for (int i = 1; i < H - 1; i += 2)
10        if (i <= t && i > t - 2 * T) op(i);
11    }
12    if (t % 2 == 0){
13 #pragma omp for schedule(dynamic,1)
14    for (int i = 2; i < H - 1; i += 2)
15        if (i <= t && i > t - 2 * T) op(i);
16    }
17 }
18 }
19 }

```

Listing 6. OpenMP based parallel algorithm for the heat equation benchmark.

The actor implementations of the algorithm in listing 1, 2 enable using the rules of parallelism explicitly. For this reason, each space coordinate  $i$  is matched by an actor. There are  $N = H-2$  actors used in both actor algorithms.

In the Templet implementation, the rules of parallelization presented in lines 5-7 of Listing 7. In lines 11 and 12, the actor informs its' neighbors  $i-1$  and  $i+1$  (if any) of the completion of the iteration ( $t, i$ ) by sending messages.

```

1 void recv(message* , actor* a)
2 {
3     int id = (int)(a - as);
4
5     if ((id == 0 || access(&ms[id - 1], a)) &&
6         (id == N - 1 || access(&ms[id], a)) &&
7         (ts[id] <= T)){
8
9         op(id+1); ts[id]++;
10    }

```



```

11  if (id != 0) send(&e, &ms[id - 1], &as[id - 1]);
12  if (id != N - 1) send(&e, &ms[id], &as[id + 1]);
    13 }
    14 }

```

Listing 7. Actor based parallel algorithm for the heat equation benchmark, Templet runtime.

In the Akka implementation, the rules of the parallelization are declared in lines 7-9 of Listing 8. In lines 13-20 the actor informs its neighbors  $i-1$  and  $i+1$  (if any) that the iteration is completed  $(t, i)$  by sending messages. Note that the message handling code in Listings 7 and 8 is implemented identically for the convenience of comparison. The code block in lines 22-24 is stops the computations.

```

1 public void onReceive(Object message) {
2   if (((Integer) message) == id - 1)
3     access_ms_id_minus_1 = true;
4   if (((Integer) message) == id)
5     access_ms_id = true;
6
7   if ((id == 0 || access_ms_id_minus_1) &&
8       (id == N - 1 || access_ms_id) &&
9       (Main.time[id] <= Main.T)) {
10
11     Main.op(id + 1); Main.ts[id]++;
12
13     if (id != 0) {
14       Main.actors[id - 1].tell(id - 1, getSelf());
15       access_ms_id_minus_1 = false;
16     }
17     if (id != Main.N - 1) {
18       Main.actors[id + 1].tell(id, getSelf());
19       access_ms_id = false;
20     }
21   }
22   if (Main.time[id] == Main.T + 1 && id == Main.N - 1) {
23     Main.system.terminate();
24   }
24 }

```

Listing 8. Actor based parallel algorithm for the heat equation benchmark, Akka.

Both actor algorithms have an initialization code, which is not considered in the paper. Complete code of the Actor Templet library and the test cases are available at <https://github.com/Templet-language/newtemplet/>.

## 5. Results

Computational experiments were performed on a computer with an Intel (R) Core (TM) i3-3220T RAM 4GB, Windows 10 x64. C ++ programs compiled in Microsoft Visual 2015. For Java programs we used the JDK version 1.8 and the Akka library version 2.4.17 deployed on the same computer.

The complexity of the problem may be denoted by  $H$ . There are two space-time domain parameters of the calculation:  $W=H*2$ ,  $T=H*2$ . Both depend on  $H$ . Note that  $H$  also determines the granularity of computing. The bigger the  $H$  parameter is, the bigger chunks of data are processed sequentially.

Columns of Table 1 indicate the duration time of the algorithm in seconds:  $T_1^{JAVA}$  - a sequential Java implementation;  $T_1^{NATIVE}$  - a sequential C++ implementation;  $T_p^{AKKA}$  - a parallel Java implementation based on Akka;  $T_p^{TEMPLET}$  - a parallel C ++ implementation based on the Templet;  $T_p^{OPENMP}$  - a parallel C ++ implementation based on OpenMP.

To account for temporary fluctuations, the data presented in Table 1 has been statistically pre-processed. Each value in Table 1 is calculated by series of 19 experiments. The value includes only the significant digits, guaranteeing them from getting into the interval  $[\min, \max]$  with confidence factor of 90% ( $\min$  - minimum,  $\max$  - maximum time in a series of 19 experiments).

Table 2 shows the efficiency of the test implementation based on the proposed runtime system by the example of the implementations based on Akka and OpenMP. The  $E_{AKKA}$  and  $E_{OPENMP}$  values show the percentage of the Templet acceleration of reference implementations based on Akka and OpenMP.

Table 1. Experimental computation time of the heat equation benchmarks.

H	$T_1^{JAVA}, s$	$T_1^{NATIVE}, s$	$T_p^{AKKA}, s$	$T_p^{TEMPLET}, s$	$T_p^{OPENMP}, s$
400	1.79	1.357	0.8	0.42	0.40
500	3.5	3.028	1.2	0.92	0.9
600	6.1	5.238	1.8	1.81	1.77
700	9.8	7.37	2.7	3.01	2.92
800	14.6	12.46	3.7	4.60	4.52
900	20.9	17.73	5.4	6.5	6.4
1000	28.7	21.09	7.6	9.0	8.9

Table 2. Relative efficiency of the Templet runtime system:  $E_{AKKA} = T_p^{AKKA} / T_p^{TEMPLET}$  and  $E_{OPENMP} = T_p^{OPENMP} / T_p^{TEMPLET}$ 

H	$E_{AKKA}, \%$	$E_{OPENMP}, \%$
400	190	95
500	130	99
600	99	98
700	90	97
800	80	98
900	83	98
1000	84	99

Correctness of the parallelization was checked by piecemeal test for equality of the temperature field values calculated by sequential and parallel method. We used equal random initial field values for parallel and sequential method. The physical interpretation of the calculation results was not carried out, since it is beyond the scope of this study.

## 6. Discussion

The experiments confirmed the high efficiency of the proposed simple implementation of the Templet runtime system for actor calculations. The Templet system has only a slight performance gap in tests performed using OpenMP, and for the small H parameter values (400..600) it is not far behind the Akka, or even exceeds it.

The advantage of actor algorithms for the Templet and Akka is the simplicity of implementation and debugging. The parallelism of the system is described in terms of a simple behavior of each individual actor. Using the OpenMP requires the understanding of the global state of computing at each time, resulting in complex boundary conditions of the algorithm cycles in Listing 6.

Our algorithm is not inferior to the implementation of OpenMP. This result is obtained despite the fact that we used the expressive possibilities of C++ Standard Library 11 to simplify the code, neglecting the efficiency. If necessary, it can be optimized by using the primitive compare-and-swap, as proposed in [2], and by work stealing algorithms [6].

The source of the simplicity of our actor model implementation is a departure from the classical approach proposed by Agha [7] and implemented in the famous actor frameworks and languages, for example, Erlang [8], Scala [9], CAF [2] and others.

Agha's approach assumes that the messages are some values that are passed between the actors. They are accumulated in the mailbox - a special system structure associated with the actor. The actor has an access to the message values.

In our implementation, messages are treated as variables that store values. A programmer is not bounded to syntactic rules of access to the message from any actor at any time. However, an access is meaningful and does not lead to violations of logic provided that the function access to the message-actor pair has returned a true value. This approach does not require the implementation of complex logic of copying values between the mailbox and the actor call frame from the runtime system.

The test also showed that despite the fact the native implementation of the test is superior to the Java implementation in terms of performance, the parallel implementation using Akka is the best for the dimensions of the test problem when the H parameter value is 600 or more. This can be attributed to the fact that the scheduling algorithm offered by Akka is more sophisticated than

the one offered by the Templet system, as well as the scheduling algorithm selected to test the implementation based on OpenMP.

## 7. Conclusion

We propose a simple implementation of the actor computation model in C++ 11, and the possibility of its usage in high-performance computing. The test example of the heat equation illustrates the high effectiveness of the proposed implementation. It approximates to the effectiveness of OpenMP, and in some cases is superior to Akka. Apart from that, it reduces the complexity of the coding.

The runtime library is used in the object-oriented Templet language [10] for the implementation of parallel computing patterns. These patterns are used in solving problems of nonlinear dynamics in the design of spacecraft [11].

This work is partially supported by the Russian Foundation for Basic Research (RFBR#15-08-05934-A), and by the Ministry of education and science of the Russian Federation in the framework of the State Assignments program (№ 9.1616.2017/ПЧ).

## References

- [1] Hewitt C. A universal modular ACTOR formalism for artificial intelligence. Proceedings of the 3rd IJCAI. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1973: 235–45.
- [2] Charousset D, Hiesgen R, Schmidt TC. Revisiting Actor Programming in C++. Computer Languages, Systems & Structures 2016; 56: 105–131.
- [3] Lightbend Inc. Akka. URL: <http://akka.io>
- [4] Charousset D, Dominik C, Schmidt TC, Hiesgen R, Wählich M. Native Actors – A Scalable Software Platform for Distributed Heterogeneous Environments. Proceedings of the 4rd ACM SIGPLAN Conference on Systems Programming and Applications (SPLASH '13) Workshop AGERE. New York, NY, USA: ACM, 2013.
- [5] Schmidt DC. Pattern-Oriented Software Architecture. Patterns for Concurrent and Networked Objects. John Wiley & Sons 2013; 2: 700 p.
- [6] Blumofe RD, Leiserson CD. Scheduling multithreaded computations by work stealing. Proceedings of the 35th annual symposium on foundations of computer science (FOCS) 1994: 356–368.
- [7] Agha G, Mason IA, Smith S, Talcott C. Towards a theory of actor computation. Proceedings of CONCUR. Lecture notes on computer science. Heidelberg: Springer-Verlag 1992; 630: 565–579.
- [8] Armstrong J. Erlang – a survey of the language and its industrial applications. Proceedings of the symposium on industrial applications of Prolog (INAP96). Hino 1996: 16–18.
- [9] Haller P, Odersky M. Scala actors: unifying thread-based and event-based programming. Theor Comput Sci 2009; 410(23): 202–220.
- [10] Vostokin SV. Templet: a markup language for concurrent actor oriented programming. CEUR Workshop Proceedings 2016; 1638: 460–468.
- [11] Doroshin AV. Heteroclinic Chaos and Its Local Suppression in Attitude Dynamics of an Asymmetrical Dual-Spin Spacecraft and Gyrostat-Satellites. The Part II – The heteroclinic chaos investigation, Communications in Nonlinear Science and Numerical Simulation 2016; 31(1-3): 171–196.

# Application of the pyramid method in difference solution d'Alembert equations on graphic processor with the use of Matlab

L.V. Yablokova<sup>1</sup>, D.L. Golovashkin<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

The paper proposes a modification of the pyramid method for constructing algorithms for the difference solution of the d'Alembert equation on a graphics processor in the event of a shortage of video memory. The authors demonstrate the effectiveness of the method on the practical example of dividing the grid area into two sub domains. Acceleration reaches the characteristic for the case of a domain entirely located in the video memory. In the article investigated the effectiveness of using the author's approach depending on the height of the pyramid and showed the boundaries of applicability of the proposed modification.

*Keywords:* The method of the pyramids; the grid area; difference solution; computing acceleration

---

## 1. Introduction

The deep interconnectedness of optics and computing technology is due to their mutual influence in the course of which at the turn of the 70-ies and 80-ies of the last century there were two independent branches of science: computer optics associated with the development of numerical methods of calculation and simulation of optical devices on a computer and optical engineering, in which the optical elements are created computing devices. The growth of the relevance of the mentioned industries due to the perfection of the architecture of computers (multithreading, multicore, vectored calculations) and technologies of formation of optical elements (transition from micro to nano-size). The first feature allowed us to use the methods of a rigorous diffraction theory [1] for calculating the nano-sized elements of optical processors, characterized by high computational complexity.

Among the numerical methods of the strict theory of diffraction, FDTD [1], deserving high universality (Maxwell's equations describe all wave electromagnetic processes) and the simplicity of understanding (based on the replacement of derivatives by difference relations) deserved wide popularity. The latter circumstance makes it possible to write the computational procedures of the method in a clear form in the popular language of matrix calculations of MATLAB [2]

Unfortunately, the software implementation of the FDTD method on modern graphics computing devices that provide faster CPU computation by an order of magnitude is encountered, when using this language, with high demands on the amount of video memory: in the production of calculations, it is necessary to use several times more volume than when Work on the central processor. This circumstance is aggravated traditionally by small video memory sizes (up to 2GB in modern budget video cards) in comparison with operating memory (not less than 4GB, even for office computers).

The authors of this publication see the application of the pyramid method as an example of the organization of calculations using the difference scheme Yee [1] on the GPU using CUDA C [3].

## 2. Difference solution of the d'Alembert equation (one-dimensional case) on a graphics processor

Traditionally, the FDTD method is understood to mean exclusively the difference solution of Maxwell's equations, which is not entirely correct. In the early 80's of the last century [1], the difference solution of the d'Alembert equation was also applied to it, which is still being done [1, 4]. We note that when solving the wave equation the problem of video memory shortage is more acute than for Maxwell's equations because of the necessity of finite-difference approximation of second, not first-order derivatives. However, the decision of the wave equation on the GPU seems more promising due to the high efficiency of vectorization of computational procedures [5].

Outlining the concept of the work, the authors decided to dwell on the one-dimensional equation of d'Alembert, seeking to demonstrate the possibilities of the pyramid method on a simple example.

So known [1] the difference scheme for this equation

$$\frac{E_i^{k+1} - 2E_i^k + E_i^{k-1}}{h_t^2} = c^2 \frac{E_{i-1}^k - 2E_i^k + E_{i+1}^k}{h_z^2} \quad (1)$$

is written with respect to the grid function defined on the domain

$D^h = \{(t_k, z_i) : t_k = kh_t, k = 0, 1, \dots, N_t = T/h_t, z_i = ih_z, i = 1, \dots, N_z = L_z/h_z + 1\}$ , where  $E$  the value of the electric field strength is,  $c$  is the speed of light in free space,  $T$  and  $L_z$  are size of the region in time and space.

Below is a fragment of the computational procedure for solving (1) in MATLAB, where  $c_1 = c^2 h_t^2 / h_z^2$ ,  $c_5 = 2\pi c h_t / \lambda$ .

% Placement of grid functions on two time layers in video memory

```
E1=zeros(1,Nz,'gpuArray'); E2=zeros(1,Nz,'gpuArray');
for k=1:2:Nt % Passage through time layers of the grid area through one
E1(2:Nz-1)=2*E2(2:Nz-1)-E1(2:Nz-1)+c1*diff(E2,2); % Calculations on the layer k
E1(2)=sin(c5*k); % Hard radiation condition on the layer k
E2(2:Nz-1)=2*E1(2:Nz-1)-E2(2:Nz-1)+c1*diff(E1,2); % Calculations on the layer k+1
E2(2)=sin(c5*(k+1)); % Hard radiation condition on the layer k+1
End
E=gather(E2); % Transfer of results to RAM
```

For  $N_z = 5 \times 10^7$  and  $N_t = 100$  the duration of calculations on the Intel Core i7 CPU was 57.08 s., On the GeForce GTX TITAN X GPU - 5.55 s. (acceleration of 10.29 times) using MATLAB 2015b and the operating system CentOS 7.2. Both used arrays occupied 762 MB in memory, however, during the computations on the CPU, the memory requirements increased by one and a half time, on the GPU the memory requirements increased threefold. Apparently, with the implementation of calculations for the design  $E1(2:Nz-1)=2*E2(2:Nz-1)-E1(2:Nz-1)+c1*diff(E2,2)$  on the CPU, the execution of the operation of numerical differentiation  $diff(E2,2)$  resulted in allocating additional memory for two copies of the array E2, and the execution on the GPU of the design as a completely required separate area of memory for all the arrays involved and for double copying E2. MATLAB takes about 0.4 gigabytes in RAM, but does not use video memory for its placement. Thus, the execution of the whole algorithm on the CPU was accompanying by the extraction of 1.52 GB. In addition, the execution of the whole algorithm on the GPU was accompanying by the extraction of 2.24 GB. Moreover, if the researcher has a video card with 2 GB of memory (like most popular video processors now) then the organization of calculations on the GPU becomes impossible. In his previous publication [7], using the difference scheme for the Maxwell equations, the CUDA C software tool the authors proposed to solve this problem using the method of pyramids.

### 3. The pyramid method application

Will this be possible in this case, given that MATLAB is not specialized for working with graphics processors and its tools in this area are very meager?

The essence of the mentioned method in the author's modification consists in splitting the grid domain into overlapping sub regions that fit in the video memory completely, with the subsequent organization of communications in the production of vector computations in each sub region separately. In this case, transfers from RAM to video and vice versa are performed not on each time layer, but through a certain number of them  $h$  (the height of the pyramid). This, on the one hand, leads to a reduction in  $h$  the number of communications. On the other hand, to the duplication of arithmetic operations in overlapping fragments of grid subdomains (the form of pyramids is available in the two-dimensional case).

A fragment of the computational procedure implementing this strategy in the case under consideration is presented below, where  $N = \frac{N_z}{2}$ .

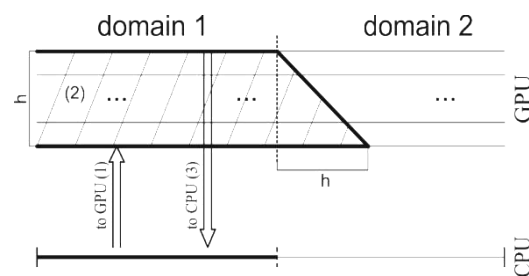


Fig. 1. The scheme of the algorithm of the pyramids to work with the first domain; (1) there is a message to GPU, (2) there is a calculate to GPU, (3) there is a message to CPU.

% creating temporary layers on CPU and GPU

```
E1=zeros(1,Nz); E2=zeros(1,Nz); E1m=zeros(1,h); E2m=zeros(1,h);
E1g=zeros(1,N+h,'gpuArray'); E2g=zeros(1,N+h,'gpuArray');
for t=1:h:Nt % Passage through the pyramids
% work with the left subdomain
E1g=gpuArray(E1(1:N+h)); E2g=gpuArray(E2(1:N+h)); % Forwarding CPU ==> GPU
for k=1:2:h % Calculations inside the first pyramids
E1g(2:N+h-k)=2*E2g(2:N+h-k)-E1g(2:N+h-k)+c1*diff(E2g(1:N+h-k+1),2);
E1g(2)=sin(c5*(t+k-1));
```

```

E2g(2:N+h-k-1)=2*E1g(2:N+h-k-1)-E2g(2:N+h-k-1)+c1*diff(E1g(1:N+h-k),2);
E2g(2)=sin(c5*(t+k));
end
E1(2:N-h)=gather(E1g(2:N-h)); E2(2:N-h)=gather(E2g(2:N-h)); % Forwarding GPU ==> CPU
E1m(1:h)=gather(E1g(N-h+1:N)); E2m(1:h)=gather(E2g(N-h+1:N));
% work with the right subdomain
E1g(1:N+h-1)=gpuArray(E1(N-h+1:Nz)); E2g(1:N+h-1)=gpuArray(E2(N-h+1:Nz));
for t=1:2:h % Calculation by layers of the pyramid
    E1g(t+1:N+h-2)=2*E2g(t+1:N+h-2)-E1g(t+1:N+h-2)+c1*diff(E2g(t:N+h-1),2);
    E2g(t+2:N+h-2)=2*E1g(t+2:N+h-2)-E2g(t+2:N+h-2)+c1*diff(E1g(t+1:N+h-1),2);
end
E1(N+1:Nz-1)=gather(E1g(h+1:N+h-2)); % Forwarding GPU ==> CPU
E2(N+1:Nz-1)=gather(E2g(h+1:N+h-2));
E1(N-h+1:N)=E1m(1:h); E2(N-h+1:N)=E2m(1:h); % Replenishment of the result
end

```

In the course of experiments with the new algorithm, the dependence of the calculation time on the height of the pyramid. The Table 1 contains the results.

Table 1. The dependence of the calculation duration of the calculation of the height of the pyramid.

Height of the pyramid, $h$	Computation time (s)	Acceleration
2	53.02	1.08
4	29.58	1.93
10	15.49	3.7
20	10.75	5.31
50	7.9	7.23

#### 4. Conclusion

Thus, the method of pyramids can be effectively using in arranging calculations for solving difference equations with the help of MATLAB on graphic processors in the case when arrays storing values of grid functions do not fit into video memory as a whole. The development of the proposed algorithm for cases of large dimensions will be the next stage of the authors' research in this direction.

#### Acknowledgements

The research leading to these results has received funding from the Russian Science Foundation grant №16-47-630560-r\_a.

#### References

- [1] Taflove A, Hagness S. Computational Electrodynamics: The Finite-Difference Time-Domain Method. Boston: Aerotech House Publishers, 2005; 1006 p.
- [2] Elsherbeni A, Demir V. The Finite-Difference Time-Domain Method for Electromagnetics with MATLAB Simulations. Scitech Publishing Inc. 2009; 426 p.
- [3] Grigorjev IS, Mejlihov EZ. Physical Values: Reference Book. Moscow: Energoatomizdat Publisher, 1991; 1232 p. (in Russian).
- [4] Malysheva SA, Golovashkin DL. Realization of the difference solution of the Maxwell equations on graphic processors by the pyramid method. Computer Optics 2016; 40(2): 179–187. DOI: 10.18287/2412-6179-2016-40-2-179-187.
- [5] Kozlova ES, Kotlyar VV. Simulation of the propagation of a short two-dimensional pulse of light. Computer Optics 2012; 36(2): 158–164.
- [6] Vorotnikova DG, Golovashkin DL. Difference solution of the wave equation on graphical processors with repeated use of pairwise sums of the differential pattern. Computer Optics 2017; 41(1): 134–138. DOI: 10.18287/2412-6179-2017-41-1-134-138.
- [7] Vorotnikova DG, Golovashkin DL. Algorithms with "long" vectors for solving grid equations of explicit difference schemes. Computer Optics 2015; 39(1): 87–93. DOI: 10.18287/0134-2452-2015-39-1-87-93.

# S.B. Popov, Doctor of Engineering (Commemorating the 60<sup>th</sup> Birth Anniversary)

V.O. Sokolov<sup>1</sup>

<sup>1</sup>*Samara Research Center of the Russian Academy of Sciences, 3a, Studenchesky pereulok, 443001, Samara, Russia*

---

## Abstract

The paper focuses on key scientific and academic accomplishments of Sergei Borisovich Popov, Doctor of Engineering.

*Keywords:* Doctor of Engineering; scientific research automation; computer vision; computer vision system; parallel image processing; distributed Big Data processing

---

## 1. Introduction

This year Sergey Borisovich Popov, Doctor of Engineering, leading researcher of the Laboratory of Mathematical Methods of Image Processing of the Image Processing Systems Institute of the Russian Academy of Sciences (IPSI RAS) – the Branch of the “Crystallography and Photonics” Federal Research and Development Center of the RAS (FRDC RAS), and in addition to his other duties, professor of the Department of Engineering Cybernetics of S.P. Korolyov Samara National Research University is celebrating his 60<sup>th</sup> Birth Anniversary. The paper focuses on key scientific and academic accomplishments of S.B. Popov.

## 2. Kuibyshev Aviation Institute

In 1981, S.B. Popov graduated the Faculty of Systems Engineering of Kuibyshev Aviation Institute named after academician S.P. Korolyov (KuAI, currently – S.P. Korolyov Samara National Research University) majoring in Applied Mathematics. He worked at KuAI from 1981 (in 1992 the Institute was renamed into S.P. Korolyov Samara State Aerospace University, SSAU) first in the capacity of an engineer and then – a senior engineer and a junior researcher. From January 1993 till December 1998, he worked in the capacity of a teaching assistant at the Department of Engineering Cybernetics at S.P. Korolyov Samara State Aerospace University (SSAU, formerly KuAI).

His graduate thesis was related to the research in which he was engaged over a long period of his activity and to which he still continues to pay much attention, i.e. scientific research automation using computer vision techniques. Within the framework of joint activities with A Department of Lebedev Physical Institute of the USSR Academy of Sciences (LPI RAS), he has developed new software for the Automated control system for spherical optical surfaces (ACSOS) "Shadow" [1, 2].

Being a part of the Research and Development Laboratory of KuAI-SSAU, he participated in the development of algorithmic and software support of the image processing system based on the PC image processing automated system [3-6].

In his research S.B. Popov developed methods of efficient organization of computing processes in image processing which paralleled these processes by combining sequence image operations into a pipeline [7-9]. These studies provided the basis for his Ph.D. thesis in Engineering “Modeling of Data Stream Processing Networks and Methods of Organizing Two-Dimensional Data Sets in Image Processing.” The thesis has presented efficient methods of image stream processing in PC-based computing systems and developed image processing software tools based thereon combining high performance of Big Data processing, relatively low value, scalability, feasibility in development and implementation of new software modules, and high adjustability to different formats of storing images. The Candidate degree in Engineering was conferred by the Dissertation Council of S.P. Korolyov Samara State Aerospace University (SSAU) on May 15, 1998 and approved by the State Higher Attestation Committee of the Russian Federation on November 20, 1998.

## 3. Image Processing Systems Institute of the Russian Academy of Sciences

From August 1998, S.B. Popov has moved to the Image Processing Systems Institute of the Russian Academy of Sciences (IPSI RAS) [10] where he has been working until present first in the capacity of a senior researcher and then, from 2013, as a leading researcher.

From 2000, he was actively involved into development of the Regional Center of High-Performance Information Processing in Samara Scientific Center of the Russian Academy of Sciences (SSC RAS) [11], developed applied software for parallel multivariable data processing on high-performance computers for scientific research in the field of Computer Optics [12,13] and Image Processing [14,15], and provided support to educational process in training programs associated with training of specialists in the field of parallel high-performance computing [16-19].

S.B. Popov’s research interests have gradually expanded in the following research areas: Big Data Image Processing [20], Mathematical Modeling of Parallel Computing, and Software for Distributed and Parallel Systems, in particular, Applied Software for High-Performance Computers [21-25].

Additionally, S.B. Popov is fully engaged in developing automation systems for complex research and tests, and in building original computer vision systems both in traditional applications (for recognition of identification numbers of railway tanks) and for unique laboratory investigations [26, 27].

In particular, under the guidance and with the active involvement of S. B. Popov, the following software tools have been developed: Automated System of Data Control, Collection and Processing in Experiments in a Wind Tunnel with a Climate chamber dynamometer in the Technical Development Directorate of JSC AVTOVAZ (Togliatti, 2002-2003), Railway Tanks Registration System in Samara-Terminal Ltd. (Syzran, 2004-2005), Computer Vision System to Control Laboratory Analysis on Quantifying Gel Particles in Polymer Solution in Kuibyshevazot company (Togliatti, 2005), Automated Computer Vision System to Control Identification Numbers of Tank Wagons in JSC Ufa Refinery (Ufa, 2008-2009), and up-grading a control system with an all-wheel drive chassis dynamometer system by Schenck in a wind tunnel (2012) for the Research and Development Centre of JSC AVTOVAZ (Togliatti).

Scientific tasks for building mathematical models and control algorithms for the all-wheel drive chassis dynamometer system by Schenck [28] and complex humidity- and temperature-control systems, being a part of the wind tunnel for testing light motor vehicles, LCVs and minivans, have been successfully solved in applications developed for the Research and Development Centre of JSC AVTOVAZ.

When creating computer vision systems for Samara-Terminal and Ufa Refinery, some original algorithms have been developed for recognition of identification numbers [29, 30] on such complex moving objects as tank wagons for transportation of contaminated crude oil and fuel oil under daylight and artificial daylight conditions with significant changes in surveillance parameters within 24 hours and throughout the year depending on a season [31-34].

New methods of thresholding and analysis of binary images being obtained therewith have been developed for the computer vision system required for laboratory tests on quantifying gel particles in polymer solution for Kuibyshevazot company that operate under conditions of a weak image-contrast ratio and in presence of considerable disturbances [26, 35]. The system used instead of an observer while carrying-out this analysis has reduced dramatically a psychovisual load on lab staff, provided documenting capability of performed lab testing, and improved accuracy and certainly of quantifying gel particles in polymer solution that finally helped adjust a process of manufacturing industrial threads and cord fabrics.

The successful implementation and long-term operation of the above mentioned computer vision systems [36, 37] are based on a human-operator base priority principle. Computer vision capacity provided therein doesn't remove the operator out of the system, but it makes him released from stress associated with the fear not to notice anything or not to manage with fixing an important event in monitoring a long-lasting dynamic process, thus providing a convenient environment for visual control and editing of an automatically generated list of tanks or fragments of occurring inhomogeneity of the laboratory analysis process.

Projects designed with the involvement of S.B. Popov have found use and successfully operated in the Central Specialized Design Bureau "Progress," FIAT Research Center (Italy), Intel (USA), and LG (South Korea) and are currently used in academic activities of Samara University.

In his scientific research S. B. Popov brings up one of the most important issues in using IT-equipment – mapping of computational mathematics problems onto the architecture of computing systems which was identified by academician G.I. Marchuk as a fundamental academic research area briefly called a mapping issue.

In particular, solving the issue of imaging computational problems onto the parallel or distributed architecture of computing systems is the most relevant issue since a focus area in improving the efficiency of the use of computing facilities is the use of parallel computing techniques [38]. The basic approach to solve the imaging issue is the analysis of a computational problem that identifies parallelism and opportunity to use distributed data and is performed on the basis of mathematically equivalent transformations of an information structure model of the solution algorithm for the problem being investigated or, more generally, of an IT model for solving the problem.

Particularly this very approach was used by S.B. Popov [39, 40] in his Doctoral thesis "Modeling and development of the structure of distributed large-size image processing systems based on the dynamic organization of data" in profile 05.13.18 – Mathematical Modeling, Numerical Computing and Software Systems (consultant – Corresponding Member of the RAS V.A. Soifer [41]), which was successfully defended at the end of 2010 in the Dissertation Council of SSAU. The thesis based on the dynamic management method, processing iterator models, and equivalence transformation rules has solved the problem of modeling and structuring of distributed image processing systems with different types of parallelism. A set of obtained scientific results is to be the solution of the fundamental scientific problem – the mapping issue for a widely used class of problems of mathematical image processing. He took his Doctoral degree in Engineering in 2011.

For the time being, in his papers S. B. Popov investigates characteristic features of the Earth's remote sensing data in the frame of Big Data and new opportunities, challenges, and research areas arising therefrom [42], considers advantages of using the Big Data technique when building distributed systems for processing multidimensional spatially dependent data, in particular, transparent expansion of functionality of such systems and improvement of their quality [43], and definition of new smart properties [44].

S.B. Popov took part in implementation of dozens of grants, state-financed and contractual research projects and was an authorized person responsible for several large research and development projects. He is also a leader of some grants financed by the Russian Foundation for Basic Research.

The projects designed with the involvement of S.B. Popov were exhibited at the Russian National Exhibition in China (November 8-13, 2006, Beijing) and were awarded with certificates of the First and Third District Exhibitions of Business Angels and Innovators (2003 – Nizhny Novgorod, 2005 – Samara).



He is the author or co-author of more than 100 research papers, including three monographs and 25 articles in the leading journals, such as *Technical Physics*, *Automation in Industry*, *Pattern Recognition and Image Analysis*, *Computer Optics*, etc., and 5 invention certificates received. S. B. Popov is one of the most active reviewers of the scientific journal “*Computer Optics*” [45, 46]. Besides, thanks to his efforts and based on the results in 2015, the journal has joined the rank of the best half (the second quartile) of the journals indexed in the Scopus database in its all focus areas.

In 2013, S. B. Popov was awarded with the Letter of Acknowledgement of Samara Regional Duma (regional legislative body) “For Strong Contribution to Development of the Federal State Budgetary Institution of Education – Image Processing Systems Institute of the RAS.”

Popov S.B. was the winner of the regional Science and Technology Award in 2014 for his research work “Development of computer vision systems for the automation of high-tech manufacturing and logistics facilities in Samara Region”.



Fig. 1. Sergei Bolrisovich Popov at the True Positive Conference.

#### 4. Teaching activities

S.B. Popov successfully continues his employment at the academic institute alongside with his teaching activities – from January 1999 he was also employed as an assistant professor of the Department of Engineering Cybernetics, SSAU, and since 2011 he has been working in the capacity of a professor of the Department of Engineering Cybernetics, SSAU.

He was awarded with the academic title of an associate professor of the Department of Engineering Cybernetics, SSAU, in accordance with the order of the Federal Education and Science Supervision Service of the Russian Federation dated 26 October 2006 No. 2212/1179-D.

S. B. Popov pays great attention to students and young researchers’ engagement in scientific activities; the diploma theses of his advised students were recognized as the best ones many times [6, 47].

In particular, he (co-authored) has written five chapters of the monograph “*Methods of Computer Image Processing*” successfully gone into two editions in 2001 and 2003 [48] in the Publishing House “Fizmatlit” (Moscow) and recommended by the Ministry of Education of the Russian Federation as a study guide for students who learn Applied Mathematics. In 2010, the monograph was supplemented with new chapters and translated into English [49, 50].

In 2006, on request of the SSAU’s Innovative Educational Program “*Development of the Center of Excellence and Training of World-Class Specialists in Aerospace and Geoinformation Technologies*” implemented within the framework of the Education National Priority Project, 4 manuals for graduate students were published [51-54].

New lecture courses developed at different times, such as *Network Programming Techniques*, *Parallel Programming*, *Parallel Programming Software Tools and Technologies*, *Data Mining*, *Big Data Processing Methods and Techniques* (in the framework of the Professional Development Programme) should be noted, too.

#### 5. Conclusion

In conclusion, I would like to wish Sergei Borisovich Popov good health, high performance, and talented students in order to continue his research and to obtain new results!

#### References

- [1] Arefyev EYu, Demidov EV, Zhivopistsev ES, Pelipenko VI, Popov SB, Sisakyan IN, Soifer VA. Automated control system for spherical optical surfaces (ACSOS) “Shadow”. Preprint 245, Institute for Physics of the USSR Academy of Sciences (LPI RAS), 1982. (in Russian)
- [2] Arefyev EYu, Zhivopistsev ES, Popov SB, Sisakyan IN, Soifer VA. Automated system of technological control of optical surfaces on the basis of micro-computer Electronica-60. Automation of experimental research. Kuybishev: KuAI, 1983; 116–121. (in Russian)
- [3] Bambulevich KE, Vasin AG, Maslov AM, Popov SB, Sergeev VV, Soifer VA. Image Processing Software IPS 1.0 RSX11M. USSR’s State Fund of algorithms and programs, No. 50850000495, 1985. (in Russian)

- [4] Arefyev EYu, Bagbaya ID, Ovchinnikov KV, Popov SB, Sisakyan IN, Soifer VA. Experiments on reconstructive tomography using an automated image processing system. *Computer Optics* 1987; 2: 31–35. (in Russian)
- [5] Arefyev EYu, Golub MA, Ovchinnikov KV, Popov SB, Sisakyan IN, Soifer VA, Tikhonov DN, Khramov AG, Shamalova GV. Verification of the phase microrelief of computer optics elements. *Soviet Physics: Technical physics* 1990; 35(6).
- [6] Popov SB, Khasanov IA. Investigation of Modifications of Algorithms for Fractal Image Encoding. *Pattern Recognition and Image Analysis* 1996; 6(1): 174.
- [7] Glumov NI, Myasnikov VV, Popov SB, Raudin PV, Sergeyev VV, Frolova NI, Chernov AV. Some Application Shells of Image Processing for IBM PCs. *Pattern Recognition and Image Analysis* 1996; 6(2): 372.
- [8] Popov SB, Sergeyev VV, Frolova NI. Architecture of the Software for Image Processing in OS/2. *Pattern Recognition and Image Analysis* 1996; 6(2): 432.
- [9] Popov SB. Scalable Automatic System of Image Processing with the Possibilities of Adaptation and Distributed Processing. *Pattern Recognition and Image Analysis* 1998; 8(3): 380–381.
- [10] Kolomiets EI. Analysis of the scientific and organizational results of the Image Processing Systems Institute of the RAS. *CEUR Workshop Proceedings* 2015; 1490: 309–326.
- [11] Shorin VP, Soifer VA, Sanchugov VI, Kazansky NL, Fursov VA, Kravchuk VV, Popov SB. Development of the Samara Network for Science and Education and High Performance Computing Center. *Proceedings of conf. "Telematics 2002"*, 2002; 162–163. (in Russian)
- [12] Volotovskiy SG, Kazansky NL, Popov SB, Serafimovich PG, Soifer VA, Fursov VA. Methodological aspects of parallel applications development in the field of computer optics and image processing. *Proceedings of conf. "Telematics 2002"*, 2002; 163–165. (in Russian)
- [13] Kazanskiy NL, Serafimovich PG, Popov SB, Khonina SN. Using guided-mode resonance to design nano-optical spectral transmission filters. *Computer Optics* 2010; 34(2): 162–168. (in Russian)
- [14] Popov SB, Soifer VA, Tarakanov AA, Fursov VA. Cluster technology for the formation and parallel filtering of large images. *Computer Optics* 2002; 23: 75–78. (in Russian)
- [15] Volotovskiy SG, Kazanskiy NL, Popov SB, Serafimovich PG. Performance Analysis of Image Parallel Processing Applications. *Computer Optics* 2010; 34(4): 567–572. (in Russian)
- [16] Kravchuk VV, Popov SB, Privalov AYU, Fursov VA, Shustov VA. Introduction to programming for parallel computers and clusters. Ed by Fursov VA. Samara: SSAU, 2000.
- [17] Soifer VA, Sergeyev VV, Popov SB, Myasnikov VV. The theoretical foundations of digital image processing. Samara: SSAU, 2000. (in Russian)
- [18] Popov SB, Skuratov SA, Fursov VA. Basics for working on a computing cluster. Samara: Samara Scientific Center of RAS & SSAU, 2004. (in Russian)
- [19] Kazanskiy NL, Popov SB, Serafimovich PG. Organization of computational experiment on high-performance systems. Samara: IPSI RAS, 2010. (in Russian)
- [20] Gashnikov MV, Glumov NI, Popov SB, Segreyev VV, Farberov EA. Software System for Transmitting Large-Size Images via the Internet. *Pattern Recognition and Image Analysis* 2001; 11(2): 430–432.
- [21] Drozdov MA, Zimin DI, Popov SB, Skuratov SA, Fursov VA. Cluster technology for determining repair filters and large image processing. *Computer Optics* 2003; 25: 175–182. (in Russian)
- [22] Nikonorov AV, Popov SB, Fursov VA. The principle of consistency of estimates in the problem of identification of color reproduction models. *Computer Optics* 2002; 24: 148–151. (in Russian)
- [23] Nikonorov AV, Popov SB, Fursov VA. Applying the estimates consistency principle in the problem of identification of color reproduction models. *Proceedings of the Samara Scientific Center of the Russian Academy of Sciences* 2002; 4(1): 159–164. (in Russian)
- [24] Nikonorov AV, Popov SB, Fursov VA. Identifying Color Reproduction Models. *Pattern Recognition and Image Analysis* 2003; 13(2): 315–318.
- [25] Nikonorov AV, Popov SB, Fursov VA. Computational aspects of the implementation of color reproduction model identification. *Proceedings of the Samara Scientific Center of the Russian Academy of Sciences* 2003; 5(1): 67–73. (in Russian)
- [26] Kazansky NL, Popov SB. A machine vision system for counting the number of gel particles in a polymer solution. *Computer Optics* 2009; 33(3): 325–331. (in Russian)
- [27] Abulhanov SR, Popov SB, Ivliev NA, Podlipnov VV. Device for Control of Apertures Surface of Pipes of Oil Assortment. *Procedia Engineering* 2017; 176: 645–652.
- [28] Ignatov NA, Kazansky NL, Kornev YuA, Popov SB. Modeling of dynamometer control system. *Journal of Samara State Technical University, Ser. Physical and Mathematical Sciences* 2005; 38: 115–121. (in Russian)
- [29] Volotovskii SG, Kazanskiy NL, Popov SB, Khmelev RV. Recognition of the numbers of railway tanks using fast localization and modification of the algorithm for comparing an object with a template using the Hausdorff metric. *Survey of Applied and Industrial Mathematics* 2005; 12(3): 714–715. (in Russian)
- [30] Volotovskii SG, Kazanskiy NL, Popov SB, Khmelev RV. Machine Vision System for the Recognition of Numbers of Railway Tank-cars with the Use of Modified Correlator in the Hausdorff Metric. *Computer Optics* 2005; 27: 177–184. (in Russian)
- [31] Bulanov AP, Volotovskii SG, Kazanskiy NL, Popov SB, Khmelev RV, Shumakov SM. Vision System for Registration of Railway Tank-cars. *Automation in industry* 2005; 6: 57–59. (in Russian)
- [32] Volotovskii SG, Kazanskiy NL, Popov SB, Khmelev RV. Machine Vision System for Registration of Oil Tank Wagons. *Pattern Recognition and Image Analysis* 2005; 15(2): 461–463.
- [33] Popov SB. The use of structured lighting in computer vision systems. *Computer Optics* 2013; 37(2): 233–238.
- [34] Popov SB. The intellectual lighting for optical information-measuring systems. *Proc. SPIE* 9533, 2015; 95330P. DOI:10.1117/12.2181168.
- [35] Kazanskiy NL, Popov SB. Machine Vision System for Singularity Detection in Monitoring the Long Process. *Optical Memory and Neural Networks (Information Optics)* 2010; 19(1): 23–30.
- [36] Kazanskiy NL, Popov SB. The distributed vision system of the registration of the railway train. *Computer Optics* 2012; 36(3): 419–428. (in Russian)
- [37] Kazanskiy NL, Popov SB. Integrated Design Technology for Computer Vision Systems in Rail-way Transportation. *Pattern Recognition and Image Analysis* 2015; 25(2): 215–219.
- [38] Popov SB. The concept of distributed storage and parallel processing of large-size images. *Computer Optics* 2007; 31(4): 77–85. (in Russian)
- [39] Popov SB. Modeling the task information structure in parallel image processing. *Computer Optics* 2010; 34(2): 231–242. (in Russian)
- [40] Kazanskiy NL, Popov SB. Distributed storage and parallel processing for large-size optical images. *Proc. SPIE* 8410, 2012; 84100I. DOI:10.1117/12.928441.
- [41] Sokolov VO. On the 70th birthday of corresponding member of the Russian academy of sciences Victor A. Soifer. *CEUR Workshop Proceedings* 2015; 1490: 1–8.
- [42] Popov SB. The Big Data methodology in computer vision systems. *CEUR Workshop Proceedings*, 2015; 1490: 420–425. DOI: 10.18287/1613-0073-2015-1490-420-425.
- [43] Protsenko VI, Serafimovich PG, Popov SB, Kazanskiy NL. Software and hardware infrastructure for data stream processing. *CEUR Workshop Proceedings*, 2016; 1638: 782–787. DOI: 10.18287/1613-0073-2016-1638-782-787.
- [44] Ilyasova NYu, Kupriyanov AV, Popov SB, Paringer RA. Particular usage characteristics of BIG DATA in medical diagnostics tasks. *Highly available systems* 2016; 12(1): 45–52. (in Russian)

- [45] Kolomiets EI. Analysis of activity of the scientific journal Computer Optics. CEUR Workshop Proceedings 2015; 1490: 138–150.
- [46] Sokolov VO. Contribution of Samara scientists into Computer Optics journal development. CEUR Workshop Proceedings 2016; 1638: 194–206. DOI: 10.18287/1613-0073-2016-1638-194-206.
- [47] Nikonorov AV, Popov SB. Comparative analysis of color reproduction models in offset color printing. Computer Optics 2002; 23: 79–83. (in Russian)
- [48] Methods of computer image processing. Ed by Soifer VA. Moscow: “Fizmatlit” Publisher; 2003. (in Russian)
- [49] Myasnikov VV, Popov SB, Sergeyev VV, Soifer VA. Computer Image Processing, Part I: Basic concepts and theory. Ed by Victor A. VDM Verlag, 2010.
- [50] Gashnikov MV, Glumov NI, Popov SB, Sergeyev VV. Image Compression. Computer Image Processing, Part II: Methods and algorithms. Ed by Soifer VA. VDM Verlag, 2010: 87–160.
- [51] Introduction to digital signal and image processing: Mathematical models of images. Ed by Soifer VA. Samara: SSAU, 2006. (in Russian)
- [52] Introduction to digital signal and image processing: Criteria for image quality and error of image discrete representation. Ed by Soifer VA. Samara: SSAU, 2006. (in Russian)
- [53] Introduction to digital signal and image processing: Improving the quality and estimation of geometric parameters of images. Ed by Soifer VA. Samara: SSAU, 2006. (in Russian)
- [54] Sergeyev VV, Gashnikov MV, Glumov NI, Popov SB. Methods of compression of digital signals and images. Samara: SSAU, 2006. (in Russian)

**Table of Contents**  
Data Science

1. The information-mathematical system of the borrower's solvency prediction V.A. Alekseeva, Yu.E. Kuvayskova.....	1-4
DOI: 10.18287/1613-0073-2017-1903-1-4	
2. Joint use of neural network technologies and decision trees for logical patterns exploration in data V.N. Gridin, V.I. Solodovnikov.....	5-10
DOI: 10.18287/1613-0073-2017-1903-5-10	
3. Control of component alterations according with the target efficiency of data processing and control system V.E. Gvozdev, M.B. Guzairov, D.V. Blinova, A.S. Davlieva.....	11-16
DOI: 10.18287/1613-0073-2017-1903-11-16	
4. Generalized Model of Pulse Process for Dynamic Analysis of Sylov's Fuzzy Cognitive Maps R.A. Isaev, A.G. Podvesovskii.....	17-23
DOI: 10.18287/1613-0073-2017-1903-17-23	
5. Capabilities of the adaptive regression modeling package SSOR G.R. Kadyrova, T.E. Rodionova.....	24-27
DOI: 10.18287/1613-0073-2017-1903-24-27	
6. The analysis of technical object functioning stability as per the criterion of monitored parameters multivarite dispersion V.N. Klyachkin, I.N. Karpunina.....	28-31
DOI: 10.18287/1613-0073-2017-1903-28-31	
7. The use of aggregate classifiers in technical diagnostics, based on machine learning V.N. Klyachkin, Yu.E. Kuvayskova, D.A. Zhukov.....	32-35
DOI: 10.18287/1613-0073-2017-1903-32-35	
8. Investigation of the genetic algorithm possibilities for retrieving relevant cases from big data in the decision support systems K. Serdyukov, T. Avdeenko.....	36-41
DOI: 10.18287/1613-0073-2017-1903-36-41	
9. Big Data incorporation based on Open Services Provider for distributed enterprises O.L. Surmin, P.V. Sitnikov, A.V. Ivaschenko, N.Yu. Ilyasova, S.B. Popov.....	42-47
DOI: 10.18287/1613-0073-2017-1903-42-47	
10. Big Data Analysis for Demand Segmentation of Small Business Services by Activity in Region V.M. Ramzaev, I.N. Khaimovich, V.G. Chumak.....	48-53
DOI: 10.18287/1613-0073-2017-1903-48-53	
11. Matrix model of data and knowledge presentation to revealing regularities of the fluid flow regime in a pipeline based on hydrodynamics parameters A. Yankovskaya, A. Travkov.....	54-58
DOI: 10.18287/1613-0073-2017-1903-54-58	
12. Prediction of Cluster System Load Using Artificial Neural Networks Y.S. Artamonov.....	59-63
DOI: 10.18287/1613-0073-2017-1903-59-63	
13. Network disruption prediction based on neural networks D.S. Taimanov.....	64-67
DOI: 10.18287/1613-0073-2017-1903-64-67	
14. Automated system for modeling traffic of multiservice networks B.Ya. Likhtsinder, A.V. Kharkovsky, S.Yu. Antsinov.....	68-71
DOI: 10.18287/1613-0073-2017-1903-68-71	
15. Semantic Analysis of Text Data with Automated System O. Chernenko, O. Gordeeva.....	72-75
DOI: 10.18287/1613-0073-2017-1903-72-75	

16. Modeling of online social networks for automated monitoring system Yu.B. Savva, Yu.V. Davydova.....	76-79
DOI: 10.18287/1613-0073-2017-1903-76-79	
17. Development and research of algorithms for clustering data of super-large volume I.A. Rytsarev, A.V. Blagov, M.I. Khotilin.....	80-83
DOI: 10.18287/1613-0073-2017-1903-80-83	
18. Research and analysis of links in social networks M.I. Khotilin, A.V. Blagov, I.A. Rytsarev.....	84-87
DOI: 10.18287/1613-0073-2017-1903-84-87	
19. The analysis of profiles on social networks V.A. Bakayev, A.V. Blagov.....	88-91
DOI: 10.18287/1613-0073-2017-1903-88-91	
20. Algorithms of the information stimulation system of Russian citizens' socio-optimal actions M.I. Geraskin.....	92-99
DOI: 10.18287/1613-0073-2017-1903-92-99	
21. Design patterns of database models as storage systems for experimental information in solving research problems D.E. Yablokov.....	100-106
DOI: 10.18287/1613-0073-2017-1903-100-106	
22. Comparative Analysis of CRM-systems E.Z. Glazunova, V.V. Kovelskiy.....	107-109
DOI: 10.18287/1613-0073-2017-1903-107-109	
23. The role of subprocess-connector in business process modeling K. Shoilekova, K. Grigorova, E. Malysheva.....	110-114
DOI: 10.18287/1613-0073-2017-1903-110-114	
24. Application of Data Mining and Process Mining approaches for improving e-Learning Processes K. Grigorova, E. Malysheva, S. Bobrovskiy.....	115-121
DOI: 10.18287/1613-0073-2017-1903-115-121	
25. Teacher attitudes in the design of learning activities through technology R. Martinez-Lopez, C. Yot, M. Sacchini.....	122-127
DOI: 10.18287/1613-0073-2017-1903-122-127	
26. Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources A.A. Zarubin, A.R. Koval, V.S. Moshkin, A.A. Filippov.....	128-134
DOI: 10.18287/1613-0073-2017-1903-128-134	

# Preface

Sergey Popov<sup>1</sup>, Dmitry Savelyev<sup>2</sup>

<sup>1</sup> Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics", Russian Academy of Sciences, Samara, Russia

<sup>2</sup> Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

Session «Data Science» was held at the 3rd International Conference on Information Technology and Nanotechnology - 2017 (ITNT-2017) in Samara, Russia, April 25–27, 2017 (<http://ru.itnt-conf.org/itnt17ru/>). This volume contains the papers presented at this session, covering a wide variety of topics such as data mining, big data technologies and systems, machine learning, neural network technologies, data and knowledge representation, natural language processing.

The goal of the ITNT-2017 Conference was to discuss problems of fundamental and applied research in information technology and nanotechnology, including:

- Computer Optics;
- Diffractive Nanophotonics;
- Image Processing;
- Computer Vision;
- Mathematical Modeling;
- Data Science.

Scientists from Austria, Belarus, Bulgaria, Denmark, Germany, Great Britain, India, Iraq, Mexico, Moldova, Russia, Spain, USA, and Finland presented over 330 reports at the ITNT-2017 Conference.

The main proceedings of the conference will be published in *Procedia Engineering* (Elsevier BV). The proceedings of the seminar, not included in *Procedia Engineering*, were selected for this volume.

We are grateful to everybody who has contributed to the seminar and look forward to meeting you again at future events. Heartfelt thanks are due to all authors, reviewers and delegates. A special thank you is due to the team of organizers for making the seminar successful and this publication possible.

## Guest Editors

- Michael Sobolewski, Polish-Japanese Institute of IT, Poland.
- Sergey Popov, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics", Russian Academy of Sciences, Samara, Russia
- Denis Kudryashov, Samara National Research University, Samara, Russia

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov – Samara National Research University, Russia

## Vice-chairs

- Vladimir Bogatyrev – Samara National Research University, Russia
- Nikolay Kazanskiy – Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Russia
- Eduard Kolomiets – Samara National Research University, Russia
- Alexander Kupriyanov – Samara National Research University, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Russia

# The information-mathematical system of the borrower's solvency prediction

V.A. Alekseeva<sup>1</sup>, Yu.E. Kuvayskova<sup>1</sup>

<sup>1</sup>*Ulyanovsk State Technical University, Severny Venec, 32, 432027, Ulyanovsk, Russia*

---

## Abstract

The paper is about research of the algorithms, methods of the classification and prediction objects' groups, depiction of the information-mathematical system, which is created on these algorithms' basis. They use variety methods of the machine learning and their compilations – aggregative classifier, all of these is for the solution of the classification's problem, particularly borrower's solvency prediction. This helps to make previous preparation of the source data, which also contents discretisation, missed data's recovery and detection of the important factors for statistics, how to use these methods of the classification and create cogeneration models, how to analyze quality of these models using statistical measures, to predict objects' groups.

*Keywords:* machine learning; aggregative classifier; statistical data analysis; classification; solvency; prediction

---

## 1. Introduction

Consider the problem of objects' binary classification [1], in which every object  $K_i (i = 1, \dots, N)$  is characterized  $m$ -measured vector of the features  $(X_1 \dots X_m)$ , which can be numeral or nonnumeric value and can create sample for the further researching. Using value of these features we need to predict value of the binary characteristics of objects  $y$ . Example for this type of matter is matters of technical diagnostics and classification of the object's condition [7,8], detection the fact of emission or absence the only signal, normal or abnormal element's condition etc.

There is a solution of the problem of binary classification by the example of credit score, which is in borrower creditworthiness assessment [10].

The increase the amount of debt on loans, increase the risk of loan default, also rivalry on the credit market – all of these need improvement of known techniques of the assessment and prediction of the borrower's solvency with the aim to more accurate assessment of the credit risk and making the right decision in the case of issuance of credit. Known methods cannot help finding more accurate models for solvation this problem.

Information-mathematical system was made in the aim of decrease the amount of borrowers' debts and to provide return of the credits, this system allow assessing borrower's creditworthiness at the stage of making decision of issuance of credit. For the assessing creditworthiness was used methods of machine learning [2] with aggregation different classifiers on the basis of decision trees, neural net, discriminant analysis, Bayesian classifier, SVM, logit regression etc.

## 2. Aggregative classifiers

Nowadays there are a lot of models and methods for the solution the problem of prediction the class of objects. Next methods was used for the analysis of credit risks' assessment: decision trees [14], neural nets [13], discriminant analysis [2], Bayesian classifier [5], SVM, logit regression [6], bagging decision trees, fuzzy inference models [10], created function method. Every method has advantages and disadvantages. For example, there is no possibility to use created function method for data's prediction, which set of characteristic values disagrees at least with one set from learning sample. For using Bayesian attitude it needs to bring given data to interval scale to variables were discrete, otherwise important information will be lost. There is no general model, which one can help assess belonging of the object to one of classes with high accuracy.

Depends on concrete case every method of machine learning can be the best one from the side of prediction's accuracy, so it offers joint using of different classifiers, which are made on variety parts of learning sample [3]. If use nine methods listed above, so it is possible to get  $2^9 - 9 - 1 = 502$  all kinds of combinations of different models using method of full enumeration.

To decide belonging borrower to one of the classes (creditworthy or not) on the basis of results of parallel application to the original sample of certain methods of the classification, aggregation results is possible on three grounds:

- by average value (the possibility of object belongs to class  $y = 1$  ("creditworthy client") shall be considered as arithmetical average of belonging probabilities of object to class  $y = 1$ , which were found out using all nine methods of classification);
- by median (first of all, expansion is ranging, which contains results base methods of classification in the combination, probability is counted through calculation result of average classifier in the case of their odd number or in the case of even number probability is counted through half-sum of results of average classifier);
- by voting (result of the aggregative classifier in this case is average result of classification's basic methods, which gave fact of the belonging object to  $y = 1$  class with  $\geq 0,1$  probability).

There is a solution algorithm for the assessment of clients' creditworthiness on the basis of aggregative classifiers, it contains next stages:

- 1) Formation and processing of original sample. This stage consists in dividing sample into learning one (for making classification models) and test one (for checking accuracy of the made models), recover missed data [9], discretisation of some characteristics and searching aspects, which influence on the output characteristic  $y$ ;
- 2) Parallel creation of nine classification models on the learning sample;
- 3) Creation aggregative classifier;
- 4) Prediction on the basis of test sample of new clients' creditworthiness using all constructed models;
- 5) Achievement of the prediction result of creditworthiness of every client. It evaluates average probability value of all constructed models on this stage;
- 6) Choice of the best model, which means model with the highest accuracy of the prediction. The accuracy is found out using certain measures [12].

### 3. Information-mathematical system of credit score

Information-mathematical system of credit score was made in the basis of listed above algorithm. It allows predict the class of the object (for example, borrowers' creditworthiness) using learning sample. Software package was devised in the programming support environment Matlab R2014a, which contains all of methods initial data computing and amount of algorithms machine learning, which are needed for solvation the classification problem. Initial data is information about clients, which is personal details and relevant class of the creditworthiness "old" clients; personal details of "new" clients; personal details, credit history and credit transaction terms and conditions of borrowers, who repay a loan.

Program allows making previous preparation of initial data: recover missed information; characteristics' discretisation; coding nonnumeric data; selection statistically worthy characteristics. All of listed above classification methods instantiates in the program, which includes aggregative classifier with the possibility of selecting criteria of aggregation (by average value, by median, by voting).

Method of  $L$ -fold cross-check is used for making classifiers for getting unbiased estimator of quality parameter. The essence of this method is in division original sample to  $L$  non-crossing parts, which are approximately equals to each other by the extent. It is possible to choose  $L$ 's value, it varies from 3 to 10. In turn every part serves as test sample, rest ones aggregates to learning sample. Summative assessment of the classifier's quality is defined by averaging mistakes in all  $L$  test sample. It allows exclude possibility of fudge to the best prediction.

In conclusion constitutes values of quality of created models for three cutoff thresholds (cutoff threshold – value, which if target is higher than the target become the one from class  $y = 1$ ): cutoff threshold 0,5; definitive cutoff threshold; custom cutoff threshold. Definitive classification threshold is the least deviation between mistakes of I-class and II-class.

Quality control of created classification models and aggregative classifiers makes with helping of next characteristics [12]: mistakes of I-class and II-class, ROC curves, area under ROC curve, MSPE and percent of right predictions creditworthy clients and right prediction percent of non-creditworthy clients.

Customer can estimate which method or method combination gives the best result for objects and make prediction for original set of characteristic values using specified criteria. Working process of aggregative classifier is making by program, so optimal method combination is formed automatically using special criteria, after this customer can differ compare results of aggregative classifier and basic classification methods.

### 4. Case study of developed system of credit score

As the first example there are results of program working on realization aggregative classifier for sample of German bank's clients, which includes 900 borrowers, who have 20 characteristics (status of current checking account, credit history, loan purpose, credit length, loan proceeds, average balance on the savings account, work experience in the last place, income in %, family status, guarantors, permanent residence in the last place, data on property, age, available loans, type of housing, number of previous loans in this bank, type of activity, number of dependents, phone availability, citizenship), and one dependent binary variable (borrower is creditworthy and non-creditworthy). This program provides previous data processing, including characteristics' discretisation and coding nonnumeric data, such as citizenship of client, education, family status etc., with numbers. Nine different classification methods and aggregative classifier are analyzed. Aggregation was made by average value. It is possible to make aggregation using all of three characteristics. A 10-fold cross-check was used in this classification.

For target sample is got optimal aggregative classifier with 0,5 cutoff threshold, which contains next methods: neural nets, logit regression, bagging decision trees, created function method, fuzzy inference models. There is results of program in tab.1. The best classification result is got with helping of aggregative classifier, because of mean-root error of aggregative classifier is less than other methods; the highest percent of right prediction of creditworthy clients is in two methods: aggregative classifier and bagging decision trees, but I-class error of aggregative classifier is lower; aggregative classifier gives average value for prediction for non-creditworthy clients, but with minimal II-class error.



Table 1. Results of German bank's borrowers' classification .

Classifier	MSE	Creditworthy ( $y = 1$ )		Non- creditworthy ( $y = 0$ )	
		Right prediction, %	I-class error, %	Right prediction, %	II-class error, %
Neural net	0,1743	84,1	56,8	44,2	15,8
Discriminant analysis	0,1862	84,5	48,0	57,0	16,7
Bayesian classifier	0,2012	76,2	32,8	62,4	28,5
SVM	0,1653	88,5	47,7	63,2	15,1
Decision trees	0,2395	79,1	46,1	57,6	26,8
Logit regression	0,1852	88,7	50,2	50,4	13,3
Bagging decision trees	0,1532	88,1	49,3	51,1	11,9
Created function method	0,4576	35,7	4,8	95,3	70,2
Fuzzy inference models	0,1845	79,3	39,1	68,2	23,5
Aggregative classifier	0,1552	88,5	36,5	61,9	11,0

There are just three levels of quality of classifiers in this table. Also, program allows form diagrams, which show areas under ROC curves (AUC). Fig.1 shows such diagram for target sample. ROC curve [12], also known as curve of errors, shows correlation between deal of right positive classifications from whole number of negative classifications with variation threshold of decision rule. AUC level allows assay diagram of ROC curve. The more AUC level is higher, the more classifier is accurate. Diagram show that aggregative classifier and bagging decision trees gives the most accurate classification result, but AUC of aggregative classifier has higher value.

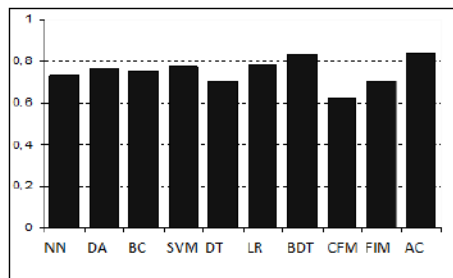


Fig.1. Areas under ROC curves for target sample.

The paper [4] analyses sample of borrowers of German banks but with bigger extent (1000 examinations). Decrease number of examination inconspicuous changes results of classification.

Researching of data about creditworthiness of Australian borrowers was made similar. Names of variables and their values were coded for Privacy Policy. Data includes one of dependent binary variable, which means creditworthiness (takes value 0 in case of non-creditworthy client or 1 in case of creditworthy client) and 14 independent characteristics. There are 690 examinations.

Optimal aggregative classifier for target sample was found 0,5 with cutoff threshold, which contains next methods: neural nets, logit regression, Bayesian classifier and fuzzy inference models. Results of working are in the Table 2. The best result was made with aggregative classifier.

Table 2. Results of classification of Australian bank's borrowers.

Classifier	MSE	Creditworthy ( $y = 1$ )		Non-creditworthy( $y = 0$ )	
		Right prediction, %	I-class error, %	Right prediction, %	II-class error, %
Neural net	0,1258	88,6	56,2	67,3	13,6
Discriminant analysis	0,2511	75,8	42,1	54,2	25,2
Bayesian classifier	0,1253	84,6	58,3	62,8	21,8
SVM	0,1648	87,2	42,2	55,1	20,1
Decision trees	0,3519	69,8	51,0	56,8	18,4
Logit regression	0,2157	78,9	42,9	61,5	12,6
Bagging decision trees	0,1642	76,8	45,2	54,5	20,3
Created function method	0,5862	58,1	31,2	66,8	25,8
Fuzzy inference models	0,1683	87,6	48,6	57,0	16,1
Aggregative classifier	0,1146	89,1	31,8	63,1	12,1

These examples show possibilities of this information-mathematical system of credit score. Method is selected from all of possible methods and it allows predict creditworthiness or non-creditworthiness of clients at the same time, minimizing mean-root error and I-class, II-class errors and maximizing AUC level. Using current method it is possible to find in which class is target using specified set of values. Also, this program allows renovate models if there is new data.

## 5. Conclusion

It considered using nine known methods of machine learning and their combinations for solvation the problem of binary classification of objects. It is not possible to explain effectiveness just one of the methods, because for different samples, even for different parts of one sample, is possible to get variety results. These methods and algorithm of making aggregative classifiers are realized in terms of information-mathematical system of credit score.

This program allows find the best model or optimal aggregative classifier, Classifier was the best one for targets samples. In the case with German borrowers the most accurate prediction was received by using combination of next methods: neural nets, logit regression, bagging decision trees, created function method, fuzzy inference models; classifier includes neural nets, logit regression, Bayesian classifier and fuzzy inference models in case with Australian data. Aggregative classifier helps to get the purpose – increase of prediction accuracy of creditworthiness clients of the bank.

This system of the credit score can be used for any problem of binary classification, for prediction of technical condition of objects [7,8] in particular and for prediction of signal presence or absence.

## References

- [1] Ayzazyan SA, Buchstaber VM, Enyukov IS, Meshalkin LD. Applied Statistics: Classification and Dimension Reduction. Moscow: Finance and Statistics, 1989; 607 p.
- [2] Alekseeva VA. Using of mining techniques in problems of binary classification. Izvestiya of the Samara Scientific Center of the Russian Academy of Sciences 2014; 16(6-2): 354–356.
- [3] Alekseeva VA. Construction of an aggregative binary classifier. Modern problems of design, production and operation of radio engineering systems 2015; 1-2(9): 211–214.
- [4] Alekseeva VA. The use of machine learning methods for binary classification. Automation of Control Processes 2015; 3(41): 58–63.
- [5] Bidyuk PI, Terent'ev AN. Construction and methods of learning Bayesian networks. Informatics and Cybernetics 2004; 2: 140–154.
- [6] Vasiliev NP. Experience in calculating the parameters of logistic regression by the Newton-Raphson method for estimating winter hardiness of plants. Mathematical Biology and Bioinformatics 2011; 6(2): 190–199.
- [7] Klyachkin VN, Karpunina IN, Kuvayskova YuE, Khoreva AS. The Machine learning methods application for technical diagnostics. Scientific Bulletin of the UVAU GA (I) 2016; 8: 158–161.
- [8] Kuvayskova YuE, Barth AD, Fedorova KA. Application of methods of fuzzy logic and machine learning in solving the problem of technical diagnostics. Informatics and Computer Science: a collection of scientific papers of the VIII All-Russian Scientific and Technical Conference of Postgraduates, Students and Young Scientists, 2016;160–166.
- [9] Little RJA, Rubyn DB. Statistical analysis of data with omissions. Moscow: Finance and Statistics, 1990; 336 p.
- [10] Shtovba SD. Identification of nonlinear dependencies using fuzzy logic in the Matlab. Scientific and practical journal Exponenta Pro: mathematics in applications 2003; 2(2): 9–15.
- [11] Shunina YuS, Alekseeva VA, Klyachkin VN. Forecasting the customers' creditworthiness through machine learning methods. Finance and Credit 2015; 27(651): 2–12.
- [12] Shunina YuS, Alekseeva VA, Klyachkin VN. Criteria of quality of qualifiers work. Bulletin of Ulyanovsk State Technical University 2015; 2(70): 67–70.
- [13] Yasnitsky LN. Introduction to Artificial Intelligence. Moscow: Publishing Center "Academy", 2005; 176 p.
- [14] Yakupov AI. Application of decision trees for modeling the creditworthiness of commercial bank clients. Artificial Intelligence 2008; 4: 208–213.

# Joint use of neural network technologies and decision trees for logical patterns exploration in data

V.N. Gridin<sup>1</sup>, V.I. Solodovnikov<sup>1</sup>

<sup>1</sup>*Design information technologies Center RAS, Str. Marshal Biryuzov 7a, 143000, Odintsovo, Moscow Region, Russia*

---

## Abstract

The issues of joint use of neural network technologies with methods of logical deduction and decision making support in data mining tasks are considered. The analysis of searching for logical patterns algorithms, their advantages and disadvantages is carried out. The description of combined algorithms for rules extraction from the trained neural networks and presentation the result in the form of a hierarchical, sequential structure of "if-then" rules is given. The representation of decision trees in the form of the semantic network facts is considered.

*Keywords:* neural network; decision trees; logical conclusion; rules extraction; data mining

---

## 1. Introduction

Data is a valuable resource, which contains a great potential opportunities for the extraction of useful analytical information. Therefore, the tasks of revealing hidden regularities, developing decision making strategies, forecasting, which requires more detailed consideration of the logical patterns exploration in classification problems, are becoming increasingly important. The peculiarity of algorithms and methods applicable for solving the data mining problems is the absence of a priori assumptions about the sampling structure and the distributions type of the analyzed indicators values. One of the closest correspondences to this condition could be the usage of an approach based on neural network technologies. This is due to the ability of neural networks for nonlinear processes modeling, working with the extremely complex dependencies, adaptability to the functioning conditions, and most importantly, the ability to extract and generalize essential features from incoming information. Thus, the network constructs rules, but these rules are contained in weighting coefficients, activation functions, and neuronal connections, but usually their structure is too complex to perceive and determine the effect of a particular characteristic on the output value. The neural network, in fact, acts as a "black box", the input of which is supplied with the initial data and the certain output result is obtained, however, it is not provided any rationale why this decision was made. To solve this problem, it is proposed the joint use of the neural network technologies with logical deduction methods, in particular decision trees, as a means of decision-making support, logical patterns exploration and the result presentation in the form of a hierarchical structure of classifying rules. And the use of semantic networks provides additional opportunities in construction of the deduction mechanisms and presentation the decision-making process.

## 2. Searching for logical patterns in the data

We will understand by logical regularity an easily interpreted rule that allocates a lot of objects of one class from the training sample and practically does not allocate objects of other classes. Logical patterns are elementary "building blocks" for a wide class of classification algorithms. Rules, that express regularities, are formulated in the language of first-order logic predicates and have the following form:

IF (condition\_1) AND (condition\_2) AND ... AND (condition\_N) THEN (conclusion),

where condition i could be  $x_i = c_1$ ,  $x_i < c_2$ ,  $x_i > c_3$ ,  $c_4 < x_i < c_5$  etc.,  $x_i$  - variable,  $c_1, c_2, c_3, c_4, c_5$  - some constants.

For nominative data, the following predicates are used: « $\Rightarrow$ » and « $\langle \rangle$ ».

### 2.1. The limited search algorithms

The limited search algorithms are used for logical regularities search in data, for solving classification and forecasting problems [1]. The main idea of this method is to analyze the frequency of occurrence of various combinations of simple logical events. At the initial stages, short associative chains are searched for, which are complicated in the process of the system's functioning, by adding new elements to them. Based on the analysis, the system makes a conclusion about the usefulness of this or that combination and, thus, establishes the logical patterns in the data. Its main disadvantage is the fact that this algorithm is capable in an acceptable time to find a solution for only a small dimension data.

### 2.2. Decision trees

Decision trees relate to the methods of logical regularities searching in data, and are the main approach applicable in decision making theory. They represent the hierarchical structure of "if-then" classifying rules, which have the form of a tree. Their main advantage is the simplicity and clarity of the decision-making process description. The disadvantage of their use in the problem of logical patterns search is the fact, that they are not able to find the most complete and accurate rules in the data and only

implement the simplest principle of sequential viewing of attributes and form fragments of regularities. Also, for large volumes of multidimensional data, these algorithms can produce a very complex tree structure that has many nodes and branches. Such trees could be very difficult for analyzing and understanding. Accordingly, the rules and patterns discovered by such a tree would be difficult for comprehension. In addition, a branchy tree with many nodes divides the training set into a large number of subsets consisting of a small number of objects. While it is much more preferable to have a tree with a small number of nodes, for each of which correspond a large number of objects from the training sample. To solve this problem, branch cutoff algorithms are often used [2], but they can not always lead to the desired result. However, methods of searching regularities with the help of decision trees allow us to find such connections that are concluded not only in certain features, but also in a combination of features, which in many cases gives these methods a significant advantage over classical methods of multivariate analysis.

Figure 1 shows an example of such a decision tree, and the corresponding logical deduction, where  $\theta_1, \theta_2, \theta_3$  - predicates,  $x, y, z$  - variables,  $\alpha, \beta, \chi$  - constants.

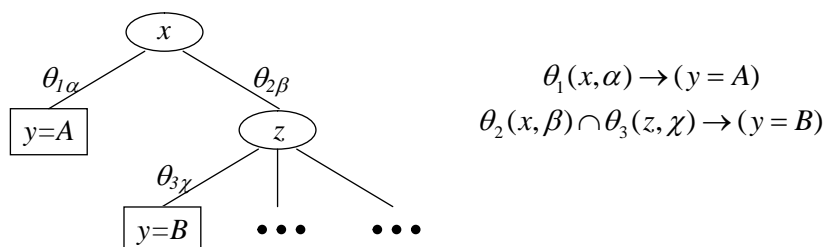


Fig. 1. An example of a decision tree.

Rules that express regularities are formulated in the form of expressions: «IF A THEN B» or in the case of a set of conditions: «IF (condition 1)  $\wedge$  (condition 2)  $\wedge$  ...  $\wedge$  (condition N) THEN (the output node value)».

The decision trees construction is usually carried out by following ways:

- on the expert assessments basis;
- using sample processing algorithms (CLS, ID3 (Interactive Dichotomizer), C4.5, CART (classification and regression trees) etc. );
- using genetic algorithms and evolutionary programming.

Each of these approaches has its advantages and disadvantages and can be used to solve its specific tasks.

### 2.3. Genetic algorithms

The most difficult problem in search for logical regularities in data sets is to find the elementary events, representing the terms of the conditional part "IF". At present, genetic algorithms (GA) are increasingly used to solve this problem, which include algorithms: Bucket-Brigade, REGAL, G-NET, HIDER, SIAO1 and some others. However, they are not without a number of drawbacks: a fixed set of rules and their length, as well as accuracy and completeness in most of them are not taken into account.

### 2.4. Neural network methods

The usage of the approach based on neural network data processing technologies is caused by the ability of neural networks to model non-linear processes, act with extremely complex dependencies, adaptate to operating conditions, work with noisy data and with the lack of a priori information. And most importantly, they are able to learn from experience, generalize previous precedents into new cases and extract significant features from incoming information. Thus, the network constructs rules, but these rules are contained in weighting coefficients, activation functions, and neuronal connections, but usually their structure is too complex to perceive. Moreover, these parameters can represent non-linear, non-monotonic relationship between the input and target values in a multilayer network. Thus, as a rule, it is not possible to separate the effect of a certain attribute to the target value, cause of this effect can be mediated by the values of other parameters. The neural network, in fact, acts as a "black box", the input of which is supplied with the initial data and the certain output result is obtained, however, it is not provided any rationale, why this decision was made.

## 3. Getting logical patterns from a trained neural network

Let the problem consists in the classification of a certain set of data with the help of a perceptron and the subsequent analysis of the obtained network in order to find the classifying rules that characterizes each of the classes.

First, let's consider this problem with the example of a single-layer perceptron, which consists of five Boolean inputs and one output neuron. This network can be accurately interpreted by a finite number of "if-then" rules, since a finite number of possible input vectors are defined for it.

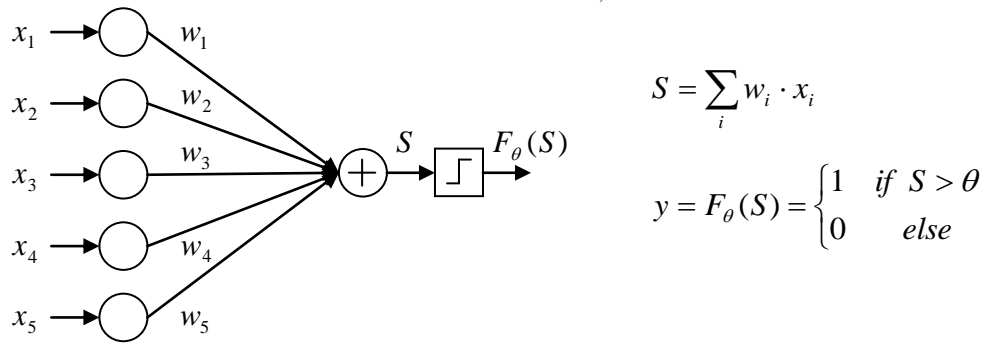


Fig. 2. Single-layer perceptron with five Boolean inputs and one output.

Let the weights take on the following values:  $w_1 = 6$ ,  $w_2 = 4$ ,  $w_3 = 4$ ,  $w_4 = 0$ ,  $w_5 = -4$ , and the bias  $\theta = 9$ . In this case, the following set of rules can be extracted from the network:

$$\begin{aligned} x_1 \wedge x_2 \wedge x_3 &\rightarrow y \\ x_1 \wedge x_2 \wedge \neg x_5 &\rightarrow y \\ x_1 \wedge x_3 \wedge \neg x_5 &\rightarrow y \end{aligned}$$

Thus, the decision-making procedure is to predict the value  $y = \text{true}$ , if the activation of the output neuron is 1, and  $y = \text{false}$ , if the activation is 0.

Generally speaking, it is possible to distinguish two approaches for extracting rules from the multilayer neural networks [3]. The first approach is to extract a set of global rules that characterize the output classes directly through the values of the input parameters. An alternative is to extract local rules by separating a multi-layer network into a collection of single-layer networks. Each extracted local rule characterizes a separate hidden or output neuron, taking into account elements that have weighted connections with it. Then all got rules are combined into a set that determines the behavior of the entire network as a whole. The local approach is illustrated at Figure 3.

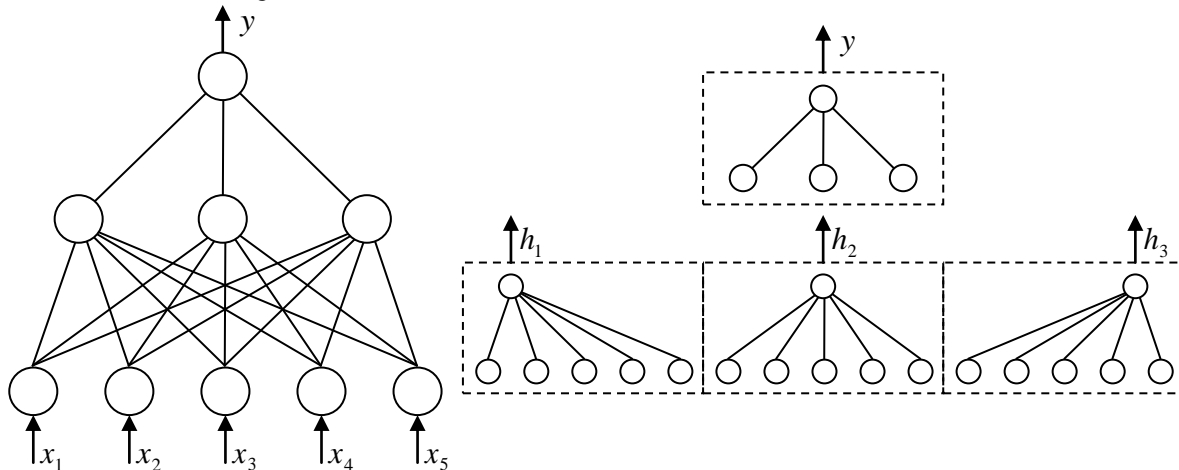


Fig. 3. A local approach for extracting rules. Multilayer neural network is divided into a set of single-layered. Rules for description of each component are extracted, which are combined into a set that characterizes the multi-layer network.

Let's consider the problem of extracting rules in a more general form.

Let  $X$  denote a set of  $n$  properties  $X_1, X_2, \dots, X_n$ , and  $\{x_i\}$  is the set of possible values that a property  $X_i$  can take. And  $C$  denotes the set of classes  $c_1, c_2, \dots, c_m$ . The associated pairs of input and output vector values are known for the training sample  $(x_1, \dots, x_n, c_j)$ , where  $c_j \in C$ .

### 3.1. Local approach for rules extraction

NeuroRule is one of the algorithms for extracting rules from neural networks, which were trained to solve the classification problem [4]. This algorithm includes three main steps:

Step 1. Neural network training.

At the first stage, a two-layer perceptron is trained until sufficient classification accuracy would be obtained. At the initial time, a large number of the hidden layer neurons are selected. Unnecessary neurons and connections would be discarded after training.

Step 2. Thinning of a neural network.

The trained neural network contains all possible connections between input neurons and hidden layer neurons, as well as between hidden and output neurons. Usually the total number of these links is so large that it is impossible to extract observable for user classifying rules from their values analysis. Thinning consists of removing unnecessary connections and neurons, which absence does not increase the network classification error. The resulting network usually contains much less neurons and connections between them, and the operation of such a network is able to be investigated.

### Step 3. Rules extraction.

At this stage, the rules that take form of  $\langle \text{IF } (x_1 \Theta q_1) \text{ AND } (x_2 \Theta q_2) \text{ AND } \dots \text{ AND } (x_n \Theta q_n) \text{ THEN } c_j \rangle$  are extracted from the thinned neural network. Here  $q_1, \dots, q_n$  are constants and  $\Theta$  is the relational operator ( $=, \geq, \leq, >, <$ ). First the preparation for rules extraction is taking place, which includes coding of all continuous quantities for input and interior network values. Also the coding process is performed for features of the classified objects if they have continuous values. To represent them, it is possible to use binary neurons and a coding principle such as a thermometer. The resulting values of the hidden layer neurons are clustered and replaced with values that determine the centers of this clusters. It is important to select a small number of such clusters. The objects classification accuracy by the network is checked, after such discretization of the hidden neurons functionality. If it remains acceptable, than the preparation for rules extraction comes to the end. Further, the rules extraction is taking place, during which the movement through the network occurs from the classifying output neurons to the network inputs. It is assumed that these rules are fairly obvious while verified and are easily could be applied to the large databases.

However, this algorithm establishes rather strict limitations on the architecture of the neural network, the number of elements, connections and the type of activation functions. So for the hidden neurons the hyperbolic tangent is used and their states change in  $[-1,1]$  interval, and for the output neurons the Fermi function with the state interval  $[0,1]$  is applied.

### 3.2. Global rules extraction

The lack of universality and scalability could be mentioned as the main drawbacks of most algorithms of rules extraction. In this regard, TREPAN algorithm [5] gets the most interest. It lacks these shortcomings and does not impose any requirements to the network architecture, input and output values, learning algorithm, etc. This approach builds a decision tree on the base of knowledge embedded in the trained neural network, and it is enough that the network is a kind of "black box", "expert" or "oracle", to whom it is possible to ask questions and get answers. Moreover, this algorithm is sufficiently universal and can be applied to a wide range of other trained classifiers. It also scales well and has no sensitive to the input attributes dimension and the size of the network.

This algorithm builds the decision tree, that approximates the functionality of the trained neural network, and consists of two following stages.

#### Preliminary stage:

1. Construct and train a neural network that will later act as an "Expert" or "Oracle".
2. Initialize the root of the tree  $R$  as a leaf.
3. Use the entire training set of examples  $S$  to construct a distribution model  $M_R$  of the input vectors that reach the node  $R$ . Compute value  $q = \max(0, \text{minSamples} - |S|)$ , where  $\text{minSamples}$  is the minimum number of training examples used in each node of the tree,  $S$  is the current training sample ( $|S|$  is the training sample volume). Thus,  $q$  is the number of additional examples that need to be generated.
4.  $q$  new learning examples are generated randomly on the base of the attributes distribution evaluation from  $S$ .  $query_R$  is the set of  $q$  examples generated by the model  $M_R$ .
5. Use neural network as an "Oracle" to classify both new  $query_R$  and old examples from the set  $S$  to a particular class. For each vector of attributes  $x \in (S \cup query_R)$ , put a class label  $x = Oracle(x)$ .
6. Initialize the *Queue*, by placing the set  $\langle R, S, query_R, \{empty\_constr\} \rangle$ .

#### Main stage:

7. Take the next set  $\langle N, S_N, query_N, constr_N \rangle$  from beginning of the *Queue*, where  $N$  is the node of the tree,  $S_N$  is the training sample in the node  $N$ ,  $constr_N$  is a set of restrictions on the certain attributes of the training examples for reaching the node  $N$ .
8. Use  $F, S_N, query_N$  for construction branching  $T$  in a node  $N$ .

Here  $F$  is a function for estimating the node  $N$ . It has the following form  $F(N) = R(N) \cdot (1 - f(N))$ , where  $R(N)$  is the probability of reaching node  $N$  by an example, and  $f(N)$  is the correctness evaluation of these examples processing by a tree. Thus, the best node is chosen, which branching has the greatest impact to the classification accuracy of the generated tree. The separation of the examples, which reach this internal node of the tree, is carried out depending on the  $m-of-n$  test [5,6]. Such a test is considered to be passed when it is satisfied, at least  $m$  from  $n$  conditions. On the other hand, it is possible to split the set  $S$  as in the usual algorithm for decision tree construction.

9. Create next-generation nodes for each branch  $t$  of branching  $T$ :
  - a. Create  $C$  as a new child node in a relation to  $N$ .

- b. Add a restriction from the branch  $t$  to  $constr_C = constr_N \cup \{T = t\}$ .
  - c. Generate set  $S_C$ , which contains examples from the set  $S_N$  that satisfy the condition on the branch  $t$ .
  - d. Construct a model  $M_C$  for examples distribution that reach the node  $C$ .  
Calculate the number of examples to generate  $q = \max(0, \minSamples - |S_C|)$ .
  - e.  $q$  new learning examples are randomly generated on the base of the characteristics distribution evaluation from  $S_C$  and constraint values  $constr_C$ .  $query_C$  is a set of  $q$  examples, which were generated by the model  $M_C$  and constraint  $constr_C$ .
  - f. Use neural network as an "Oracle" to classify new examples  $x \in query_C$  and expose the class label  $x = Oracle(x)$ .
  - g. Initially, it is assumed that the node  $C$  is a leaf. Use  $S_C$  and  $query_C$  to determine the class label for  $C$ .
  - h. Check the necessity of the further branching of the node  $C$ . Put the set  $\langle C, S_C, query_C, constr_C \rangle$  into the *Queue* if the local stop criterion is not satisfied. A local criterion in this case is the probability that in a given node there are instances of one class.
10. If the *Queue* is not empty and the global stop criterion is not fulfilled, then go to step 7, otherwise return the tree with the root  $R$ .

The maximum tree size and the overall classification quality evaluation of examples by a tree are used as the global criterion for completing the algorithm.

The generalization ability of artificial neural networks, which allows to obtain more simple decision trees is the main advantage of this approach. In addition, the applying of such an "Oracle" allows to compensate the lack of data, which is usually could be observed at lower levels during the decision trees construction by the sample processing algorithms. Thus, it is possible to extract structured knowledge not only from extremely simplified neural networks, but also from arbitrary classifiers that makes possible the appliance of this algorithm in a wide range of practical problems.

#### 4. Rules representation in a form of a semantic network

A simple semantic network, sometimes called a computational semantic network, is actually a bipartite graph.

Lets there are a finite set  $A = \{A_1, \dots, A_r\}$ , which is called attributes, and the finite set  $R = \{R_1, \dots, R_n\}$  of relations. The scheme or intensional of the ratio  $R_i$  ( $i = 1, \dots, n$ ) is the set of pairs:

$$INT(R_i) = \{ \dots, [A_j, DOM(A_j)], \dots \},$$

where  $R_i$  is the name of the relation,  $DOM(A_j)$  is the domain of  $A_j$  ( $j = 1, \dots, r$ ), i.e. the set of attribute  $A_j$  values of the relation  $R_i$ . The union of all domains is called the base set of the model or the set of objects, on which the relations  $R$  are specified.

An extensional of  $R_i$  relation is the set:

$$EXT(R_i) = \{F_1, \dots, F_p\},$$

where  $F_k$  ( $k = 1, \dots, p$ ) is the fact of the relationship  $R_i$ . The fact is set by an aggregate of attribute-value pairs, called attribute pairs. Fact is a concretization of a certain relationship between the specified objects. In a graphical interpretation, fact is a subgraph of a semantic network that has a star-shaped structure. The root of a subgraph is a vertex of a predicate type, labeled with a unique label that includes the name of the corresponding relationship. From the vertex of the fact connections are coming out, which are marked with the names of attributes of this fact. They are leading to the vertices of the base set and are the values of these attributes.

It is worth noting that the semantic network can be represented as a storage of facts that were derived from the decision trees processing, i.e. the decision tree is transformed into a semantic network. In this case, each fact is presented in the form of a ready-made deduction output. This provides additional opportunities for analysis. For example, even the usual means of databases can find the facts that relate to different relationships but have the same attributes and values that characterize them. To store semantic networks in the database, or rather their extensional, it is possible to use a table of the form:

Table 1. Table of extensional.

Field name	Data type	Field Properties
CodeValue	Numeric	Key field
FactMark	Numeric	The fact mark
FactAttributes	Text	Attribute of fact
AttributeValue	Text	Attribute Value

Thus, the decision tree view will be obtained in the form of a fact table, where a record with the fact attributes values exists for each value of the output deduction variable. Since there is a mapping of one representation to another, so from the formal point of view these representations are identical.

The conclusions could be drawn in a semantic network, which are far from obvious for a decision tree. Let's take this simple example. The decision tree determines the conclusion about the establishment of the parental relations in the first generation. It is necessary to define the parent relationship in the second generation. The solution of this problem for the semantic network is

obvious. It is necessary to highlight the facts of parental relations, then to delete objects from the intensional that do not match the parents and repeat the process of highlighting the facts of the parent relationship.

This means that it is possible in the semantic network introduce means of constructing functional dependencies similar to how it is done in functional programming languages, for example, such as LISP [7]. You can consider the fact as a list of atoms, each of which is assigned a value either directly or in the process of output. If operations for such lists manipulation are entered then it is possible to create and modify decision trees with formal methods.

Thus, it seems advisable to combine in a single system a representation in the form of decision trees and a class of semantic networks, which makes it possible to visually display the decision-making process and gives additional possibilities in constructing the deduction mechanisms.

## 5. Conclusion

In this paper, the problems of logical regularities search in the classification tasks were considered. A joint use of neural network technologies with logical deduction methods, in particular decision trees, as a means of logical patterns exploration and the result presentation in a hierarchical structure form of classifying rules, is proposed. Two main approaches are identified. The first is to extract local rules, where the multilayer network is divided into a set of single-layered. Each local rule characterizes a separate hidden or output neuron, taking into account elements that have weighted connections with it. Then the rules are combined into a set that determines the behavior of the entire network as a whole. However, this approach often establishes fairly strict limitations on the network architecture, the number of elements, links and the type of activation functions, which negatively affects the universality and scalability.

An alternative is to extract a set of global rules that characterize classes at the output directly through the values of the input parameters. In the framework of this approach a modified algorithm for decision trees construction on the base of the trained neural networks is developed. It does not impose any requirements on architecture, learning algorithm, input and output values and other network parameters. The construction of the tree is base on knowledge that embedded into the trained neural network, and it is enough that the network is a kind of "Black Box" or "Expert", for which it is possible to ask questions and get answers. The generalization ability of artificial neural networks, which allows to obtain more simple decision trees and, if it is necessary, to compensate the lack of initial data are the main advantages of this approach. Moreover, this algorithm is sufficiently universal, well scalable and not sensitive to the input attributes dimension and the size of the network. This circumstance acquires special significance in the light of the rapid development of deep learning technology. Thus, it is possible to extract structured knowledge not only from extremely simplified neural networks, but also from arbitrary classifiers that makes possible the appliance of this algorithm in a wide range of practical problems.

In addition, an algorithm for converting already formed decision trees into semantic networks in the form of a bipartite graph has been developed. This provides a solution tree view in the form of a fact table and allows a quick search by known attributes.

The automation tools introduction into the data mining systems could shorter the time, improve the quality and effectiveness of the decisions making.

## Acknowledgements

Work is carried out with the financial support of the RFBR, the project 15-07-01117a.

## References

- [1] Dyuk V, Samojlenko A. Data Mining. SPb: Piter, 2001; 368 p.
- [2] BaseGroup Labs company WebSite, Trees of decisions - general principles of work. URL: <https://basegroup.ru/community/articles/description> (10.01.2017).
- [3] Gridin VN, Solodovnikov VI, Evdokimov IA, Filippkov SV. Building decision trees and extracting rules from trained neural networks. *Iskusstvennyj intellekt i prinyatie reshenij* 2013; 4: 26–33.
- [4] Ezhov AA, Shumskij SA. Neurocomputing and its application in economics and business. M.: MIFI, 1998; 224 p.
- [5] Craven MW, Shavlik JW. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA 1996; 8: 24–30.
- [6] Murphy PM, Pazzani MJ. ID2-of-3: Constructive induction of M-of-N concepts for discriminators in decision trees. *Proceedings of the Eighth International Machine Learning Workshop*, Evanston, IL 1991; 183–187.
- [7] Hyuvenen EH, Seppyanen I. *Mir LISPA. Methods and systems of programming*. M.: Mir, 1990; 320 p.



# Control of component alterations according with the target efficiency of data processing and control system

V.E. Gvozdev<sup>1</sup>, M.B. Guzairov<sup>1</sup>, D.V. Blinova<sup>1</sup>, A.S. Davlieva<sup>1</sup>

<sup>1</sup>Ufa State Aviation Technical University, Karl Marx street, 12, Ufa, Russia

---

## Abstract

In this article we describe the solution to make an estimate of allowed alterations in model components parameters that make up a data processing and control system. They reflect properties of physical and information components of the aforementioned system. This solution is derived according to the limitations of possible changes in integral property that describes the system behavior in different modes of operation. The proposed solution makes it possible to solve not only the direct problem - making a conclusion whether the vulnerability of the data processing system is negligible with respect to alterations in parameters of this systems components but the inverse as well, finding tolerable levels in alterations in systems components parameters from an acceptable level of uncertainty of targeted efficiency. The proposed solution follows known demands: inner properties of the system must be so that customer demands are fulfilled.

*Keywords:* data processing and control system; system vulnerability; target system efficiency; parameter alterations; uncertainty in systems components

---

## 1. Introduction

Nowadays information environment is a key factor in the life of our society, which leads to critical importance of managing the functional vulnerabilities of data processing and control systems. (DPCS) [1]. Concept of "security" is tightly coupled with the concept of "reliability" [2]. Reliability estimate is done by comparison of actual functional properties of the system and base functional properties defined in the development specification. These base functional properties in turn are the reflection of wishes and demands of the users onto the functional properties of DPCS. One of the properties of "general reliability" is survivability, the property of the system to continue functioning under influence of external and internal malicious factors, such factors include alterations in parameters of components of the system [3,4]. The opposite to survivability is vulnerability - a parameter that describes the possibility of the system being damaged by external and internal causes of different nature. These can be functional, economical, management, physical and so on. One of the distinct properties of modern DPCS is that they have infinitely many internal states. This is caused by infinite combinations of input data as well as other external factors that can cause errors of different nature. The cause of these defects are flaws of different types made on certain stages of software lifecycle [5-8]. The latter promotes development of vulnerability research techniques of DPCS by analysis of external behavior of the system [9-11]. This work presents a technique to study the influence of alterations in systems parameters, running in different modes; on it's vulnerability in cases when the external behavior of the system is defined by the target efficiency property.

## 2. Target efficiency as indirect property of functional vulnerability

Functional vulnerability is the factor that negatively effects consumer qualities and consumption of resources needed to maintain DPCS.

Target efficiency shows the degree of match between actual functions of the system and it's target functions [2]. Due to this fact target efficiency can be viewed as integral quality property of DPCS, used in different modes. This quality property shows the correlation between the expected by users behavior and external behavior of the given system. Decline in the target efficiency is an indirect property of deviation of the system from state  $S_k$  and base state  $S_0$ , in other words it is an indirect vulnerability criteria. Random nature of the target allows the usage of statistical methods to research the target efficiency [3].

In [3] robustness of the system is defined as follows: "...conditional probability of end system state  $S_k$  will not deviate from base state  $S_0$  more than a given value  $\varepsilon_0$  when event  $\omega$  happens". The same reference contains formal relations between functional vulnerability and robustness.

$$v_f = 1 - \text{Rob} , \quad (1)$$

where  $\text{Rob} = 1 - P[\|S_k - S_0\| < \varepsilon_0 | \omega]$ . Here  $\omega$  is an attribute of unwanted event.

Based on the given definition of robustness relations between robustness and functional vulnerability it is possible to state that there exists a direct dependency between robustness and target efficiency. Existence of such dependency lets us postulate the following: functional vulnerability of the system  $v_f$  is contained in tolerable limits  $\varepsilon_0^{(v_f)}$ , if the probability of deviation of target efficiency property  $\mathcal{A}_\Sigma$  does not rise over a base value  $\varepsilon_0^{(v_f)}$ :

$$\exists P[\mathcal{A}_\Sigma - \mathcal{A}_\Sigma^{(0)} < \varepsilon_0^{(\mathcal{A}_\Sigma)}] \Rightarrow v_f < \varepsilon_0^{(v_f)} . \quad (2)$$

Alteration of systems components parameters is the inherent property of any DPCS. Statistical uncertainty as a probability of state parameters being in tolerable intervals comprise the metric property of such alteration. On the other hand change in the parameters is the cause of statistical uncertainty of target efficiency property:

$$v_f = f_1(H_D) , \tag{3}$$

$$H_{\mathcal{D}} = f_2(H_D) , \tag{4}$$

Here  $H_{\mathcal{D}}$  is the metric uncertainty characteristic of the target efficiency;

$f_1(\bullet)$  – a direct functional relationship, making a relationship  $v_f$  between the metric characteristics of the uncertainty  $H_D$  and the state parameters vector components  $D$  .

$f_2(\bullet)$  – a direct functional relationship that makes a relationship  $H_{\mathcal{D}}$  between the metric characteristics of uncertainty  $H_D$  . The character  $f_2(\bullet)$  is defined by the structure of the system.

From (3) and (4) it can be concluded that, from the limitations on the statistical uncertainty of the system's target efficiency, there is a limitation on the value of the statistical uncertainty of the system state parameters. In other words, if the limitations on the variability (the uncertainty characteristic) of the average target efficiency's index are kept, then it can be argued that the vulnerability of the system is within the permissible limits.

### 3. Task statement and assumptions

The initial data of the problem are:

- (A) Description of the system states set  $S_i(i = \overline{1;N})$  , with each state matching the characteristics of the target efficiency  $\mathcal{D}_i$  ;
- (B) The same characteristics of the statistical uncertainty of the target efficiency for each states  $H_i(\mathcal{D})$  ;
- (C) A system model that characterizes the relationships  $\lambda_{ij}$  between the states  $i$  and  $j$  ( $i, j = \overline{1;N}, i \neq j$ ) ;
- (D) A rule that allows us to estimate the average proportion of the time  $p_i(i = \overline{1;N})$  the system is in the  $i$ -th state;
- (E) Uncertainty characteristics of relations  $H_{ij}(\lambda)$  ;
- (F) The rule for estimating the average target efficiency  $\mathcal{D}_{\Sigma}$  as a function of  $\mathcal{D}_i$  and  $p_i$  :

$$\mathcal{D}_{\Sigma} = \varphi(\mathcal{D}_i, p_i) ; \tag{5}$$

- (G) The rule for estimating the statistical uncertainty characteristics of the average target efficiency  $H_{\mathcal{D}}$  based on  $H_i(\mathcal{D}), H_{ij}(\lambda)$  :

$$H_{\mathcal{D}} = f_2(H_i(\mathcal{D}), H_{ij}(\lambda)) \tag{6}$$

Note that in (5) components  $H_i(\mathcal{D}), H_{ij}(\lambda)$  are the components of the state parameters vector (see (H)).

- (H) Limitations on the variability  $\Delta H_{\mathcal{D}}$  of the uncertainty characteristic of the average target efficiency  $\varepsilon_0^{(\mathcal{D}_{\Sigma})}$  .

It is required: to estimate the limits on the possible values of uncertainty characteristics  $\Delta H_i(\mathcal{D}), \Delta H_{ij}(\lambda)$  based on the limitation on the values  $\Delta H_{\mathcal{D}}$  of the uncertainty characteristic of the average target efficiency.

Assumptions:

- (A) The apparatus of Markov processes is used as a basis for modeling the state of DPCS [12];
- (B) Intervals  $\mathcal{D}_i \in [\mathcal{D}_i^{(l)}, \mathcal{D}_i^{(u)}]$  ;  $\lambda_{ij} \in [\lambda_{ij}^{(l)}, \lambda_{ij}^{(u)}]$  are used as characteristics of the statistical uncertainty of the target efficiency  $H_i(\mathcal{D})$  and relationships  $H_{ij}(\lambda)$  accordingly. The index "l" corresponds to the lower limit of the interval; index "u" - the upper;
- (C) Linear convolution is used as an estimate of the target efficiency

$$\mathcal{D}_{\Sigma} = \sum_{i=1}^N p_i \cdot \mathcal{D}_i , \tag{7}$$

which is the average value of the target efficiency;

- (D) Probability is a uncertainty characteristic of the target efficiency

$$H_{\mathcal{D}} = P[a^{(l)} < \mathcal{D}_{\Sigma} < a^{(u)}] , \tag{8}$$

where,  $a^{(l)}, a^{(u)}$  are determined by (Fig. 1):

$$P[0 < \mathcal{D}_{\Sigma} \leq a^{(u)}] = P[a^{(l)} < \mathcal{D}_{\Sigma} < \infty] = \varepsilon_0^{(\mathcal{D}_{\Sigma})} / 2 . \tag{9}$$

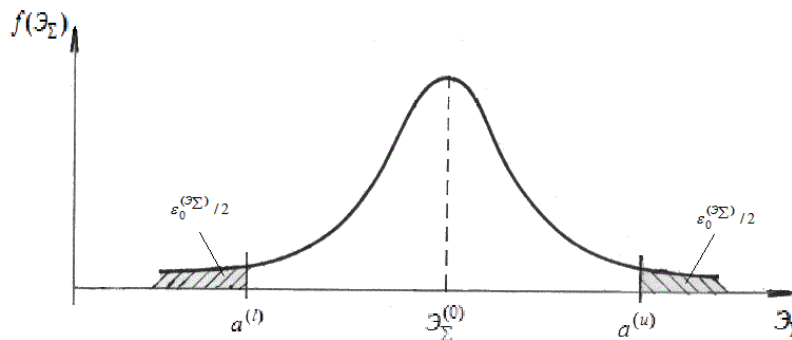


Fig. 1. The graphical illustration to the problem of estimating confines on the of uncertainty characteristics values of the target efficiency.

In Fig. 1  $\mathfrak{A}_\Sigma^{(0)}$  corresponds to the basic values  $\{\mathfrak{A}_i\}, \{\lambda_{ij}\}$ .

#### 4. Solution for the task<sup>1</sup>

The basis solution for the task is the construction of the dependence (6), which connects the statistical characteristics of the uncertainty of the average target efficiency of the system with the statistical characteristics of the uncertainty of the components of the system model. The basis for constructing the dependence (6) is a statistical experiment, the scheme of which is shown in Fig.2.

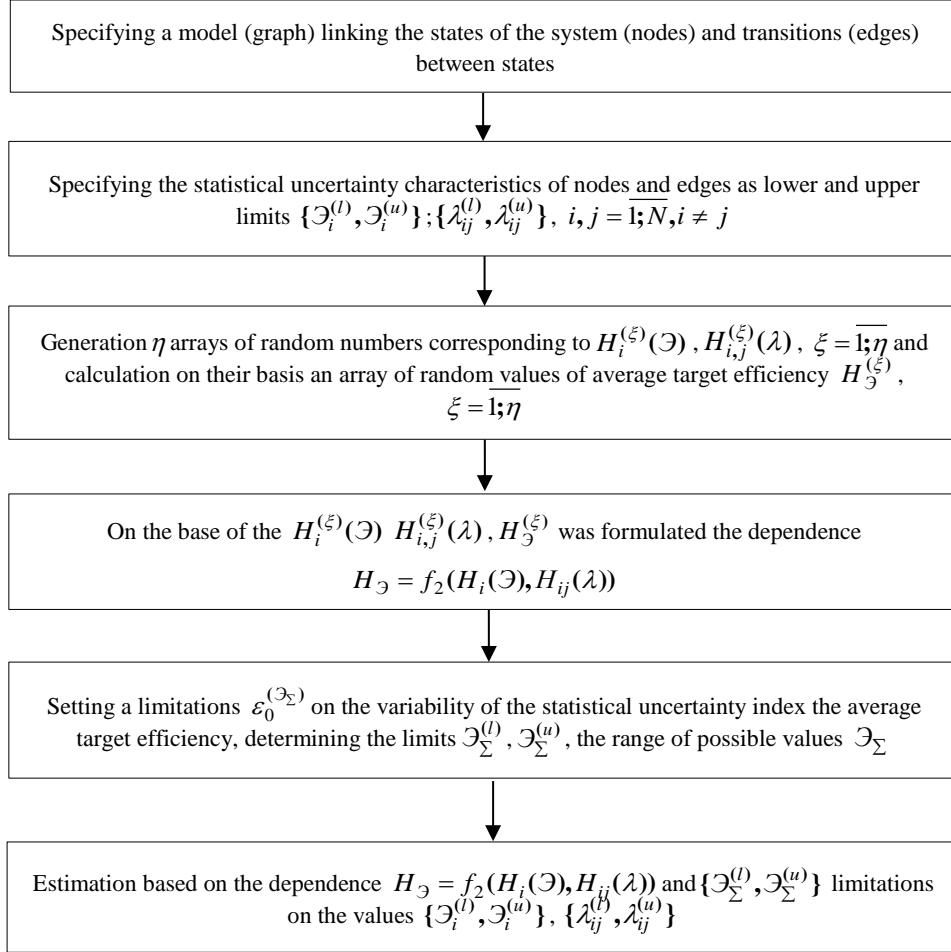


Fig. 2. Scheme of statistical experiment.

In the experiment as the uncertainty characteristic of the average target efficiency  $H_{\mathfrak{A}}$  was the distribution density  $f(\mathfrak{A}_{\Sigma})$  of the average target efficiency  $\mathfrak{A}_{\Sigma}$ . As the uncertainty characteristics of system components  $H_i(\mathfrak{A}), H_{ij}(\lambda)$  were interval limits of possible values  $\{\mathfrak{A}_i^{(l)}, \mathfrak{A}_i^{(u)}\}, \{\lambda_{ij}^{(l)}, \lambda_{ij}^{(u)}\}$ . These limits were determined by the rules:

$$\mathfrak{A}_i^{(l),(u)} = \mathfrak{A}_i^{(\emptyset)}(1 \pm \alpha_{\mathfrak{A}}); \lambda_{ij}^{(l),(u)} = \lambda_{ij}^{(\emptyset)}(1 \pm \alpha_{\lambda}),$$

where the sign "-" corresponds to the lower limit of the interval of possible values of the graph component characteristic; the "+" sign corresponds to the upper limit;

the index " $\emptyset$ " corresponds to the basic values of the characteristic.

Note that the interval uncertainty characteristics estimates in accordance with the principle of maximization of entropy [13, 14] can be associated a law of random variable distribution.

Fig. 3 shows the model (states graph) of the system [15, 16]. Table 1 shows the base values of average target efficiencies  $\{\mathfrak{A}_i^{(b)}, i = \overline{1;4}\}$ . The base values of the transitions intensities  $\{\lambda_{ij}^{(b)}, i, j = \overline{1;4}, i \neq j\}$  were taken to be the same and equaled ten. During the study,  $\alpha_{\mathfrak{A}}, \alpha_{\lambda}$  took a different value ( $\alpha_{\mathfrak{A}} \in [0;1], \alpha_{\lambda} \in [0;1]$ ). Fig. 4 shows estimates of the distribution densities  $f(\mathfrak{A}_{\Sigma})$ , corresponding to different  $\alpha_{\mathfrak{A}}, \alpha_{\lambda}$  for different uncertainty distribution by the graph components:

- (A) The statistical uncertainty corresponds to the graph nodes, the nominal values of the transitions intensities correspond to the edges;

<sup>1</sup> In the development of the program for conducting a statistical experiment and processing the results of the experiment, the undergraduate student of the Department of Technical Cybernetics of the Ufa State Aviation Technical University Teslenko V.V. actively participated.

- (B) The statistical uncertainty corresponds to the graph edges, the nominal values of the average target efficiency correspond to the nodes;
- (C) The statistical uncertainty corresponds to both the edges and the nodes of the graph.

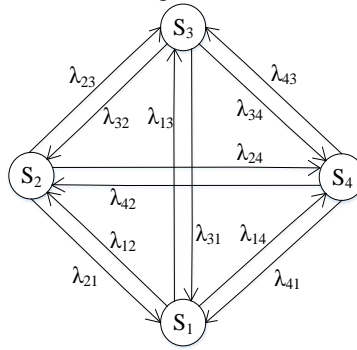


Fig. 3. States graph of the system.

Table 1. Characteristics of graph nodes.

System state	$S_1$	$S_2$	$S_3$	$S_4$
Base value of average target efficiency	10	20	30	40

At estimating  $f(\Theta_\Sigma)$  value  $\eta$  was taken  $10^4$ . To determine the number of grouping intervals in the histograms construction, the Sturges rule was used:

$$n = \text{int}(1 + 3.31\lg\eta)$$

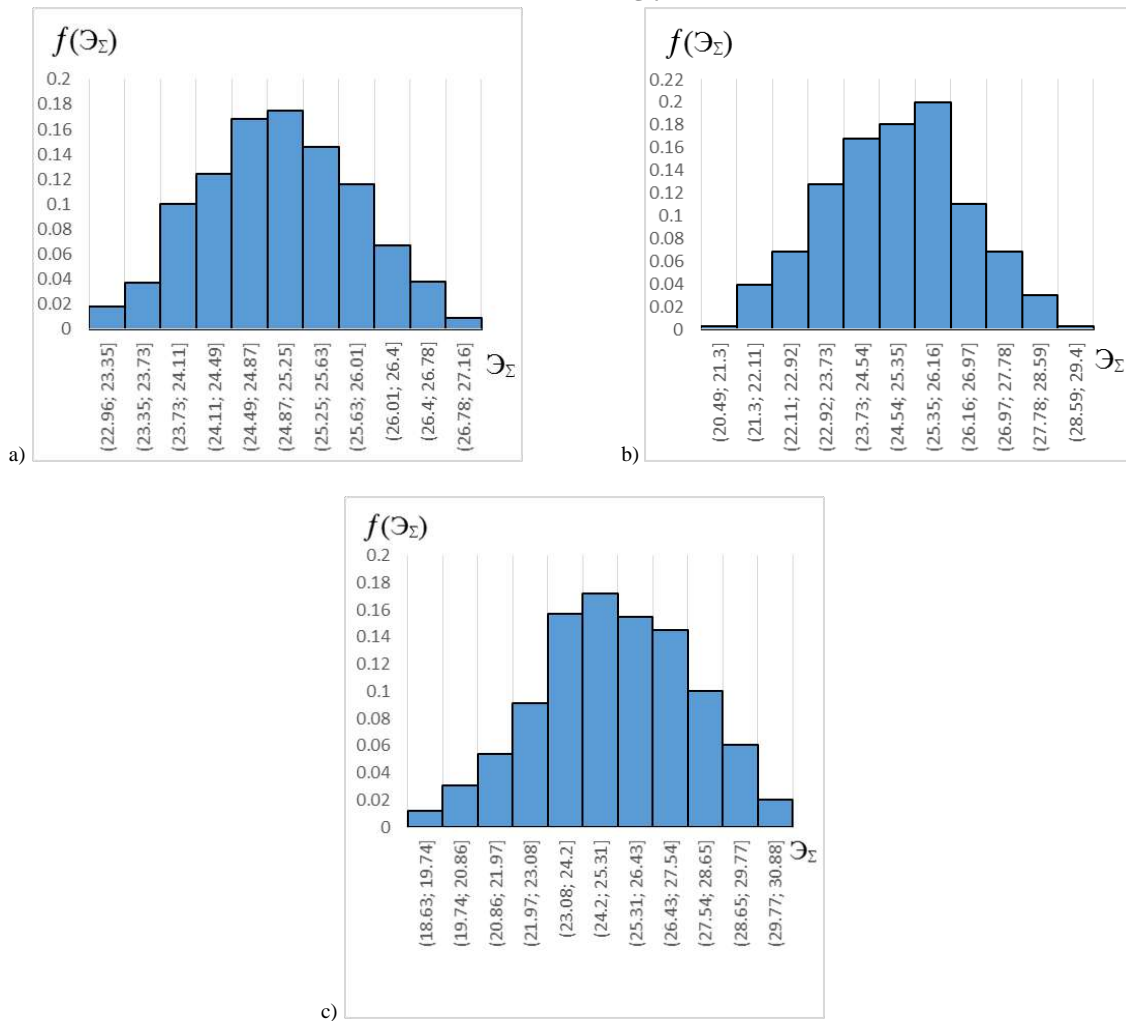


Fig. 4. Estimates of the distribution densities by: a)  $\alpha_\mathcal{J} = 0, \alpha_\lambda = 0.2$ ; b)  $\alpha_\mathcal{J} = 0.2, \alpha_\lambda = 0$ ; c)  $\alpha_\mathcal{J} = 0.2, \alpha_\lambda = 0.2$ .

The constructed estimates  $f(\Theta_\Sigma)$  became the basis for constructing  $H_\mathcal{J}$ , for various combinations of characteristics the statistical uncertainty of the graph components. Fig. 5 shows the resulting dependencies, corresponding to different values  $\varepsilon_0^{(\Theta_\Sigma)}$

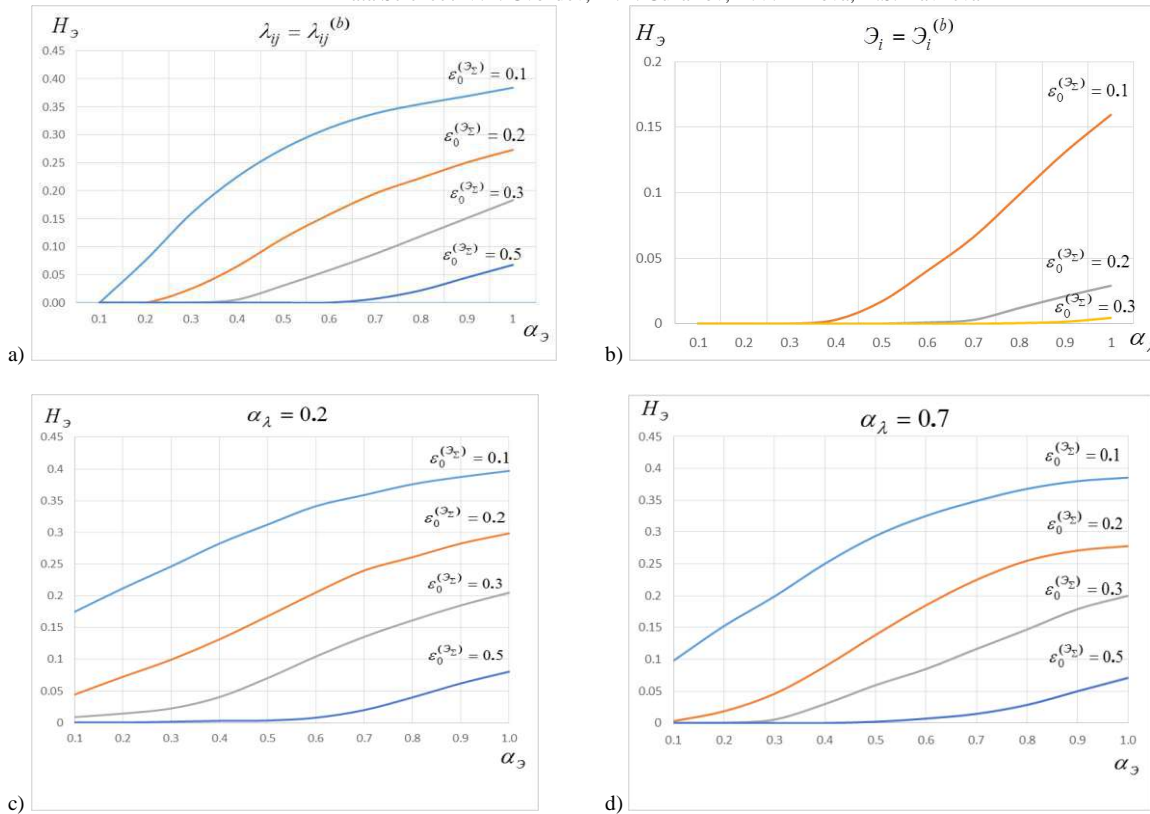


Fig. 5. Uncertainty characteristics dependences of average target efficiency on uncertainty characteristics of graph components.

The resulting dependences  $H_3$  allow us to solve a direct problem: an estimation of uncertainty characteristics of target average efficiency  $H_3$  on the information basis on model components parameters variability; and the inverse problem: an estimation limitations on the parameters variability of graph's nodes and edges based on the limitations on the characteristics of the target average efficiency.

An example of solving a direct problem. Given:  $\varepsilon_0^{(3\Sigma)}$ ;  $\alpha_3$ . It is believed that the variability of transitions intensities is absent. It is required to estimate the expected uncertainty  $H_3$ . The scheme for solving the problem is shown in Fig. 6.

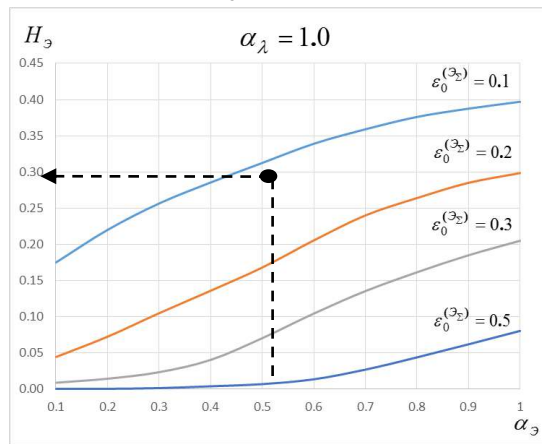


Fig. 6. An example of solving a direct problem.

An example of solving an inverse problem.

Given:  $H_3$ ;  $\varepsilon_0^{(3\Sigma)}$ ;  $\alpha_\lambda$ . It is required to estimate the possible value  $\alpha_3$ . Fig. 7 shows an example of solving an inverse problem with the selected value  $\alpha_\lambda$ . Thus, the proposed technique allows us to formalize the procedure for making conclusions about the vulnerability of the data processing and control system based on the analysis the characteristics of the external behavior of the system.

## 5. Conclusion

Nowadays DPCS play more and more of a substantial role as a vital component in systems that control complex objects. This promotes posing the problem of developing theoretical basis and development tools for managing the functional security of DPCS. One of the key tasks in solving such a problem is analyzing vulnerabilities of DPCS. They are affected by internal

properties (construction, component properties) and by external environment in which the system is operated. The allowed level of the vulnerability is determined by whether the deviation of the systems behavior from the base behavior is affecting the quality of control of a complex object.

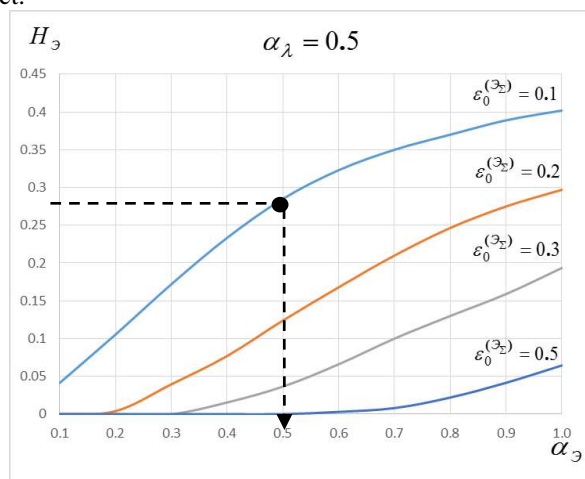


Fig. 7. An example of solving an inverse problem.

In this article a solution is given to estimate the possible deviation in components parameters of the model OF DPCS (such parameters reflect physical and information properties of the system). This solution is based of the limitations on alternation of the integral factor that shows the behavior of the system in different modes of operation, this behavior is the target efficiency of the system. Proposed solution follows known demands: inner properties of the system must be so that customer demands are fulfilled (Kano's model). The described solution makes it possible to solve not only the direct problem - making a conclusion whether the vulnerability of the data processing system is negligible with respect to alternations in parameters of this systems components but the inverse as well, finding tolerable levels in alterations in systems components parameters from an acceptable level of uncertainty of targeted efficiency's index.

## Acknowledgements

This work was supported by RFBR grant No. 17-07-00351 "Methodological basics of dependability assurance of transmission systems telemetry information with use of intelligent data analysis technologies".

## References

- [1] Lipaev VV. Functional security of software. Moscow: SYNTEG, 2004; 348 p.
- [2] Antamoshkin AN, Morgunova ON. Technique to study the effectiveness of complex hierarchical systems. Vestnik SibGAU 2006; 2 (9): 9–13.
- [3] Makhutov NA, Reznikov DO. Vulnerability assessment of technical systems and its place in the risk analysis procedure. Problems of risk analysis 2008; 5(3): 72–85.
- [4] Mladen AV. Software Reliability Engineering. Proceedings of the Annual Reliability and Maintainability Symposium. Los Angeles, California, USA, January 24-27, 2000.
- [5] Abde IMoez W, Nassar D, Shereshevsky M, Gradetsky N, Gunnalan R, Ammar HYuB, Mili M. Error Propagation in Software architectures. Proceedings of the Software Metics. 10th International Symposium, Washington, DC, USA. IEEE Computer Society, 2004; 384–393.
- [6] Khoshgoflaar TM, Munson JC. Predicting software development errors using complexity metrics. IEEE of Selected Areas in Communications 1990; 8(2): 253–261.
- [7] Bellini P, Bruno I, Nesi P, Rogai D. Comparing fault-proneness estimation models. Proc. of 10th IEEE International Conference on Engineering of Complex Computer Systems, 2005; 205–214.
- [8] Maevsky DA, Yaremchuk SA. Estimating the number of software defects based on the complexity metrics. Electrotechnic and computer systems: Scientific and Technical Journal 2012; 7(83): 113–120.
- [9] Michael CC, Jones RC. On the uniformity of error propagation in software. Proceedings of the Annual Conference on Computer Assurance, 1996; 68–76.
- [10] Liu XF. Software quality function deployment. IEEE Potentials 2000; 19(5): 14–16.
- [11] Shindo H. Application of QFD to Software and QFD Software Tools. Pre-Conference Workshops of the Fifth International Symposium on Quality Function Deployment and the First Brazilian Conference on Management of Product Development. Belo Horizonte, Brazil, 1999.
- [12] GOST R 51901.5-2005 Risk management. Application guide of reliability analysis methods. Moscow, Standartinform, 2005; 44 p.
- [13] Kuzin LT. Fundamentals of cybernetics. Vol. 1. Mathematical foundations of cybernetics. Textbook for students of technical colleges. M.: Energy, 1973; 504 p.
- [14] Jaynes ET. Information theory and statistical mechanics. The Physical Review 1957; 106(4): 620–630.
- [15] Gvozdev VE, Blinova DV, Davlieva AS, Teslenko VV. Construction of basic functioning efficiency models of the hardware-software complexes, based on the mathematical statistics methods. Software Engineering 2016; 7(11): 483–489.
- [16] Gvozdev VE, Blinova DV, Davlieva AS, Teslenko VV. Effect assessment of the system parameters variability on the functional vulnerability indexes of the hardware-software systems. Information Technologies for Intelligent Decision Making Support (ITIDS'2017). Proceedings of the 5th International Conference, Ufa, Russia, May 16-19, 2017. (in press)

# Generalized Model of Pulse Process for Dynamic Analysis of Sylov's Fuzzy Cognitive Maps

R.A. Isaev<sup>1</sup>, A.G. Podvesovskii<sup>1</sup>

<sup>1</sup>*Bryansk State Technical University, 50 let Oktyabrya boul. 7, 241035, Bryansk, Russia*

---

## Abstract

Pulse process as a means of dynamic analysis of cognitive models of semi-structured systems is studied. There is introduced and substantiated a generalized model of pulse process for Sylov's fuzzy cognitive maps, suggested its implementation for various semantic interpretations of concept interactions. The results of experimental validation of the proposed models are given.

*Keywords:* cognitive modeling; fuzzy cognitive map; dynamic analysis; pulse process

---

## 1. Introduction

One of the approaches to study semi-structured systems, widely used at the present time, is a cognitive approach. In accordance with the definition given in [1], this approach focuses on the development of formal models and methods supporting the intelligent process of solving problems because these models and methods take into account human cognitive capabilities (perception, conception, cognition, understanding, explanation) in solving management problems. Methods of structured, target and simulation modeling on the basis of cognitive approach are commonly combined by the general term "cognitive modeling". In general, cognitive modeling refers to the study of the structure of a system and the processes of its functioning and development by analyzing its cognitive model. The cognitive model of a system is based on a cognitive map, which reflects the subjective view of the researcher about it (individual or collective) as a set of semantic categories (called factors or concepts) and a set of cause-and-effect relations between them.

A cognitive model is an effective tool for exploration and estimation analysis of the situation. It does not give the opportunity to obtain accurate quantitative characteristics of the system under study, but it allows to assess the trends related to its functioning and development, and to identify significant factors influencing these processes mostly. Thanks to this we can search, generate and develop effective solutions for managing the system, as well as identify risks and develop strategies to reduce them.

Cognitive modeling starts with creating a cognitive map of the system under study on the basis of information received from experts. The next step includes directly modeling, which main objectives are forming and testing hypotheses of the system structure under study that can explain its behavior as well as developing strategies for situations in order to reach the specified targets.

The tasks solved by means of cognitive modeling can be divided into two groups:

1. The tasks of structured and target analysis:
  - finding the factors which have the most significant influence on targets;
  - identification of contradictions between the targets;
  - identification of feedback loops.
2. The tasks of dynamic analysis (scenario modeling):
  - self-development ("what if nothing is done?");
  - managed development:
    - direct task ("what if ...?");
    - inverse task ("how to do ...?").

Thus, with the help of scenario modeling it is possible to predict the state of the simulated system under different management actions as well as the search for alternative management solutions to bring the system to the target state.

The most common mathematical apparatus used to represent cognitive models and being the base of the methods for their analysis is fuzzy logic. Because of this there appeared a class of cognitive models based on different types of fuzzy cognitive maps (FCM) – a very detailed overview of such models can be found in monograph [3]. One of FCM varieties, well-proven in practical problems of analyzing and modeling of ill-structured organizational, social and economic systems are Sylov's FCM firstly proposed in [7] and representing the development of signed cognitive maps [6]. For this type of FCM there was developed quite a wide range of methods of structured and target analysis based on the study of such FCM factors as consonance, dissonance and action. A detailed description of these methods can be found in the original monograph [7], and some examples of their application in the study of different organizational and social systems – in papers [2, 4]. The problem of developing and improving dynamic analysis methods of Sylov's FCM was given far less attention. This article discusses the approach to dynamic analysis of this type FCM with the use of a generalized model of pulse process. The proposed approach is based on the notion of pulse process, originally introduced in [6] for the class of signed cognitive maps, generalizing this concept by extending it to the class of FCM, and is a development of the approach, first mentioned in [5] and is described in more detail in monograph [4] (section 3.2).

## 2. Formal definition and structure of Sylov's fuzzy cognitive map

As it has already been mentioned, the cognitive model is based on formalization of cause-and-effect relations which occur between the factors characterizing the system under study. The result of formalization is representing the system in the form of cause-and-effect network, called a cognitive map and having the following form:

$$G = \langle E, W \rangle,$$

where  $E = \{e_1, e_2, \dots, e_K\}$  is a set of factors (also called concepts),  $W$  is a binary relation over set  $E$ , which specifies a set of cause-and-effect relations between its elements.

Concepts can specify both relative (qualitative) characteristics of the system under study, such as popularity, social tension, and absolute, measurable values – population size, cost, etc. Besides, every concept  $e_i$  is connected with state variable  $v_i$ , which specifies the value of the corresponding index at a particular instant. State variables can possess values expressed on a certain scale, within the established limits. Value  $v_i(t)$  of state variable at instant  $t$  is called the state of concept  $e_i$  at the given instant. Thus, the state of the simulated system at any given instant is described by the state of all the concepts included in its cognitive map.

Concepts  $e_i$  and  $e_j$  are considered to be connected by relation  $W$  (designated as  $(e_i, e_j) \in W$  or  $e_i W e_j$ ) if changing the state of concept  $e_i$  (cause) results in changing the state of concept  $e_j$  (effect). In this case we say that concept  $e_i$  has an influence on concept  $e_j$ . Besides, if increasing the value of the state variable of concept-cause leads to increasing the value of the state variable concept – effect, then the influence is considered positive (amplification), and if the decrease – negative (inhibition). Thus, relation  $W$  can be represented as a union of two disjoint sets  $W = W^+ \cup W^-$ , where  $W^+$  is a set of positive relations and  $W^-$  is a set of negative relations.

Fuzzy cognitive model is based on the assumption that the influence between concepts may vary in intensity, besides, this intensity may be constant or variable in time. In order to take into account this assumption,  $W$  is set as a fuzzy relation, besides, the way of its setting depends on the adopted approach to formalization of cause-and-effect relations. Cognitive map with fuzzy relation  $W$  is called a fuzzy cognitive map.

Sylov's fuzzy cognitive map represents FCM, characterized by the following features.

1. State variables of concepts can possess values on the interval  $[0, 1]$ .
2. The intensity of interactions is considered constant, so relation  $W$  is specified as a set of numbers  $w_{ij}$ , characterizing the direction and degree of intensity (weight) of influence between concepts  $e_i$  and  $e_j$ :

$$w_{ij} = w(e_i, e_j),$$

where  $w$  is a normalized index of influence intensity (characteristic function of relation  $W$ ) with the following properties:

- a)  $-1 \leq w_{ij} \leq 1$ ;
- b)  $w_{ij} = 0$ , if  $e_j$  does not depend on  $e_i$  (no influence);
- c)  $w_{ij} = 1$  if the positive influence of  $e_i$  on  $e_j$  is maximum, i.e. when any changes in the system related to concept  $e_j$  is univocally determined by the actions associated with concept  $e_i$ ;
- d)  $w_{ij} = -1$  if negative influence is maximum, i.e. when any changes related to concept  $e_j$  are clearly constrained by the actions associated with concept  $e_i$ ;
- e)  $w_{ij}$  possesses the value from the interval  $(-1, 1)$ , when there is an intermediate degree of positive or negative influence.

It is easy to notice that FCM of this structure can be graphically represented as a weighted directed graph, which points correspond to the elements of set  $E$  (concepts) and arcs – to nonzero elements of relation  $W$  (cause-and-effect relations). Each arc has a weight which is specified by the appropriate value  $w_{ij}$ . In this case, relation  $W$  can be represented as a matrix of dimension  $n \times n$  (where  $n$  is a number of concepts in the system), which can be considered as the adjacency matrix of the graph and is called a cognitive matrix. In addition, each point of the graph also has a weight which corresponds to the concept state and can change over time.

## 3. Pulse process as a means of dynamic analysis of cognitive maps

The basis of dynamic analysis of cognitive maps is modeling of dynamics of concept states over time. Besides, concept state may change, firstly, due to changes of state of other concepts influencing this one, and, secondly, due to external influences. External influence on the concept is understood as change of its state relative to the current one under the impact of external factors, i.e. irrelatively the concepts included in the cognitive map. At the same time external influence can be targeted, i.e. it comes from the subject carrying out management of the system, and untargeted, i.e. due to external factors to the system which are beyond control. Accordingly, in the first case we will talk about control action, and in the second case – about perturbation action (or perturbations).

To describe the dynamics of concept states we will use pulse processes. The basis of this approach is the assumption that changes of all concept states occur at discrete moment of time. State change of concept of  $e_i$  at instant  $t$  will be called pulse and denote as  $p_i(t)$ . Thus,

$$p_i(t) = v_i(t) - v_i(t-1).$$

Additionally, it is assumed that influence is transmitted by one step: changing the state of the concept-cause at instant  $t$  result in changing the state of the concept-effect at instant  $t + 1$ .

Let us first give the model of pulse process for signed cognitive maps, i.e. maps which take into account only the directions of influence and do not take into account their intensity. For these maps values  $w_{ij}$  can only take values  $-1, 0$  or  $1$ , and respectively, the graph arcs are marked by signs “+” and “-“. The model of pulse process was proposed in [6]:



$$p_i(t+1) = \sum_{j=1}^K \text{sgn}(w_{ji}) p_j(t),$$

accordingly

$$v_i(t+1) = v_i(t) + \sum_{j=1}^K \text{sgn}(w_{ji}) p_j(t).$$

Thus, changing the state (pulse) of each concept in the current step is determined by the pulses of all concepts influencing it and by the ratio of influence signs. Besides, transfer of positive influence is neutralized by simultaneous transfer of negative influence, and vice versa.

In [4, 5], a modified model of pulse process for Sylov's FCM was proposed. The model takes into account the transfer of influences between concepts and external influences:

$$v_i(t+1) = \min \left( v_i(t) + u_i(t+1) + q_i(t+1) + \sum_{j=1}^K w_{ji} p_j(t), 1 \right), \quad (1)$$

where  $u_i(t+1)$  is control action on concept  $e_i$  at instant  $t+1$ ;  $q_i(t+1)$  – is disturbance  $e_i$  at instant  $t+1$ .

#### 4. Generalized model of pulse process

In the framework of model (1) it is assumed that the state change of concept  $e_j$  is equal to the difference between its state at the current step and the previous step:

$$p_j(t) = v_j(t) - v_j(t-1).$$

Thus, in dynamic modeling in order to determine the state of dependent concepts we take into account *absolute change* of states of influencing concepts. This approach is acceptable, but at the same time, it is not the only possible one. In this regard, it is advisable to consider other, alternative approaches to interpret the interaction of concepts and propose alternative models of pulse process on their basis.

However, it is necessary to define a number of requirements to models of pulse process, which must be met by all proposed models in the future, regardless of the assumptions which they are based on.

Firstly, the model of pulse process should unambiguously determine the state of arbitrary concept  $e_i$  at instant  $(t+1)$ , using for this purpose the following available information:

- the state of the same concept  $e_i$  at instant  $t$ ;
- the state of concepts  $e_j, \dots, e_k$ , influencing concept  $e_i$ , at instant  $t$ ;
- the state of these concepts influencing  $e_i$ , at instant  $(t-1)$ ;
- connection weights (influence intensity)  $w_{ji}, \dots, w_{ki}$  between all dependent concepts and  $e_i$ ;
- control and disturbance influences on  $e_i$  at instant  $(t+1)$ , if there are any.

Or, more formally:

$$v_i(t+1) = f(v_i(t), v_j(t), \dots, v_k(t), v_j(t-1), \dots, v_k(t-1), w_{ji}, \dots, w_{ki}, u_i(t+1), q_i(t+1)). \quad (2)$$

Secondly, the following conditions should be met:

- the values of state variables of concepts should belong to the interval  $[0, 1]$ , that is  $v_i(t+1) \in [0, 1]$ ;
- if influence intensity between concepts  $e_j$  and  $e_i$  is equal to 0, then changing  $e_j$  state should not cause changing  $e_i$  state;
- if the state of influencing concepts at the previous step did not change ( $v_j(t) = v_j(t-1)$  for all  $j$ ), and there is no control and disturbance influence, then the state of the dependent concept at the current step should not change:  $v_i(t+1) = v_i(t)$ ;
- when increasing (decreasing) the state of influencing concept and the positive relation, the state of dependent concept should *not decrease (not increase)*:  $v_i(t+1) \geq v_i(t)$  if  $w_{ji} > 0$  and  $v_j(t) > v_j(t-1)$ ;  $v_i(t+1) \leq v_i(t)$  if  $w_{ji} > 0$  and  $v_j(t) < v_j(t-1)$ ;
- when increasing (decreasing) the state of influencing concept and the negative relation, the state of dependent concept should *not increase (not decrease)*:  $v_i(t+1) \leq v_i(t)$  if  $w_{ji} < 0$  and  $v_j(t) > v_j(t-1)$ ;  $v_i(t+1) \geq v_i(t)$  if  $w_{ji} < 0$  and  $v_j(t) < v_j(t-1)$ ;
- more significant change of the influencing concept with other things being equal should result in more significant change of the dependent concept:  $p_i^1(t+1) \geq p_i^2(t+1)$ , if  $p_j^1(t) \geq p_j^2(t)$ ;
- higher intensity of the influence with other things being equal should result in more significant change of the dependent concept:  $p_i^1(t+1) \geq p_i^2(t+1)$ , if  $w_{ji}^1 \geq w_{ji}^2$ .

The expression (2) together with the above mentioned conditions we call a generalized model of pulse process. This model, on the one hand, comprises model (1) as a possible particular case, and on the other hand, it creates the base for building other implementations of pulse process model.

## 5. Implementation of the generalized model of pulse process

Let us consider the alternative implementations of the described generalized model of pulse process, involving different interpretations of concept interaction.

### 5.1. Model of impulse process, based on relative changes of concept states

We assume that concept influence on the system is determined not by changes of its condition in general, but how significant is this change relative to the previous state of this concept. In other words, we consider the relative state changes of the concepts, not absolute one.

With this purpose we will consider pulse  $p_i(t)$  as a relative state change of concept  $e_i$  at instant  $t$ :

$$p_i(t) = \frac{v_i(t) - v_i(t-1)}{v_i(t-1)}.$$

Thus, the value of pulse  $p_i(t)$  shows *what fraction* of its state concept  $e_i$  changes at instant  $(t-1)$ .

Now, let us define the way of transferring influence between directly related concepts. Let there is a relation between concepts  $e_j$  and  $e_i$ , which strength is equal to  $w_{ji}$ . At the beginning, knowing  $p_j(t)$  – relative change of state  $e_j$  at instant  $t$ , let us define the relative change of state  $e_i$  at instant  $(t+1)$ .

It is necessary to consider the conditions of the generalized model, and the following additional conditions:

- if  $p_j(t) = 0$  or  $w_{ji} = 0$ , then  $p_i(t+1) = 0$ ;
- if  $w_{ji} = 1$ , then  $p_i(t+1) = p_j(t)$ .

The following product satisfies these conditions:

$$p_i(t+1) = w_{ji} p_j(t).$$

Finally, let us define the state of concept  $e_i$  at instant  $(t+1)$ . Note that

$$p_i(t+1) = \frac{v_i(t+1) - v_i(t)}{v_i(t)}.$$

So,

$$v_i(t+1) = v_i(t) + v_i(t) w_{ji} p_j(t).$$

The resulting model is easily generalized in the case of multiple influencing concepts:

$$v_i(t+1) = v_i(t) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t).$$

As one of the conditions of the generalized model is that the concept states are within the interval  $[0, 1]$ , then we should add the following constraints to the model:

$$v_i(t+1) = \max \left( \min \left( v_i(t) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t), 1 \right), 0 \right).$$

Besides, control and disturbance influences on  $e_i$  should also be defined in terms of relative changes. For example, control influence  $u_i(t+1) = 0,1$  means “to increase the value of a state variable of  $i$ -concept by 10% of its current value”.

Thus, we obtain the final version of the model:

$$v_i(t+1) = \max \left( \min \left( v_i(t) + v_i(t) u_i(t+1) + v_i(t) q_i(t+1) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t), 1 \right), 0 \right). \quad (3)$$

### 5.2. Multiplicative model of pulse process

Let us consider another model which also takes into account relative changes of concept states but implies slightly different interpretation of these changes. This model is not equivalent to that one described above, but they both come from similar preconditions.

In this case, relative state change of concept  $e_j$  shows *how much* this concept has changed at instant  $t$  compared with its condition at instant  $(t-1)$ :

$$p_i(t) = \frac{v_i(t)}{v_i(t-1)}.$$

Let us define the way of transferring influence between directly related concepts. In addition to the terms of the generalized model, in this case, the following conditions should be taken into account:

- if  $w_{ji} = 1$ , then  $p_i(t+1) = p_j(t)$ ;
- if  $w_{ji} = 0$  or  $p_j(t) = 1$ , then  $p_i(t+1) = 1$ ;

- if  $w_{ji} = -1$ , then  $p_i(t+1) = \frac{1}{p_j(t)}$ .

Operation of exponentiation satisfies these conditions:

$$p_i(t+1) = (p_j(t))^{w_{ji}}$$

Now it is easy to determine the state of concept  $e_i$  at instant  $(t+1)$ :

$$v_i(t+1) = v_i(t)(p_j(t))^{w_{ji}}$$

Generalization of the model in the case of multiple influencing concepts will be the following:

$$v_i(t+1) = v_i(t) \prod_{j=1}^K (p_j(t))^{w_{ji}}$$

This model does not use negative values (excluding connection weights used as indexes), thus, fulfillment of condition  $v_i(t+1) \geq 0$  is guaranteed. To fulfill the other condition of the generalized model, namely  $v_i(t+1) \leq 1$ , we add the constraint:

$$v_i(t+1) = \min \left( v_i(t) \prod_{j=1}^K (p_j(t))^{w_{ji}}, 1 \right)$$

Control influence and disturbance in this model should be specified on the basis of interpretation “a concept state has changed  $n$  times”. For example, control influence  $u_i(t+1) = 2$  means “to increase the concept state 2 times in comparison with the current state”.

So, here is the final version of the model:

$$v_i(t+1) = \min \left( v_i(t) u_i(t+1) q_i(t+1) \prod_{j=1}^K (p_j(t))^{w_{ji}}, 1 \right) \tag{4}$$

### 6. Experimental validation of the model of pulse process under study

For the purpose of experimental validation and comparison of the examined models, let us perform dynamic analysis of cognitive maps using each of them with the same initial data.

Fig. 1 shows a fragment of the cognitive map used for the experiment. Connection weights have the following values:  $w_{12} = 0,9$ ;  $w_{23} = -0,8$ ;  $w_{31} = 0,7$ . Initial concept states are specified as the following:  $v_1(1) = 0,2$ ;  $v_2(1) = 0,3$ ;  $v_3(1) = 0,8$ .

Let there is control influence on concept 1, which results in its transfer into state  $v_1(2) = 0,6$ . Under the influence of the initial pulse, concept states begin to change in accordance with the rules defined by each model of pulse process.

Fig. 2-4 give schedules of changes of concept states during operation of three models of pulse process. The horizontal axis shows modeling steps, the vertical axis shows the state of the appropriate concept. Schedules have the following signs:

- “Model 1” – the results obtained using the additive model (1);
- “Model 2” – the results obtained using the additive model (3) based on the relative state changes of concepts;
- “Model 3” – the results obtained using the multiplicative model (4) based on the relative state changes of concepts.

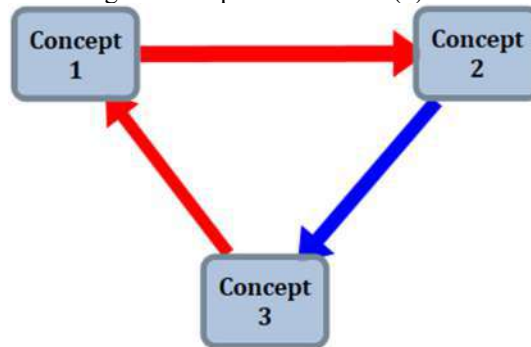


Fig. 1. Fragment of a fuzzy cognitive map used for the experiment.

Of the greatest interest for the interpretation is the transfer of influence between directly related concepts, differently occurring in the framework of different models, which result in different results in the end. Thus, in models 2 and 3, implying relative change of concept states, the state of the second concept on the 3rd modeling step increased more than in model 1. Similarly, relative changes result in more significant decrease in the state of the third concept on the 4th step. Similar regularity is typical for the subsequent steps.

Describing the results in general, the following should be noted:

- all models operate correctly regarding the influence transfer: the direction of changing concept states correspond to the signs of influences;
- all models are stable: pulse decays with time, which results in transferring the system in a stable state;

- according to the results of modeling the state of each concept has changed in the same direction for all models (the states of the first and second concepts have increased, the state of the third one has decreased in comparison with the initial one), these results are generally consistent with intuitive understanding of the nature of system changes, which also proves the correctness of models;
- differences in predictions obtained by means of different models are quite well explained by the assumptions (concerning the nature of influences between concepts), which they are based on.

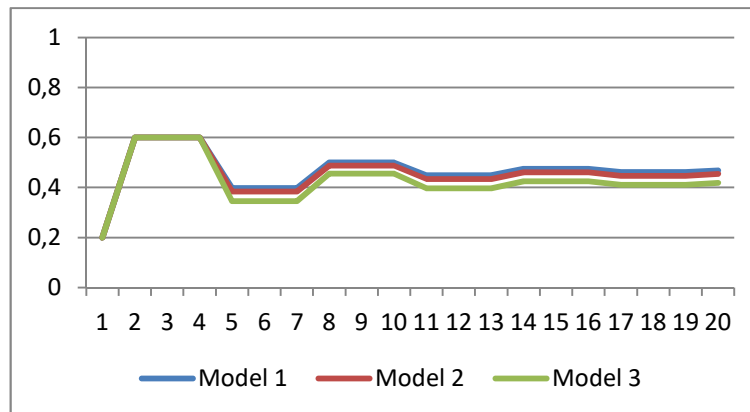


Fig. 2. Dynamics of state change of concept 1.

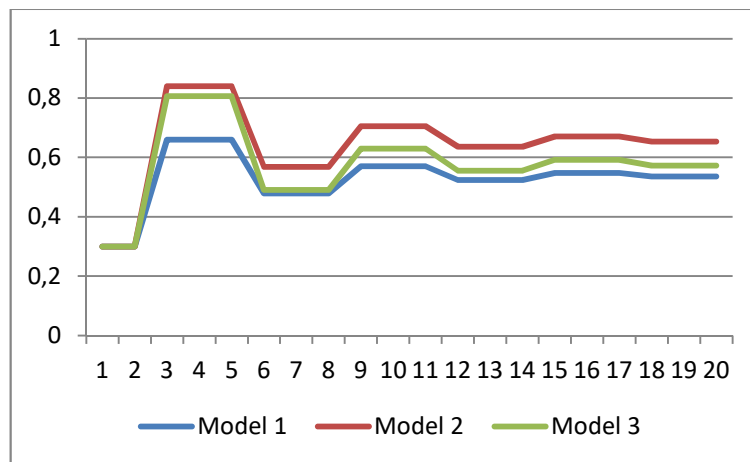


Fig. 3. Dynamics of state change of concept 2.

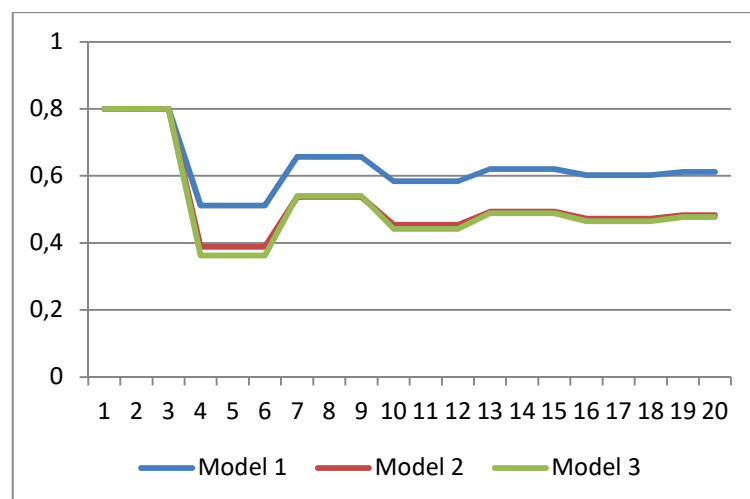


Fig. 4. Dynamics of state change of concept 3.

## 7. Conclusion

The paper considers a generalized model of pulse process for Sylov's fuzzy cognitive maps. This model, on the one hand, represents a generalization of previously developed models, and on the other hand, can serve as a basis for building other variations of pulse process.

Also, there are proposed alternative implementations of this generalized model of pulse process, involving different interpretations of concept interactions. Experimental validation of these implementations is carried out, and its results confirm their correctness and operability.

Among the possible directions for further research, the following are of the greatest interest:

- identifying characteristics and making requirements to the methods of expert identification of FCM parameters in different models of pulse process;
- identifying characteristics and making requirements to the methods of identification of FCM parameters on the basis of statistical data in different models of pulse process;
- development of methods for selecting an optimal model of pulse process on the basis of the analysis of available statistical and expert data.

## References

- [1] Avdeeva ZK, Kovriga SV, Makarenko DI. Cognitive modeling for solving semi-structured management systems (situations). *Managing Large Systems* 2007; 16: 26–39. (in Russian)
- [2] Averchenkov VI, Kozhukhar VM, Podvesovskii AG, Sazonova AS. Monitoring and Prediction of Regional Demand for Highest Scientific Degree Specialists: monograph. Bryansk: Bryansk State Technical University Press, 2010; 163 p. (in Russian)
- [3] Borisov VV, Kruglov VV, Fedulov AS. Fuzzy Models and Networks. Moscow: “Goryachaya Liniya – Telekom” Publisher, 2012; 284 p. (in Russian)
- [4] Erokhin DV, Lagerev DG, Laricheva EA, Podvesovskii AG. Strategic Enterprise Innovation Management: monograph. Bryansk: Bryansk State Technical University Press, 2010; 196 p. (in Russian)
- [5] Podvesovskii AG, Lagerev DG, Korostelyov DA. Application of fuzzy cognitive models for construction of alternatives set in decision problems. *Bulletin of Bryansk State Technical University*, 2009; 4(24): 77–84. (in Russian)
- [6] Roberts FS. *Discrete Mathematical Models with Application to Social, Biological and Environmental Problems*. Prentice-Hall, Englewood Cliffs, 1976.
- [7] Sylov VB. *Strategic Decision Making in Fuzzy Environment*. Moscow: “INPRO-RES” Publisher, 1995; 228 p. (in Russian)

# Capabilities of the adaptive regression modeling package SSOR

G.R. Kadyrova<sup>1</sup>, T.E. Rodionova<sup>1</sup>

<sup>1</sup>Ulyanovsk State Technical University, Severnyi Venets str., 32, 432027, Ulyanovsk, Russia

---

## Abstract

The original statistical package «The system of searching for optimal regressions» is presented in the paper. The package allows performing high-precision statistical (regression) modeling of processes or phenomena with the subsequent use of models for the forecast of their output characteristics. The approach implemented in the package reduces the dimension of the model, increases the accuracy of parameter determination and improves the quality of the forecast. The effectiveness of the approach is directly proportional to the dimensionality, the degree of noisiness, and the multicollinear nature of the initial data. The features of the package make it a perspective mathematical tool for high-precision statistical calculations.

*Keywords:* regression modeling; prediction; methods of structural identification; model quality criteria; software package

---

## 1. Introduction

The software package «The system of searching for optimal regressions» (SSOR) is a specialized system implementing the strategy of adaptive regression modeling (ARM) [1].

At the initial stage, the ARM–approach provides the application of linear regression analysis (RA), which assumes the model postulating, least–squares method (LS) estimation, statistical analysis of the model and its components. LS–estimates  $\hat{\beta}$  and forecasts  $\hat{Y}$  are considered to be the best linear estimates (BLE) under certain assumptions. Unfortunately, these assumptions are violated in many cases. This leads to a distortion of LS–estimates  $\hat{\beta}$ , entails an uncontrolled increase in random and systematic errors of forecasts  $\hat{Y}$ .

At the following stages the ARM–approach includes checking the compliance of the RA–LS hypotheses, ranking the violations by the degree of distortion of the properties of the best linear estimates or depending on the purpose of the model (forecast, description or description and forecast), consistent adaptation to violations by applying appropriate computational procedures, repeated check of violations and ranking if necessary.

The main difficulties in the practical implementation of the RM–approach are as follows: selection of a global (or integrated) criterion of optimality; satisfactory solution of the problem of structural identification in conditions of high dimensionality; selection of the optimal route for checking the application conditions of the RA–LS scheme and the corresponding adaptation. The SSOR package resolves these problems [3, 8].

The main purpose of the package is to obtain regression models of processes, phenomena or functioning of objects with their subsequent use for forecasting output characteristics (responses) [9, 14]. The need for such a system is generated by great difficulties in performing such work, which requires both a multivariate calculation, and the application of various methods for estimating parameters and structural identification and analysis of residuals in the selected scenario for verifying compliance with the LS assumptions.

## 2. Adaptive RM

In developing and using forecast models the main goal is to achieve the properties of the best linear estimation (consistency, unbiasedness, efficiency) for the predicted value  $\hat{Y}$ . These properties are primarily ensured by the selection of the corresponding (optimal according to the given criterion) structure of the model from the set (on the basis of the postulated model) of competing structures. Thus, the problem of not only parametric, but also structural identification is solved.

In most cases, the postulated model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon; \quad i = \overline{1, n} \quad (1)$$

is not optimal (adequate) to observations. If linear dependence (1) is considered to be suitable, the dimensionality of the model will be the main problem.

On the one side, for fear of losing significant factors, a researcher tries to include as many of them as possible in the right-hand side of the model (1). Therefore, as a rule, the model is overdetermined, which leads to: a) economic costs; b) the inclusion of non-informative, low-information and duplicating variables. The latter leads to an increase in the variance of the forecast  $\hat{Y}$  for the forecast model and to a decrease in the accuracy of the estimation of the  $\hat{\beta}$ -coefficients in the parametric model.

On the other side, an underdetermined model that does not contain significant factors leads to a systematic error  $\Delta$  in the forecast. Here the problem arises of measuring the displacement  $\Delta$  with the magnitude of the random error of the forecast in the overdetermined model. Most often the random error is greater than the systematic error. In addition, in some estimation methods other than LS, the model is deliberately burdened with bias to reduce the forecast error.

Thus, putting forward the hypothesis (1), a researcher encounters a set of competing models (structures) containing  $x_0 (x_0 = 1)$  and some regressors from the set  $\{x_1, \dots, x_{p-1}\}$ . Since each variable  $x_j (j = \overline{1, p-1})$  can either enter the equation or not we'll

have  $2^{p-1}$  models. From this set of structures one or more competing models must be selected according to a given quality criterion.

If a standard regression analysis is used, in applied statistics after analyzing the model as a whole and its individual terms we use one-criterion search for the optimal structure. If it is impossible to apply a complete search of structures, one or another known type of incomplete search is used according to one of the model quality criteria (mean square error  $\sigma$ , selective coefficient of multiple correlation  $R$ ,  $F$ -criterion etc.).

The most preferable for structural identification is the error in the control sample. This criterion to the maximum extent reflects the real random and systematic errors of the forecast (response) and does not have a systematic move in relation to the dimension. The error in the control sample is, naturally as «true», as the «true» control values  $y_i$  are burdened, in their turn, with various errors.

The application of the RM approach requires the development of multi-criteria search algorithms. In the general case, in order to obtain an adequate data processing model, it is necessary to solve the multicriteria optimization problem by successive adaptation to violations of the RA–LS conditions.

In some cases two-criterion methods are quite effective. The SSOR a step-by-step regression method (inclusion-exclusion scheme) is implemented using in addition to the  $F$ -criterion a number of other measures of comparison [10].

### 3. Criteria for the quality of a model

One of the most important tasks in the analysis of data is the problem of choosing a criterion for comparing competing descriptions.

The SSOR, in addition to the quality criteria of the model on the training sample  $(t, \sigma, F, R)$ , the possibility of calculating the error on the control sample and the error on the «sliding» control sample is given [1,19].

The quality of the RA model is usually determined by the following criteria:

- mean square error  $\sigma$ , which is used both to assess the adequacy of the model, and to compare different models with each other;

- selective multiple correlation coefficient  $R$ , which is used as a linear link measure (1): the larger the value of  $R$  ( $0 \leq R \leq 1$ ), the stronger the connection, i.e. the better the approximating function corresponds to observations, the high value of  $R$  also guarantees the suitability of the forecast model;

- $F$ -criterion, when  $F > 4F_T(\alpha; p-1, n-p)$  ( $F_T$  – critical value, taken from the table for the  $F$ -criterion) model is recognized as worthy of attention for its use for forecasting.

These quality criteria characterize the adequacy of the model only with respect to the sampling points used for its construction (training sample). This is the first stage in the study of the model in which an experimenter must be convinced that the model corresponds to observations.

If the model is intended for forecasting, then one must be sure of its suitability for determining the region that does not coincide with the sampling points  $y_i$ .

Control points are used to assess external adequacy (forecast accuracy). The initial sample is divided into training and control. The first sample builds a model or set of models; the second one estimates its adequacy or discrimination by statisticians is made.

The error in the control sample is based on an analysis of the discrepancies between the forecast  $\hat{Y}$  and the known observed value  $Y$  for objects that did not participate in obtaining the model.

Since when working with small samples there is no possibility to divide them into a training sample and control one with a sufficiently large number of points, we suggest to use a criterion based on a «sliding» control sample to assess external adequacy. If, sequentially, we deduce each of the sampling objects from it, assuming that this object is a control one, and recalculating the model parameters again, the differences between  $y_i$  and  $\hat{y}_i$  for a sliding control point  $\Delta_i = \text{«Observation minus the forecast»}$  ( $i = 1, n$ ; where  $n$  – total number of objects) can be used to calculate the error on a «sliding» control sample.

Consecutive exclusion of objects, corresponding to the removal of certain rows from the data matrix makes it possible to formulate an artificially new sample (check or control) of the same volume as the original one.

All procedures of structural-parametric identification included in the package realize the calculation of the statistics considered and the search for the optimal structure of the model. More detailed criteria for comparing competing models are considered in [19].

### 4. The software package SSOR

In organizing the optimal RM-strategy it is necessary to take into account the availability of various samples, assumptions, classes of functions, estimation methods, quality measures and their sets for the principle of multicriteria, structural identification methods competing in accordance with the principle of non-conclusive solutions of adaptation strategies to the violation of assumptions.

The practical application of RM first of all requires the full automation of all declared procedures. For this purpose the corresponding software was developed.

The SSOR package includes the following modules:

- 1) control module;

- 2 request generation module;
- 3) library of functional procedures;
- 4) script block;
- 5) system configuration block;
- 6) data editor block;
- 7) table formation block;
- 8) guide.

The main tool for positive impact on the predictive properties of the model is the algorithm for finding its optimal structure. The package includes the following structural-parametric identification procedures:

- multiple linear regression,
- comb regression,
- robust estimation,
- complete search of structures,
- incomplete search of structures (search with restriction on the number of included regressors in the model),
- search of normal systems,
- step-by-step regression with inclusion-exclusion,
- random search with adaptation,
- random search with a return.

These procedures can be performed both in automatic mode to process a number of data samples and to process a single sample of data according to the realized optimal scenario [18].

The package implements the procedure for constructing and analyzing the residue schedule, which is a useful statistical tool for testing the adequacy of the estimated regression model to the available data.

Competitiveness of SSOR with other statistical packages can be described as:

- using new methods of structural identification: full search, partial search of overdetermined and normal systems, multi-criteria method of step-by-step regression with inclusion-exclusion;
- using flexible tool for building comparative tables;
- using, in addition to the classical quality criteria of the model in the training sample, the quality criteria of the model in the control sample and the errors in the «sliding» control sample, which allows an external model adequacy assessment (forecast accuracy) to be performed.

At present work is under way to enhance the capabilities of the SSOR and its intellectualization [15, 17].

## 5. Using the SSOR package

The SSOR package can be used to solve the problems of the least-squares method (problems of recovering dependencies from excessive indirect observations) and regression analysis in any areas (ecology, technological processes, economics, sociology etc.), various tasks requiring restoration of the empirical relationship between the output process parameter and the input set.

In processing aerospace photographs [2] and solving a number of photogrammetric problems [4] the use of SSOR by computational experiments allowed to obtain the following results:

1. Obtaining models for transforming coordinates from small samples with a variance of the accuracy estimation of 1.2-100 times smaller than the variance in the standard approach, which corresponds to an increase in the approximation accuracy when applying PM up to several times.

2. Increase of accuracy in the use of RM is ensured by the procedure of structural identification. The implementation of the latter implies the formation of a set of competing structures based on the initial perspective model and the search for the optimal structure according to a given quality criterion.

3. Search for a model that is optimal by error on a «sliding» control sample leads to a model with better predictive properties than models that are optimal for the mean square error and allows to solve the problem of selecting a set of regressors that is informative by the t-criterion. In 70% of all cases models that are optimal for the mean square error contain low-information terms. Models that are optimal by error on a «sliding» control sample contain only insignificant terms only in 17% of all cases. Models containing little informative terms obtained by mistake on a «sliding» control sample contain one insignificant regressor, while models derived from the mean square error usually have two or more little informative terms. It was estimated that an improvement in external accuracy makes a significant difference in the quality of the error in the «sliding» control sample. The analysis showed that application of this criterion gives a significant improvement in predictive properties compared to the mean square error. The stability of the conclusions with respect to the observations included in the control sample was verified and confirmed by 10 random experiments for each of the three randomly selected images.

The SSOR package was successfully used to process laser [5] and high-dimensional radiointerferometric data [6, 7], for assessing the quality of drinking water [11, 16], for processing socio-economic indicators [12, 13].

## 6. Conclusion

The SSOR package can be useful in the development of forecast models in high-precision areas of knowledge, in technological processes with input characteristics that contain interdependent, non-informational or little informational factors



and in socio-economic phenomena and environmental situations. The application of the package provides an increase in the accuracy of forecasting using the optimal model up to several times.

## References

- [1] Valeev SG, Kadyrova GR. The system of searching for optimal regressions: tutorial . Kasa : FEN, 2003; 160 p.
- [2] Kadyrova GR, Bilibina NA, Bugaevskii LM, Valeev SG. Regression models for transforming images in aerocosmic pictures. *Izvestiya Vuzov. Geodezy and Aerophotography* 1997; 1: 56–66.
- [3] Valeev SG, Kadyrova GR. Automatic system for solving least-squares method tasks. *Izvestiya Vuzov. Geodezy and Aerophotography* 1999; 6: 124–130.
- [4] Valeev SG, Kadyrova GR. Optimal reduction models in photographical astronomy. *Izvestiya Vuzov. Geodezy and Aerophotography* 2002; 3: 58–69.
- [5] Valeev SG, Rodionova TE. The method of stepwise orthogonalization of the basis and its using during least-squares task. *Izvestiya Vuzov. Geodezy and Aerophotography* 2003; 6: 3–14.
- [6] Valeev SG, Rodionova TE, Zharov VE. Methodic of statistical processing of RSDB-observings. *Izvestiya Vuzov. Geodezy and Aerophotography* 2008; 1: 13–18.
- [7] Valeev SG, Rodionova TE, Zharov VE. Computational experiments for processing of RSDB-observings. *Izvestiya Vuzov. Geodezy and Aerophotography* 2008; 2: 94–100.
- [8] Valeev SG, Kadyrova GR, Turchenco AA. Software system for optimal regression searching. *Issues of modern science and practice. Technical science* 2008; 4(14): 97–101.
- [9] Kadyrova GR. Estimation and prediction of the state of a technical object based on regression models of regressions. *Automation of management processes* 2015; 4(42): 90–95.
- [10] Kadyrova GR. Modification of the stepwise regression method for obtaining mathematical models for predicting the behavior of an object. *Automation of management processes* 2016; 3(45): 65–70.
- [11] Rodionova TE. Using adaptive-regression modelling for describing the functioning of technical object. *Izvestiya of the Samara Russian Academy of Sciences scientific center* 2014; 16(6-2): 572–575.
- [12] Rodionova TE, Rybkina MV. Using mathematical modeling for the analysis of the social sphere influence on the quality of life of the population (on the example of the Ulyanovsk region). *Economic analysis: theory and practice* 2014; 32(383): 61–66.
- [13] Rodionova TE, Rybkina MV, Ananeva NA. Research of inflation impact on socio-economic factors. *Quality. Innovation. Education* 2015; 9(124): 48–51.
- [14] Kadyrova GR. Software System of searching for optimal regression models of forecast . *Way of science* 2014; 7 (7): 10–11.
- [15] Kadyrova GR. The system of searching for the optimal model. *State of affairs and development prospects. Modern science potential* 2015; 4(12): 8–10.
- [16] Rodionova TE, Klyachkin VN. Statistical methods of estimation the drinking water quality. *Reports of the Academy of Sciences of the Russian Federation* 2014; 2-3(23-24): 101–110.
- [17] Kadyrova GR. Possibilities of a software regression modeling system for estimating a model and searching for its optimal structure. *Radioelectronic engineering* 2015; 2(8): 228–233.
- [18] Kadyrova GR. Formation of strategies for finding optimal regressions // *Modern problems of design, production and operation of radiotechnical systems* 2016; 1(10): 178–180.
- [19] Kadyrova GR. Research of the quality measures of models for assessing the state of a technical object. *Synthesis, analysis and diagnostics of electronic circuits* 2016; 13: 71–83.

# The analysis of technical object functioning stability as per the criterion of monitored parameters multivariate dispersion

V.N. Klyachkin<sup>1</sup>, I.N. Karpunina<sup>2</sup>

<sup>1</sup>Ulyanovsk State Technical University, 432027, Ulyanovsk, Russia

<sup>2</sup>Ulyanovsk Civil Aviation Institute, 432071, Ulyanovsk, Russia

---

## Abstract

The assessment of any technical object functioning stability is often limited by the monitoring of midrange constancy and monitored parameters dispersion. For that, the methods of multivariate statistical monitoring, used for the assessment of process stability, are offered. Midrange multivariate process monitoring is accomplished with the help of algorithms, based on Hotelling's chart statistics. While assessing the dispersion stability, one can use generalized variance based algorithms - covariance matrix determinant. The approaches described here to increase the efficiency of multivariate dispersion monitoring.

*Keywords:* multivariate statistical monitoring; generalized variance; specialized structures; exponentially weighted moving average

---

## 1. Introduction

Technical object functioning stability often testifies to its serviceability. Destabilization may immediately lead to a failure or emergency situation [1]. The fault fastest detection is our main task. For example, the hydraulic unit vibration monitoring is done with the help of the chain of detectors [2]. The reading of these detectors indicates the stability or instability of the monitored hydraulic unit operation. In water purifying system potable water physical -chemical properties are monitored (color index, chlorides and aluminum content, etc) [3]: it is vitally important to keep the properties within the limits.

Destabilization appears as the alternation of statistical midrange characteristics and monitored parameters dispersion, so to detect the fault, process statistical monitoring methods and algorithms could be used [4-6]. The most frequent destabilization features, connected with midrange changes, are either step -wise displacement or trend i.e. gradual midrange decrease or increase. To detect this type of destabilization, monitoring a single parameter, Shewart's charts for midrange values and individual observation are used. To monitor multiple correlated parameters, algorithms based on Hotelling's chart statistics are used. To increase the efficiency of Hotelling's chart, there are several methods offered. [7]. One of them is finding specialized structures in the chart, probability of which is commensurate with the probability of false warning: trends, dramatic changes, events of approaching the control lines or abscissa. One more approach is the use of alert control line: several points in a row between the control lines show the availability of a midrange destabilization.

Similar methods could be used to find destabilization in investigating multivariate dispersion of object function parameters. The main fault types detected in object operation as per dispersion criterion are step-wise or gradual increase in monitored parameters dispersion. Monitoring one parameter the dispersion is characterized by a swing, standard deviation or variance. The main feature of multivariate dispersion is generalized variance- covariance matrix determinant.[8,9]. Sometimes effective variance is used [10].

There is a method of object operation stability analysis as per multivariate dispersion criteria, including the analysis of the detectors reading under the conditions of steady (flawless) object operation, covariance matrix assessment; the selection of possible statistical tools for the future dispersion monitoring; the assessment of average run length for various statistical tools, taking into account all possible deviations; statistical tests; minimum run length tools selection; constant monitoring of object operation with the goal of multivariate dispersion stability diagnostics. The up-dated information technologies and modern software products enable fast diagnostics of object operation fault with the help of the developed algorithms.

## 2. Generalized variance based algorithm for monitoring multivariate dispersion

To verify the hypothesis about the equality of covariance matrix  $\Sigma$  to selected value  $\Sigma_0$ , generalized variance, i.e. covariance matrix determinant, could be used [4,8]. For each time moment  $t$  selected covariance matrix  $S_t$  is generated, the elements of which are as follows:

$$s_{jkt} = \frac{1}{n-1} \sum (x_{ijt} - \bar{x}_j)(x_{ikt} - \bar{x}_k), \quad (1)$$

$x_{ijt}$  is the result of  $i$  - observation as per  $j$ -exponent in  $t$ -sample ( $i = 1, \dots, n$ ,  $n$  - sample size,  $j, k = 1, \dots, p$ ,  $p$  - number of monitored parameters,  $t = 1, \dots, m$ ,  $m$  - number of samples, taken to analyze the process as per learning sample). The matrix determinant (1)  $|S_t|$  is generalized variance of  $t$  instantaneous sample.

The assessment of average covariance is computed as per the collection of samples too.

$$\bar{s}_{jk} = \frac{1}{m} \sum_{t=1}^m s_{jkt}, \quad (2)$$

which make covariance matrix  $S$ ; its determinant  $|S|$  is used as the assessment of destination generalized variance  $|\Sigma_0|$ . While plotting the control chart the selected values of the generalized variance  $|S_t|$  for each  $t$ -sample are singled out on it.

The control lines of the generalized variance chart are determined from the ratios:

$$\left. \begin{array}{l} UCL \\ LCL \end{array} \right\} = |\Sigma_0| (b_1 \pm u_{1-\alpha/2} \sqrt{b_2}), \quad (3)$$

where  $u_{1-\alpha/2}$  is normal inverted distribution of order  $1 - \alpha/2$ ,  $\alpha$  is a confidence level (probability of false alert); the coefficients are computed as per the following formulae :

$$b_1 = \frac{1}{(n-1)^p} \prod_{j=1}^p (n-j); \quad (4)$$

$$b_2 = \frac{1}{(n-1)^{2p}} \prod_{j=1}^p (n-j) \left[ \prod_{k=1}^p (n-k+2) - \prod_{k=1}^p (n-k) \right], \quad (5)$$

the assessment of destination generalized variance  $|\Sigma_0|$  is found as per the learning sample .If the lower control line  $LCL$  as per formula (3) is negative, zero value is taken.

Destabilization of the process is witnessed by at least one point getting beyond one of the control lines on the chart of the generalized variance , i.e. the process is steady when the in equation below is satisfied:

$$LCL < |S_t| < UCL, \quad (6)$$

where  $t$  is the number of monitored samples. For example, Fig.1 shows the chart of generalized variance: lower control line is zero, no points beyond the control line: the process is steady.

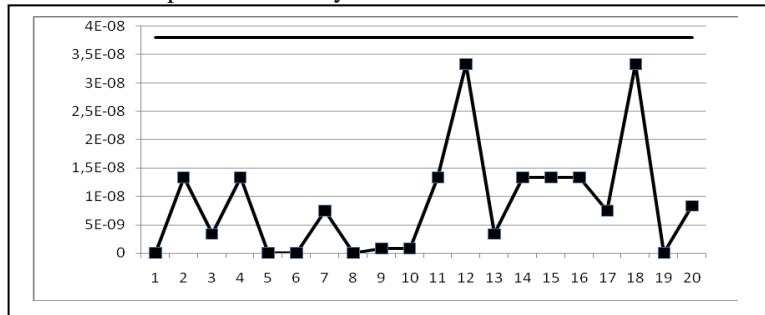


Fig.1. Generalized variance chart.

### 3. Methods to improve efficiency of faults detection as per multivariate dispersion

#### 3.1. Searching the structures of special form

The process is considered steady as per criteria of multivariate dispersion if on the chart of generalized variance there are no points beyond the control lines, i.e. the condition is followed (6). This condition is important, but very often insufficient to ensure the process stability. Sometimes on the chart there are special form structures, which testify process instability: these are the structures, the probability of which is commensurate with the probability of false alert. For example, several successive points increasing or decreasing indicate the trend of process monitored parameter. The specialists have no unanimous opinion regarding the structures to be used for stability assessment. Western Electric [4,5] four criteria are widely popular; one of them for example is as follows: at least eight successive points located on one side of the central line show the process instability. ISO distinguish eight criteria [6], six criteria are offered for Hotelling's chart [7].

Generalized variance algorithm is based on normal inverted distribution, and as a rule, practical calculations are done on the basis of three sigma rule: in formula (3) we take  $u_{1-\alpha/2} = 3$ . To find the fault one can use the same specialized structures types as for Schewart's chart. They are: 1) at least one point getting out beyond the control lines, 2) at least two out of three points located on one side of the central line, getting out beyond the twin sigma limits 3) at least four out of five successive points located on one side of the central line, getting out beyond one sigma limit 4) at least eight successive points located on one side of the central line , 5) six decreasing or increasing points in a row (trend), 6) fourteen in turn increasing and decreasing points (cycles) etc.

The probability of eight points in a row on one side of the central line may be detected as follows. Firstly, we check the criterion fault (6) i.e. a point getting out beyond one of the control lines, probability of this event while using tree sigma rule is equal to  $0,0027/2 = 0,00135$ . The probability of one point getting to one side of the central line is equal to 0,5. Then the probability of eight points on one side of the central line provided all the points are located within the control lines is equal to  $(0,5 - 0,00135)^8 = 0,003823$ , this is commensurate with the probability of false alert 0,0027.

Plotting the control charts on PC, the search of specialized structures of any type on the charts is easily computerized, without any difficulties. But, take into account that the increase of criteria number will lead to decrease of observations number among false alerts. Using only structure 1 to detect the unsteady state this number is equal to  $1/\alpha \approx 370$  samples, structures 1 and 4 get 153 samples selecting four structures from the 1<sup>st</sup> to 4<sup>th</sup> lead to 92 samples. This value is acceptable, but the use of additional criteria may bring the number of observations among the false alerts to unacceptably small value.

#### 3.2. Exponentially weighted moving average chart for generalized variance

To detect the step-wise increase in the dispersion exponentially weighted moving average algorithm for generalized variance is seldom used; the corresponding values are determined as per formula

$$E_t = (1 - k) E_{t-1} + k|S_t|; \quad (7)$$

where  $0 \leq k \leq 1$  is a smoothening parameter,  $E_0 = |\Sigma_0|$ . The process is considered steady if the found values are within the control lines

$$\left. \begin{array}{l} UCL \\ LCL \end{array} \right\} = |\Sigma_0| \pm H\sigma_{E_t}, \quad (8)$$

where  $H$  is a parameter, which determines the location of the control lines;  $\sigma_{E_t}$  is a mean square deviation of  $E_t$  values, determined as per formula:

$$\sigma_{E_t}^2 = \frac{\sigma_{|S|}^2}{n} \frac{k}{2-k} [1 - (1-k)^{2t}]. \quad (9)$$

Fig. 2 shows the chart of exponentially weighted moving average for generalized variance, plotted with the same data as in Fig.1 chart. It is seen, that in observations 18-19 there is a fault in the process (intentionally simulated little dispersion), which was not found by the chart of generalized variance.

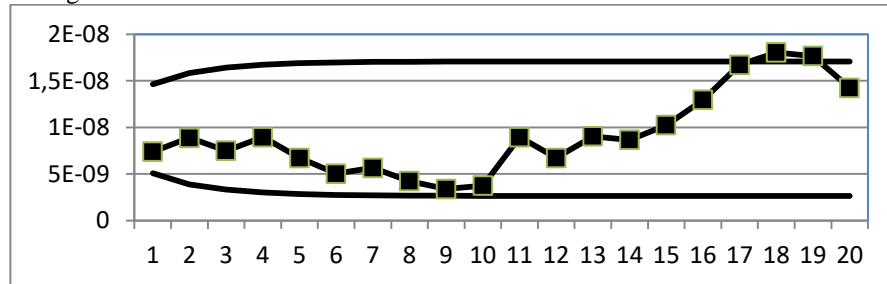


Fig.2. The chart of exponentially weighted moving average for generalized variance.

### 3.3. The offered way to assess the object functioning stability

The conducted investigation made it possible for us to offer the following way of object operation stability assessment:

1. Under the conditions of flawless stable operation the detectors readings are taken and main statistical characteristics are calculated: mean values vector and covariant matrix (characteristics of learning sample).
2. A set of all possible statistical tools is selected for further monitoring. Non correlated data are monitored by the tools based on Schewart's chart. To monitor mean level of correlated parameters, Hotelling's chart is used, to monitor multivariate dispersion, generalized variance chart is used.
3. When necessary the exponentially weighted moving average algorithm based on Hotelling statistics and generalized variance is used.
4. Constant monitoring of object operation is done in order to detect destabilization. Specialized structures are searched on the charts, which prove the possible fault in the process.

## 4. Results and discussion

The computational investigation was done based on the example of hydraulic unit serviceability with the use of vibration dispersion stability criterion [12]. The detectors readings were correlated; simulated dispersion increase was captured by the exponentially weighted moving averages chart (Fig.2)

To assess the stability of object functioning stability as per the criteria of multivariate dispersion one may use the control charts of generalized variance. But these charts do not always detect the faults on time. There were offered methods for the charts sensitivity improvement: the search of non- random structures and the use of algorithm of exponentially weighted moving averages can significantly increase the monitoring efficiency.

## 5. Conclusion

The offered way of object functioning destabilization diagnostics, based on the process statistical control methods, enables timely detection of the operation faults, connected with its parameters dispersion alternation, and prevents an emergency when necessary.

## Acknowledgements

The investigation is done with financial support of RFFI (РФФИ), projects №16-48-732002.

## References

- [1] KlyachkinVN, Karpunina IN. The use of statistical control methods to assess units operation stability. Reports of RF Higher School Academy of Science 2016; 3: 65-72.

- [2] Klyachkin VN, Kuvaiskova YuYe, Aleshina AA. Simulating of hydraulic unit vibration on the basis of adaptive dynamic regression. *Computerizing. Modern technologies* 2014; 1: 30–34.
- [3] Kuvaiskova YuYe, Bulyzhev YeM, Klyachkin VN, Bubyr DS. Predicting the water supply source status in order to ensure water quality. Reference book. *Engineering Journal with attachments* 2016; 5: 37–42.
- [4] Montgomery DC. *Introduction to statistical quality control*. New York: John Wiley and Sons, 2009; 754 p.
- [5] Ryan TP. *Statistical methods for quality improvement*. New York: John Wiley and Sons, 2011; 687 p.
- [6] Klyachkin VN. *Models and methods of statistical control of polyvalent process*. M.: Fizmatlit, 2011; 196 p.
- [7] Klyachkin VN, Kravtsov YuA. The detection of faults during process multivariate statistical monitoring. *Software and systems* 2016; 3: 192–197.
- [8] Svyatova TI, Klyachkin VN. Multivariate statistical monitoring of dispersion process. *Radio technology* 2014; 11: 123–126.
- [9] Klyachkin VN, Svyatova TI. Methods of statistical monitoring of process as per criteria of multivariate dispersion. *Radio industry* 2015; 4: 147–153.
- [10] García-Díaz Carlos J. The ‘effective variance’ control chart for monitoring the dispersion process with missing data. *Industrial Engineering* 2007; 1(1): 40–45.

# The use of aggregate classifiers in technical diagnostics, based on machine learning

V.N. Klyachkin<sup>1</sup>, Yu.E. Kuvayskova<sup>1</sup>, D.A. Zhukov<sup>1</sup>

<sup>1</sup>*Ulyanovsk State Technical University, 432027, Ulyanovsk, Russia*

---

## Abstract

While solving the problems of technical diagnostics with machinery learning involvement, there is binary classification of an object state performed: the objects are subdivided into “good” and “bad” with the help of models, received as per learning samples. The quality of classification, which specifies the efficiency of machine learning, depends on several factors, such as: the scope of original sample, method of machine learning, method of dividing the sample into learning and validating parts, selection of value indicators, etc. Sometimes it is reasonable to use aggregate methods of classification, which are, in fact, the joined results of classification basic methods. To search the best aggregate method, one iterates over all possible basis sets.

*Keywords:* binary classification; “good” and “bad” state; aggregate method

---

## 1. Introduction

The object technical diagnostics is done to increase the reliability of the system, and it is often limited by the assessment of its serviceability [1,2]. The main aim is object state recognition. By recognition we mean an object state, typed as per one of the classes - diagnoses. As a rule, the problem solution is restricted by classifying the object as good, i.e. able to perform the desired functions, or bad. This is the task of binary classification. The selection of parameters, characterizing the system state, is important. The assessment of the system state is done during its operation, the information receiving is somehow difficult, to take a decision - different methods of recognition are used. The solution of technical diagnostics problems is also connected with the technical object state forecasting [3,4].

The recognition of the object technical state is usually done, based on the results of indirect indicators of the object operation under the conditions of available limited information. The known results of the system state assessment are used: for the selected values of monitored parameters the system is assessed as “good” or “bad” (serviceable or unserviceable). So, there are many objects (situations) with selected indicators, and many possible states of the system. There is ascertain unknown dependence between the object operation indicators and its actual state. The finite universe of pair - “set of indicators, state”- is known, i.e. the original data retrieving. It is required to store the dependence, i.e. to build an algorithm, capable to generate a rather accurate response. Anyway there is a risk of getting false warning or missing the target. This is the task of machine learning, or learning from examples (with a tutor) [5,6].

To measure the accuracy of classification there is performance functional introduced, e.g. the mean error can be used: original sample is divided into learning one, with the help of which the algorithm of the needed dependence is identified, and validating one (test), with the help of which the mean error is assessed [7,8].

## 2. Methods of machine learning, used for binary classification

The methods of machine learning are widely used in different fields: speech recognition, medical diagnose, loan score and others. From the point of view of technical diagnostics, the machine learning is only binary classification task: as per the selected vector of object parameters, it is necessary to determine, which state of the object is (“bad” or “good”).

For solving the problems of technical diagnose, the method of learning from the examples is used. Bayesian classifier is one of them, as well as K-Nearest Neighbors (KNN) method, neural network, logistic regression, discriminate mining, Support Vector machine method, decision trees, etc.

The problem of technical object classification is solved as follows: the object is found “good”  $Y = 1$ , if the probability model is  $P\{Y = 1 | X\} > 0,5$ , and “bad”  $Y = 0$  – if opposite. As threshold any number different from 0,5 can be used.

For example:

Using logistic regression, we assume that the probability of “good” object is equal to  $Y = 1$ :

$$P\{Y = 1 | X\} = f(z),$$

$$z = q_0 + q_1 x_1 + \dots + q_n x_n, \tag{1}$$

where  $q_0, \dots, q_n$  is model parameters (1),  $f(z)$  is logistic function:

$$f(z) = \frac{1}{1 + e^{-z}}. \tag{2}$$

As variable  $Y$  takes one of pair values (0,1), the probability of “bad” state is equal to  $Y = 0$ :

$$P\{Y = 0 | X\} = 1 - f(z). \tag{3}$$

So, the logistic regression is based on the following expression:

$$\log \frac{P\{Y = 1 | X\}}{P\{Y = 0 | X\}} = \frac{f(z)}{1 - f(z)} = q_0 + q_1 x_1 + \dots + q_n x_n. \tag{4}$$

To identify parameters  $q_0, \dots, q_n$ , as a rule, maximum likelihood method is applied. This method is aimed at likelihood function maximization with the help of gradient descent method, Newton Raphson method and others.

With the time, while new data are received, the earlier found parameters can become out-dated. To update them, different procedures can be used, e.g. those based on pseudogradient use [9].

The short coming of the logistic model is its sensitivity to factors correlation, that is why, the presence of strongly correlated input variables is unacceptable in the model.

The advantage of them model is the possibility to take in to consideration the limitations of probability value, which cannot be out of frame 0 and 1, the possibility of conducting investigation and assessment of the factors, affecting the result.

While using another widely applied model – discriminant mining – to determine the object  $m$  class, linear discriminant functions are used:

$$\begin{aligned} o_1(x) &= q_0^1 + q_1^1 x_1 + \dots + q_n^1 x_n, \\ o_2(x) &= q_0^2 + q_1^2 x_1 + \dots + q_n^2 x_n, \\ &\dots \\ o_m(x) &= q_0^m + q_1^m x_1 + \dots + q_n^m x_n, \end{aligned} \quad (5)$$

Where  $o(x)$  is «counting», as per which this or that class is identified. In result, that class is chosen, which counting is the highest. The model parameters are assessed with the help of learning sample. In case of two classes, there coincides with the result of linear regression.

The point is, that, in advance, one cannot say, which method, out of those mentioned above, will ensure the correct solution of the problem, that is why several methods or combination of methods are used. The decision is taken, based on the results of the performance functional for the validation set.

### 3. Aggregate methods

Aggregating methods (combined application of several methods) is of special interest, indeed, as this way compensates the disadvantages of one model with the help of the others, thus improving the forecasted accuracy. Now we will consider the follow in gset, let us say, of base 7 models [10,11]: neural network, logistic regression, discriminate mining, Bayesian classifier, Support Vector machine method, decision trees, bagging trees etc.

Let's make all possible combinations of base models, consisting of two, three... etc. models. In case of seven models taken, the total amount of all the combinations, starting from two and finishing with all seven models, will make:  $C_7^2 + C_7^3 + C_7^4 + C_7^5 + C_7^6 + C_7^7 = 120$  models.

Let  $Y_j^m$  be the result of serviceability assessment of  $j$ -object. The result was determined with  $m$  base model,  $j = 1, \dots, l$  and  $m = 1, \dots, M$ , where  $M$  – the number of base models in combination. Now let us see the following ways of base models aggregating (joining).

#### a. Aggregating, based on mean value

In this case

$$Y_j^{AK\_mean} = \frac{\sum_{m=1}^M Y_j^m}{M}, \quad (6)$$

where  $Y_j^{AK\_mean}$  is the result of aggregate classifier based on mean value.

#### b. Aggregation as per median value

Firstly, we will grade the row, containing the results of the base models in combination  $Y_j^m$ . If the number of base models is odd:

$$Y_j^{AK\_median} = Y_j^{\frac{M+1}{2}}, \quad (7)$$

where  $Y_j^{AK\_median}$  is the result of aggregate classifier as per median value.

#### c. Aggregation as per voting

This works as follows: if the majority of the models consider the object is “good”, then the result of aggregate classifier is a mean value of the results of the models, voting for the serviceable class. In opposite case, the object is “bad” (unserviceable) ( $Y=0$ ).

#### 4. Diagnose quality assessment

The accuracy of classification is assessed with the help of performance functional, i.e. validation set classification mean error can be used.

When the results are presented in the terms of object “good” or “bad” class probability, to assess the methods quality it is possible to find error dispersion  $\sigma^2$ , which indicates the deviation of forecasted value from actual ones:

$$\sigma^2 = \frac{1}{l} \sum_{r=1}^l (P(Y_r) - \hat{P}(X_r))^2, \quad (8)$$

where  $P(Y_r)$  is actual probability of serviceability class of  $r$  object ( $P(Y_r) = 0$ , if the object is “bad” or  $P(Y_r) = 1$  for the “good” object),  $\hat{P}(X_r)$  is actual probability of serviceability class of  $r$  object,  $l$  is number of objects.

The performance functional mainly depends on the way of validation set constructing. If necessary, check the influence of the way of validation set constructing on the error dispersion.

In order to optimize the diagnostic of technical object functioning, the algorithm of serviceability forecasting is offered. This is the use of machine learning models combination and generating optimum decision on their basis.

The algorithm main stages:

1. Generating and preliminary processing of the original data, dividing them into learning and validation set.
2. Constructing the base classification model on a learning set.
3. Building all the possible models combination on the same learning set in addition to all base models with three mentioned above methods aggregation.
4. On the validation set, through all the constructed models, new (monitored) objects serviceability is forecasted.
5. For each model or model combination, the forecasted mean square root error is calculated, and the best model, providing the error minimum, is selected.

#### 5. Results and discussion

The computational investigation was done based on the example of St. Petersburg sewage plant operation [12]. The parameters of the water supply source and the dosage of chemical agent, used for purifying, were monitored. At least one parameter of potable water quality, found out beyond the acceptable limitations, was considered to be the system malfunction. Seven base methods of binary classification were used, and the best result was shown by the method of support vector machines (SVM); mean classification error was equal to 0,238. Mean value aggregation had an error equal to 0,196 (combination of SVM with discriminate mining and neural network). While using the selection of tangible value parameters through the method of stepwise regression, the results deteriorated a little, but even that, minimum mean error of aggregation was 0,207. The set of base classification changed: the logistic regression method was added to those three available.

The performance quality of classification is determined by the scope of the original sample, selected by machine learning method (one of base or aggregate), by the method of dividing the original sample into learning and validation one (either by random selection, or by taking a certain part of original sample for validation one; sometimes the procedure of sliding exam is reasonable, evidently, the scope of validation sample plays a certain role hereto), method of tangible value (e.g. stepwise regression) and some other factors.

To provide the efficiency of machine learning for the technical object diagnostic, it is necessary to work out the system in order to investigate the influence of these factors on the performance quality of classification with original sample, which could ensure the optimum approaches.

#### 6. Conclusion

The given above investigation showed, that the methods of machine learning could be used for solving the problems of technical diagnostics, i. e. identifying the state of serviceability of the investigated object. Here some problems may arise. The problems are those connected with generating a rather big scope original sample, with dividing the sample into learning and validating, with assessment of this or that method efficiency, with possibility to use aggregate classifiers.

#### Acknowledgements

The investigation is done with financial support of RFFI (РФФИ), projects №16-48-732002 and №16-38-00211 mol\_a.

#### References

- [1] Birger IA. Technical diagnostics. M.: Mechanical engineering, 1978; 240 p.
- [2] Zhukov DA, Klyachkin VN. The tasks of machine learning efficient cyprovision for technical objects diagnostics. Modern problems of radio aids design, industry and operation 2016; 1(10): 172–174.
- [3] Klyachkin VN, Bubyр DS. Forecasting of technical object state based on piecewise linear regression. Radiotechnics 2014; 7: 137–140.
- [4] Klyachkin VN, Kuvayskova YuE, Bubyр DS. Forecasting of technical object state with the use of time series system. Radiotechnics 2015; 6: 45–47.
- [5] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005; 525 p.
- [6] Merkov AB. Patterns recognition. Introduction to the methods of statistical learning. M.: Editorial URSU, 2011; 256 p.
- [7] Klyachkin VN, Kuvayskova YuE, Alekseeva VA. Statistical methods of data mining. M.: Finance and Statistics, 2016; 240 p.
- [8] Klyachkin VN, Karpunina I.N, Kuvayskova YuE, Khoreva FR. Machine learning method in technical diagnostics. Academic Bulletin of UCAS 2016; 8: 158–161.



- [9] Krashennnikov VR, Klyachkin VN, Shunina YuR. Updating of aggregate classifiers on the basis of pseudogradient procedure. Academic Bulletin of computer and information technologies 2016; 10(148): 36–40.
- [10] Shunina YuR, Klyachkin VN. Forecasting of bank customers creditability on the basis of machine learning methods and Markov chains. Software products and Systems 2016; 2: 105–112.
- [11] Shunina YuR, Alekseeva VA, Klyachkin VN. Forecasting of bank customers creditability on the basis of machine learning method. Finances and Credit 2015; 27(651): 2–12.
- [12] Kuvayskova YuE, Bulzhev YeM, Klyachkin VN, Buby DS. Forecasting of water supply source state in order to ensure water quality. Reference manual. Engineering pamphlet with attachment 2016; 5: 37–42.

# Investigation of the genetic algorithm possibilities for retrieving relevant cases from big data in the decision support systems

K. Serdyukov<sup>1</sup>, T. Avdeenko<sup>1</sup>

<sup>1</sup>Novosibirsk State Technical University, Prospekt K. Marksa 20, 630073, Novosibirsk, Russia

---

## Abstract

In present paper we consider the advantages and disadvantages of case-based reasoning (CBR) approach for knowledge representation of the application domain. One of the CBR shortcomings is the insufficient speed of real-time retrieval of cases, as well as the insufficient relevance of the retrieved cases to the current situation. To solve these problems, we offer the use of genetic algorithm. We propose formal statement of the genetic algorithm to the CBR retrieving stage. The results of the investigation are presented. The major advantage of the genetic algorithm is that it gives a more compact set of retrieved cases extracted which possesses, however, characteristic features of the current situation. This property can be very useful when extracting such cases from big data. In conclusion, the perspectives of applying the method for adaptation of cases have been given.

*Keywords:* decision support systems; case-based reasoning; genetic algorithm; data mining; big data

---

## 1. Introduction

The emergence of Decision Support Systems (DSS) in the mid-1960s was associated with a model-oriented approach, popular at the time. The first DSS were limited to the types of models implemented in them, mostly deterministic, that practically did not use operative information about the circumstances in which decisions had to be made. In addition, solutions produced as a result of the operation of such systems were based on deterministic optimization models and were not always understandable and explainable from the point of view of the decision-maker, which significantly hampered their practical application.

The situation has dramatically changed since the 1990s, when DSS began to integrate first with operational databases, and then with specialized warehouses built using OLAP (On-line Analytical Processing) technology. There appeared an opportunity of operative decision-making on the basis of the objective information saved up in data warehouses. In the modern era of the Internet, when the situation in which decisions have to be made changes literally before our eyes, the DSS concept undergoes drastic changes. Information becomes so much (even the stable expression "big data" has appeared) that the data itself has ceased to be of great value. What is really valuable is knowledge, which can be extracted from the data to solve an actual problem in a particular problem domain. The task of obtaining such (informationally saturated) qualitative knowledge, and organizing them in a specially designed knowledge base, is now, in the era of large data, more relevant than ever before. It is knowledge in the form of human-understandable (cognitive) constructions, cleared of information garbage, that make it possible to organize decision support at a completely different qualitative level, when the decision-maker not only receives recommendations from the DSS, but also understands why in a particular situation it is necessary to make such a decision.

Methods of representation and organization of knowledge are traditionally developed within the Artificial Intelligence (AI) scientific discipline. In the process of development of this scientific field, two basic approaches to the declarative representation of knowledge were formed: rule-based and case-based. The emergence, in the 1970s, of expert systems based on knowledge, is associated with the approach to the representation of knowledge in the form of rules. It is with these systems that the first real commercial successes in the field of artificial intelligence are connected. By 1992 about two thousand expert systems based on rules were implemented [1].

However, despite the success, even at the very beginning of development of systems based on rules, their shortcomings became obvious. The main problem was the problem of acquisition of knowledge from sources of information and presenting them in the form of rules. Most often, experts intuitively make decisions based on their extensive experience, without thinking, what kind of rule they apply in this or that case. Splitting the expert's specific behavior into separate blocks, called rules, is the key problem (bottleneck) in the development of rule-based systems. Another problem is the discrepancy between the real complexity of the problem domain and the very simple rule structure in the early expert systems. At present, this problem is partially solved by introducing an object-oriented description of the rule parts.

On the other hand, since the 1980s, the alternative paradigm for presenting knowledge and reasoning has attracted more and more adherents. Case based reasoning (CBR) allows solving new problems by adapting the experience of solving similar problems in the past, just as a person does in real conditions. In the paper [2] the foundations of this method are given, it is suggested to generalize knowledge about past cases and save them in the form of scenarios that can be used to develop solutions in similar situations. Later, Schank [3] continued to investigate the role played by the memory of previous situations (precedents) presented in the form of a certain knowledge container, in decision making and in the learning process.

At present, in the studies on AI, CBR is one of the key directions that is rapidly developing. The following generally accepted definition of a case could be given: "a case is a description of the problem or situation in conjunction with a detailed description of actions taken in a given situation for solving current problem". Thus, the case as a unit of knowledge includes the following:

- description of the situation;

- the decision that was made in this situation;
- the result of applying the solution.

There are various ways of presenting cases - from simple (linear representation) to more complex hierarchical representations. The case generally includes a description of the problem, as well as a solution to the problem. In case the cases from the knowledge base were used to solve specific practical problems, an additional component in the description of the case may be the result (or forecast) of the case use (positive or negative). It is interesting to note that in [4], which is often referred to as a philosophical basis of the precedent approach, it is noted that natural concepts of the application domain can often not be described by a simple linear set of properties (features), but require more complex structures for their description.

CBR-approach to the knowledge representation allowed to overcome a number of limitations inherent to the systems based on rules [5]. It does not require an explicit model for representing knowledge of the application domain, so the complex problem of knowledge acquisition is transformed into the task of accumulating cases of decision making. The implementation of the system is reduced to identifying the significant features of the case and the subsequent description of the decision-making cases in accordance with these features, which, of course, is much simpler task than building an explicit knowledge model of the application domain.

By now, the following advantages of CBR have become apparent:

- The ability to use the accumulated experience directly, without the direct involvement of a specialist who proposed a solution to a similar problem, from the case base;
- Reduction of the time for the development and making a new decision due to the available experience in solving similar problems;
- For very similar problems, the probability of making an erroneous decision is reduced;
- You can use well-developed database technology to store large volumes of cases;
- It seems promising to apply machine learning methods to the CBR-systems to the extracting knowledge in an explicit form, as well as to expanding the case base.

At the same time, there are fundamental limitations to the traditional CBR. First, when describing cases, specialists are often limited only to general knowledge or description of the problem, without deepening into the process of deriving a decision and confining themselves only to the results. Thus, the structure of the decision-making cases in this problem area does not correspond to its complexity. Secondly, as the knowledge base accumulates, the number of cases grows, which negatively affects the performance of the DSS and, accordingly, the quality of the decision made. Based on these shortcomings, it is possible to highlight the most actual requirements for the CBR-system design:

- The need for clear indexing and organization of systems for cases comparison ;
- Requirement for the selection of relevant precedents, and not just similar ones based on the closeness concept;
- Interpretability of the retrieved cases in relation to the specific problem to be solved;
- Formulation of the a solution even if there are no similar cases in the knowledge base.

Reasoning by the analogy based on CBR consists in solving the current problem in accordance with the following four steps forming the so-called CBR-cycle, or the 4R-cycle (Retrieve, Reuse, Revise, Retain) shown in fig. 1. The main stages of the CBR-cycle are:

**Retrieve** the most appropriate or similar case (a subset of cases) from the case base (knowledge base);

**Reuse** the retrieved cases to solve the current problem;

**Revise** (or adapt) cases, if necessary, to obtain a more specific and accurate solution;

**Retain** the solution in the knowledge base as a new case for its further use.

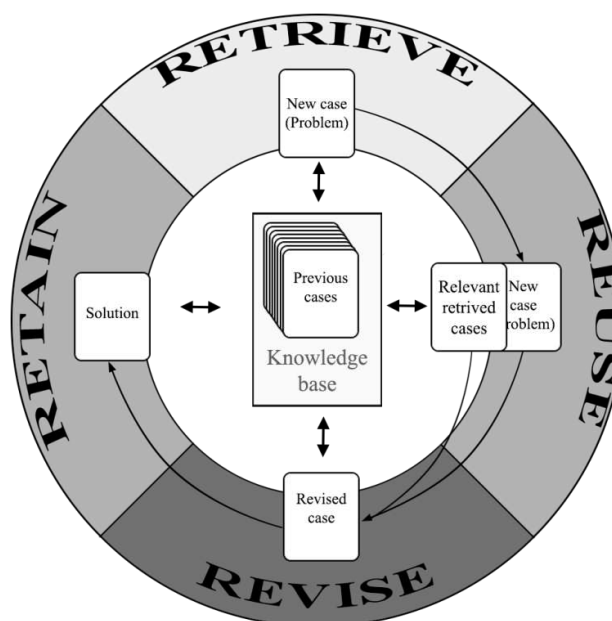


Fig. 1. CBR-cycle.

The first and most investigated stage of the CBR-cycle is to retrieve cases. The main problem in retrieving cases is a choice of the method by which the similarity measure is calculated. Often, the closest neighbor method is used for this purpose, which is based on measuring the degree of coincidence of the feature values determining the case. In papers [6-9], measures based on the introduction of the weight function, taking into account the significance of each of the features forming the case, have been proposed. However, despite numerous studies in this field, there remain problems of insufficient relevance of the retrieved cases, as well as insufficient speed of their extraction. In addition, the problem of adapting the retrieved cases to the real conditions in which decisions are made is still very far from any acceptable solution. It seems to us promising to use evolutionary approaches, in particular, the genetic algorithm, both from the point of increasing the speed of retrieving cases and from the point of view of relevance of the extracted cases to the current problem. An interesting area of research is seems the adaptation of the retrieved case to the current situation through evolutionary development.

In this paper, we consider an approach to solving the problem of retrieving and adapting cases based on the genetic algorithm. Section 2 provides a formal problem statement and scheme of the genetic algorithm for solving the problem of retrieving cases. Section 3 shows the results of research of the implemented algorithms for the two data samples. In section 4 we formulate conclusions on the work and propose perspectives for further research.

## 2. Problem statement in terms of the genetic algorithm

Suppose that in the DSS we have the knowledge base for decision support consisting of  $n$  cases  $Case_i, i = \overline{1, n}$ . Let us set the task of retrieving a subset of cases  $Retrieved = \{Case_{i_1}, Case_{i_2}, \dots, Case_{i_m}\}, i_k \in \{1, \dots, n\}$ , that best fit the current problem  $Target$ , determined by a set of features  $Target^j, j = \overline{1, m}$ . Note that each case  $Case_i$  is determined by a set of features some of which  $Case_i^j, j = \overline{1, m}$ , exactly corresponds to the characteristics of the target problem.

The genetic algorithm solves the problems of searching in complex decision spaces on the basis of evolutionary principles [10]. We formulate the task of retrieving cases in terms of a genetic algorithm as follows. Suppose that the population of individuals contains  $P$  chromosomes  $X_p, p = \overline{1, P}$ , each of which is a binary vector of the dimension  $n$ , consisting of genes encoding the presence or absence of an appropriate case  $Case_{i_k}$  in a subset of the retrieved cases  $Retrieved$ :

$$X_p = [X_p^1, X_p^2, \dots, X_p^n]^T,$$

where  $X_p^i = \begin{cases} 1, & \text{if chromosome } X_p \text{ corresponds to retrieving the case, } Case_i \in Retrieved \\ 0, & \text{if chromosome } X_p \text{ corresponds not to retrieving the case, } Case_i \notin Retrieved \end{cases}$

Thus, each chromosome corresponds to a certain subset of the retrieved cases and is characterized by a definite value of the generalized unfitness function  $UF(X_p)$ , that has the more value the less similar are the target problem and the subset of the retrieved cases in general:

$$UF(X_p) = \sum_{i=1}^n gap(Target, Case_i), \tag{1}$$

where  $gap(Target, Case_i)$  is the discrepancy between the target problem and the case  $Case_i$ , computed as a weighted sum of discrepancies by all features

$$gap(Target, Case_i) = \sum_{j=1}^m w_j * \delta(Target^j, Case_i^j), \tag{2}$$

where the weights  $w_j$  define the significance of the considered features.

The values discrepancies  $\delta(Target^j, Case_i^j)$  between separate features are calculated in various ways for categorical and quantitative characteristics. For categorical variables we have

$$\delta(Target^j, Case_i^j) = \begin{cases} 0, & \text{if the value of the } j - \text{th feature coincides for the target problem and the case} \\ 1, & \text{if the values of the } j - \text{th feature for the target problem and the case differ} \end{cases}$$

For the numerical features we have  $\delta(Target^j, Case_i^j) = \frac{|Target^j - Case_i^j|}{\max_i |Case_i^j| - \min_i |Case_i^j|}$ .

In fig. 2 we present composition of one population. One population consists of a set of chromosomes, which, in turn, consist of genes. Each gene is given a value of 0 if it is unfitted or 1 if it is fitted. Accordingly, the color of the cell will also depend on the value of the unfitness function. Thus, the more retrieving cases are there in the chromosome the better it is suitable for crossing.

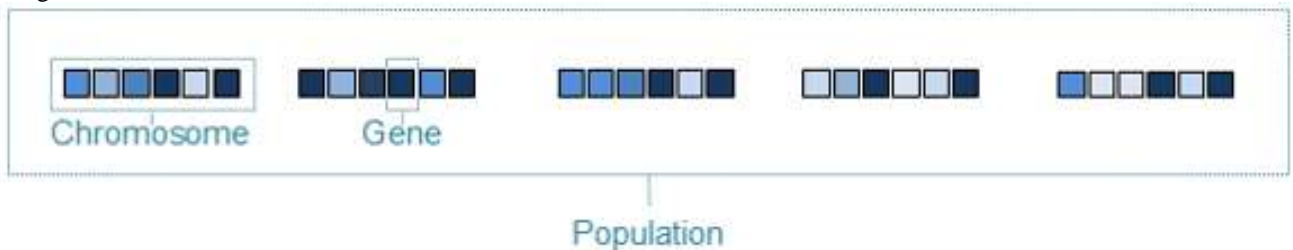


Fig. 2. Composition of one population.

Based on this formalization, the method sequentially implements three operations of the genetic algorithm: selection, crossover and mutation, as presented in fig. 3. The initial population is randomly generated. Further selection of individuals is performed on the basis of the unfitness function  $UF(X_p)$ . The lower is the chromosome unfitness function, the higher is its

reproductive capacity. We use a single-point crossover to cross chromosomes, and we also assume a mutation probability of 0.05.

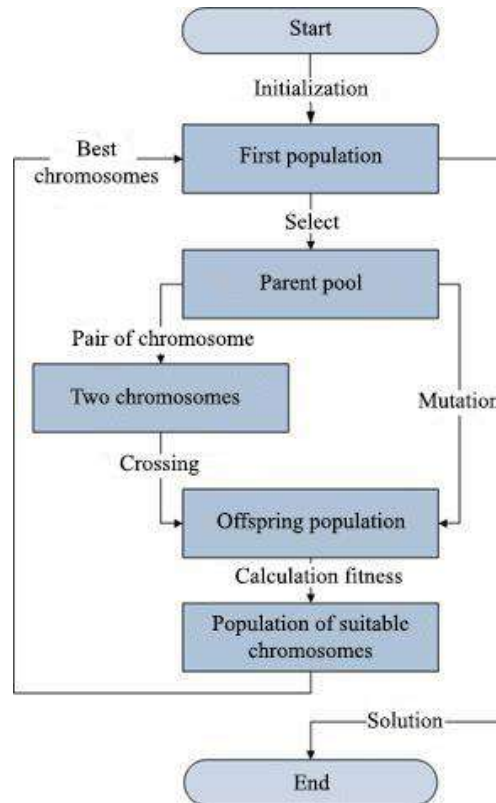


Fig. 3. Scheme of the genetic algorithm.

The next section presents the results of computations of the implemented algorithm on the test data.

### 3. Results of the genetic algorithm performance

#### 3.1. Investigation of the genetic algorithm for the numerical features

In this subsection we investigate the proposed approach with the data set Iris (available at <http://archive.ics.uci.edu/ml/>) with 150 cases, characterized by 4 numerical features ( $n=4$ ), which are classified into 3 classes (Iris Setosa, Iris Versicolour or Iris Virginica) through the classifying attribute. The features are as follows:

- Length of sepals;
- Width of sepals;
- Length of petals;
- Width of petals.

In our test we are interested in the retrieving cases similar to the request: length of the sepals 6.3 cm, width of the sepals 2.3 cm, length of the petals 4.4 cm, width of petals 1.3 cm. The flower with these parameters refers to the species Iris Versicolour. The genetic algorithm settings are as follows: population is 100 chromosomes, 10 generations, 50% crossover position and 5% mutation chance.

In table 1 we give the results obtained for three variants that differ from one another in the flexibility of the query. In the first test, we are interested in the exact match of the features of the request (current situation) and the retrieved cases (0% discrepancy). For numeric attributes, the probability of the exact coincidence of all four features is very small, so we get a small number of retrieved cases, but also a short computation time. In the second test, we allow 10% deviation of the features of the retrieved cases from the query, and in the second test, we allow 20% such deviation. In this case, we obtain a consistent increase in the number of retrieved cases, and accordingly increase of the computation time.

As for the quality of the retrieving, we can judge it by the quality of the classification of the retrieved cases. As you can see, the correct classification takes place in 83% of cases for 20% and 10% feature deviations, i.e. allowing a flexible query, we support the representativeness of the retrieved sample. If we set 0% feature deviation, the accuracy of the classification of the retrieved cases is reduced.

#### 3.2. Investigation of the genetic algorithm for big data

In this test we use Adult database from UCI [15]. The case base contains 32561 cases. Database contains the following features:

- Age (numeric feature);
- Workclass (categorical feature);
- Fnlwgt – final weight (numeric feature);
- Education – last education (categorical feature);
- Education-num – number of different education types (numeric feature);
- Marital-status (categorical feature);
- Occupation (categorical feature);
- Relationship (categorical feature);
- Race – ethnic group (categorical feature);
- Sex (categorical feature);
- Capital-gain – incomings (numeric feature);
- Capital-loss – expenses (numeric feature);
- Hours-per-week (numeric feature);
- Native-country (categorical feature);
- Annual income (numeric feature).

Table 1. Investigation of accuracy and speed of the genetic algorithm for Iris data.

No	Results of deviation								
	20% feature deviations			10% feature deviations			0% without deviation (exact)		
	Time (s)	The number	Class	Time (s)	The number	Class	Time (s)	The number	Class
1	0.56	3	Iris Versicolour	0.35	1	Iris Versicolour	0.05	1	Iris Versicolour
2	0.31	2	uncertainty	0.15	2	Uncertainty	0.09	1	Iris Versicolour
3	0.48	3	Iris Versicolour	0.23	1	Iris Versicolour	0.12	1	Iris Setosa
4	0.54	3	Iris Versicolour	0.1	1	Iris Versicolour	0.1	2	Iris Versicolour
5	0.12	1	Iris Versicolour	0.12	1	Iris Versicolour	0.12	2	uncertainty
6	0.22	1	Iris Versicolour	0.09	1	Iris Versicolour	0.09	1	Iris Versicolour
7	0.37	1	Iris Versicolour	0.17	2	Iris Versicolour	0.05	2	H/H
8	0.38	1	Iris Versicolour	0.26	1	Iris Versicolour	0.07	1	Iris Versicolour
9	0.43	1	Iris Versicolour	0.12	1	Iris Virginica	0.11	1	Iris Versicolour
10	0.31	3	uncertainty	0.13	1	Iris Versicolour	0.09	1	Iris Setosa
Av	0.372	1.9	Iris Versicolour (~83%)	0.172	1.2	Iris Versicolour (~83%)	0.089	1.3	Iris Versicolour (~70%)

We use only two of 14 features in the request (current situation) - numerical feature Age (equal to 30 years) and categorical feature Education (bachelor). As the classifying attribute we use annual income with two classes of more than and less than \$50 000 ( $\leq 50K$  or  $> 50K$ ).

In the first test we carried out the research similar to those made with Iris dataset. The genetic algorithm settings are as follows: population is 100 chromosomes, 10 generations, 50% crossover position and 5% mutation chance. The results are given in table 2 and are similar to those obtained in section 3.1.

Table 2. Investigation of accuracy and speed of the genetic algorithm for UCI data.

No	Results of deviation								
	20% feature deviations			10% feature deviations			0% without deviation (exact)		
	Time (s)	The number	Class	Time (s)	The number	Result	Time (s)	The number	Class
1	136	1849	$\leq 50K$	127	1253	$\leq 50K$	101	89	$\leq 50K$
2	121	1807	$\leq 50K$	114	1224	$\leq 50K$	95	85	$\leq 50K$
3	116	1858	$\leq 50K$	120	1226	$\leq 50K$	94	89	$\leq 50K$
4	115	1814	$\leq 50K$	138	1237	$\leq 50K$	102	90	$\leq 50K$
5	116	1803	$\leq 50K$	112	1212	$\leq 50K$	94	82	$\leq 50K$
6	137	1780	$\leq 50K$	134	1250	$\leq 50K$	89	86	$\leq 50K$
7	134	1857	$\leq 50K$	126	1261	$\leq 50K$	100	74	$\leq 50K$
8	135	1805	$\leq 50K$	125	1224	$\leq 50K$	89	100	$\leq 50K$
9	151	1804	$\leq 50K$	115	1227	$\leq 50K$	89	90	$\leq 50K$
10	146	1820	$\leq 50K$	108	1241	$\leq 50K$	97	85	$\leq 50K$
	130.7	1819.7	$\leq 50K$ (~65%)	121.9	1235.5	$\leq 50K$ (~70%)	95	87	$\leq 50K$ (~63%)

In the second test we compare genetic algorithm with conventional CBR. In the table 3 we give computation time, the number of retrieved cases and accuracy of classification for conventional CBR, and the genetic algorithm with 1,2 and 5 generations. The genetic algorithm settings are as follows: population is 10 chromosomes, 50% crossover position, 5% mutation chance and 5% numerical feature (age) deviation. As a result one can see that for the genetic algorithm we have obtained less amount of the retrieved cases with the same classification accuracy and insignificant increase in the computation time. This indicates a greater representativeness of a set of the retrieved cases when using the evolutionary approach for retrieving.

Table 3. Comparison of genetic algorithm with conventional CBR.

No	Conventional CBR			GA with 1 generation			GA with 2 generations			GA with 5 generations		
	Comp. time, s	The numb.	Accu-racy	Comp. time, s	The numb.	Accu-racy	Comp. time,s	The numb.	Accu-racy	Comp. time,s	The numb.	Accu-racy
1	11	3534	66%	14	1767	66%	12	1768	66%	29	1726	66%
2	11	3534	66%	13	1767	66%	12	1727	66%	26	1769	65%
3	11	3534	66%	8	1757	68%	16	1793	67%	22	1792	66%
4	11	3534	66%	10	1747	65%	17	1755	65%	23	1778	65%
5	11	3534	66%	9	1761	66%	17	1816	66%	20	1760	67%
	11	3534	66%	10.8	1759.8	66%	14.8	1771.8	66%	24	1770.4	66%

#### 4. Conclusion

Thus, we considered the basic stages of reasoning by analogy, and also the peculiarities of the CBR-cycle. We proposed formal statement of the genetic algorithm for the problem of retrieving a subset of cases relevant to the problem solved. The results of the conducted research on the test data were presented that testify to good prospects for using the genetic algorithm not only in the retrieving stage but also in the adaptation stage of the CBR-cycle. It seems promising to integrate the genetic algorithm with the generation of fuzzy rules [12] to implement the adaptation of retrieved cases to the problem being solved.

#### Acknowledgements

The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/PCh.

#### References

- [1] DTI. Knowledge-based systems survey of UK applications. Department of Trade & Industry UK, 1992.
- [2] Schank RC, Abelson RP. Scripts, Plans, Goals and Understanding. Erlbau, 1977.
- [3] Schank RC. Dynamic Memory: A theory of reminding and learning in computers and people. Cambridge University Press, 1982.
- [4] Wittgenstein L. Philosophical Investigations. Blackwell, 1953.
- [5] Watson I, Marir F. Case-based reasoning: A review. The Knowledge Engineering Review 1994; 9(4): 327–354.
- [6] Bonzano P, Cunningham P, Smith B. Using introspective learning to improve retrieval in CBR: A case study in air traffic control. Proc. 2nd Int. Conf. Case-based Reasoning, 1997; 291–302.
- [7] Cercone N, An A, Chan C. Rule-induction and case-based reasoning: Hybrid architectures appear advantageous. IEEE Trans. Knowledge and Data Engineering 1999; 11: 166–174.
- [8] Coyle L, Cunningham P. Improving recommendation ranking by learning personal feature weights. Proc. 7th European Conference on Case-Based Reasoning, 2004; 560–572.
- [9] Jarmulak J, Craw S, Rowe R. Genetic algorithms to optimize CBR retrieval. Proc. European Workshop on Case-Based Reasoning (EWCBR 2000), 2000; 136–147.
- [10] Yang HL, Wang CS. Two stages of case-based reasoning - Integrating genetic algorithm with data mining mechanism. Expert Systems with Applications 2008; 35: 262–272.
- [11] Adult Data Set. UCI Machine Learning Repository. URL: <http://archive.ics.uci.edu/ml/datasets/Adult>.
- [12] Avdeenko TV, Makarova ES. Integration of case-based and rule-based reasoning through fuzzy inference in decision support systems. Procedia Computer Science 2017; 103: 447–453.

# Big Data incorporation based on Open Services Provider for distributed enterprises

O.L. Surnin<sup>1</sup>, P.V. Sitnikov<sup>2</sup>, A.V. Ivaschenko<sup>3</sup>, N.Yu. Ilyasova<sup>3,4</sup>, S.B. Popov<sup>3,4</sup>

<sup>1</sup>SEC "Open Code", 55, Yarmarochnaya Str., 443001, Samara, Russia

<sup>2</sup>Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, 14, lit. A, Birzhevaya liniya, 199034, Saint-Petersburg, Russia

<sup>3</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>4</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

---

## Abstract

There is provided a new software solution for multiple data sources integration at modern enterprises with distributed organizational structure. Open Services Provider (OSP) is a platform powered by SEC "Open code" that allows developing situational centers for decision making support based on Big Data analysis and visualization. The paper describes a problem of management of modern distributed enterprises, the proposed OSP solution and results of its probation in practice. Research is supported by Big Data engineering center at Samara University.

*Keywords:* Big Data; Open Services Provider; Complex Automation; Integration

---

## 1. Introduction

In order to provide high competitive power and soundness most large industrial enterprises need to cooperate and share their resources. This trend enforces top management to introduce modern administration technologies that are based on interaction in matrix organizational structures, and caused by it information analysis and data flows integration and exchange. In case of high autonomy of involved parties and flexibility of cooperation the process of their integration in solid information space becomes hard. To solve this problem there is proposed an IT solution based on service-oriented software architecture capable of adaptation to new integration requirements and processing the Big Data of informational interaction between several enterprises with autonomous behavior. This vision is predominantly influenced by using experience of working with the Internet and mobile devices. This paper presents the features of Open Service Provider powered by SEC "Open code" and some benefits of its implementation in practice.

## 2. Theoretical background overview

Basic challenges of information processes management at modern supply chains are concerned with a necessity to support a horizontal interaction between a number of enterprises with various goals and tasks. In order to consider this factor, matrix organizational architectures are introduced [1]. The use of matrix models is also determined by the big number of projects as this organizational structure is considered the best to support project management activities and share resources between functional structures. A number of theories are studying the control over network organizational structures, for example the theory of hierarchical management describes the problems of decision making under the circumstances of unpredictability [2]. Self-organization in networks applicable for enterprise management is investigated using a Bio-inspired approach [3, 4].

Peer-to-peer (P2P) networks [5, 6] are used in practice to simulate the work inside matrix organizational structures. P2P models are frequently used to describe and simulate interaction processes in organizations with the network structure and autonomous decision makers. Actors, representing employees in the integrated information space, are the peers of the network as they are autonomous enough to make decisions and to use their own resources for project execution.

In order to solve this problem, there is a proposition of a new model of Open Service Provider based on the technologies of business processes automation and self-organization support. The idea and principles are similar and inspired by the intelligent solutions in transportation logistics, the Internet and multifunctional centers [7, 8]. This proposed approach is close to 5PL (Fifth Party Logistics) concept, which is based on implementation of a number of services for customers and enterprises provided by a specially designed software platform. 5PL platform is open for new transportation companies and even drivers and helps them negotiate with customers in integrated information space.

Interaction of customers and service providers powered by intermediary services [9] generate and can be characterized by a big number of events that form Big Data and require modern technologies for its analysis [10]. Business processes that support such interaction should be flexible and dependent on unique customer requirements. This makes it reasonable to implement subject-oriented approach for business processes management (S-BPM), which conceives a process as a collaboration of multiple subjects organized via structured communication [11].

An OSP concept is similar to the idea of «One Internet» governance [12]. The common trend in these areas is virtualization: a web aggregator that collects information about applications from potential buyers and the information about service providers and then links them on the basis of P2P principles is introduced. This web aggregator provides the best options for implementation of services for both sides: buyers (future users) and software providers.



### 3. A conceptual model and problem statement

Let us consider the parts of the integrated information space built as a result of enterprise complex automation as a list of services  $s_j$ , where  $j = 1..N_s$  is a number of service.

Each service has corresponding problem domain  $d_i$ ,  $i = 1..N_d$ : e.g. customers management, product lifecycle management, counting, HR management, etc.

In this sphere, each service requirement can be described by a Boolean variable:

$$r_{i,j,l} = r_{i,j,l}(d_i, s_j, t_l) \in \{0, 1\}, \quad (1)$$

where  $t_l$ ,  $l = 1..N_r$  is the  $s_j$  – order submission time.

The fact of each service delivery is defined by:

$$v_{i,j,l,k} = v_{i,j,l,k}(r_{i,j,l}, g_k, c_{i,j,l,k}, \Delta t_{i,j,l,k}) \in \{0, 1\}, \quad (2)$$

where  $g_k$  represents a possible service provider (IT company),  $k = 1..N_g$ ,

$c_{i,j,l,k}$  – the costs of the service to be delivered,  $\Delta t_{i,j,l,k}$  – the period of time required by the service to be delivered, including implementation, integration, testing and QA.

In this model, we assume that multiple providers can implement and deploy one service, which is significant for a business with high competitiveness. The number of options  $v_{i,j,l,k}$  generated for each demand is limited by the current service provider capabilities and their core competence.

Options  $v_{i,j,l,k}$  are related to each other in resources: the same providers  $g_k$  can be used for different services allocation. For two service options  $v_{i,j_1,l,k}, v_{i,j_2,l,k}$ ,  $j_1 \neq j_2$ , we can also define the relations of:

- sequence  $\phi(v_{i,j_1,l,k}, v_{i,j_2,l,k})$ , one service requires for its start one or several preceding services to be completed, and
- combination  $\psi(v_{i,j_1,l,k}, v_{i,j_2,l,k})$ , the services are implemented simultaneously.

Therefore, there is a generated virtual network of services, combined with a network of options  $v_{i,j,k}$  with transitions of the sequence and relation to one demand or resource.

The proposed model allows formalizing the following challenges of OSP. Firstly, it is necessary to minimize the services delivery costs, which makes the platform attractive for users:

$$C(d_i) = \sum_{j=1}^{N_s} \sum_{l=1}^{N_e} \sum_{k=1}^{N_g} v_{i,j,l,k} \cdot c_{i,j,l,k} \rightarrow \min. \quad (3)$$

Next, the operational efficiency and performance of services should be high:

$$T(d_i) = \sum_{j=1}^{N_s} \sum_{l=1}^{N_e} \sum_{k=1}^{N_g} v_{i,j,l,k} \cdot (t_{i,l,\min}^{fin} - t_{i,l,\min}) \rightarrow \min, \quad (4)$$

where  $t_{i,l,\min}^{fin}$  is a  $d_i$  delivery time.

Finally, the individual earnings of each real service provider should also be high, which comes to a certain contradiction with the goal (3):

$$\forall g_k : \sum_{i=1}^{N_d} \sum_{j=1}^{N_s} \sum_{l=1}^{N_e} v_{i,j,l,k} \cdot c_{i,j,l,k} \rightarrow \max. \quad (5)$$

The solution of the introduced problem is specified as a set of non-zero values of Boolean variables

$$\mu(d_i) = \{v_{i,j,l,k}(r_{i,j,l}, g_k, c_{i,j,l,k}, \Delta t_{i,j,l,k}) = 1\}, \quad (6)$$

that can be referred to as an IT strategy with cost  $C(d_i)$ .

There can be multiple IT strategies for problem domains  $d_i$ , so the basic problem of OSP is to find and dynamically manage the interaction between IT services providers and users considering the challenges (3 – 5).

### 4. Solution vision

Considering the contradiction of stated problem (3 – 5), it is proposed to solve it constructively, in the form of a design of a specific IT platform that provides the users and developers of IT services with OSP functionality for interaction. The OSP solution is presented in Fig. 1. A modern enterprise contains a few departments that cooperate with each other based on the P2P principle of information exchange. The platform supports both hierarchical and matrix organizational negotiations.

Software products and solutions can be accessible by certain services implemented in the integrated information space. In the modern Internet, realization of specific features becomes more concealed for users. When users visit different sites and portals, this process seems to them like using widgets on their dashboard. To implement this idea, it is necessary to develop a functional aggregator that provides the users with a variety of services with unified API and UI. On the other hand, to implement the functionality, it is necessary to develop a unified service aggregator that should be able to involve various service providers. This aggregator should have an open architecture, support interoperability and the set of unified intelligent software solutions for decision making support and application of the unified technology of combined security and data storage.

The proposed approach allows involving all actors into the process of decision making. The users get access to new functionality immediately and directly, software providers get the opportunity to easily access possible users on a competitive basis, and enterprise top management get a powerful analytical tool that provides a realistic picture of users interaction based on real statistics. Consequently, it becomes capable of controlling the entire IT-infrastructure of the company. An IT department in

this case forms the goals and objectives for service providers and users. Due to it, the company management obtains an opportunity to monitor and influence functional aggregators, receive an overall picture of their work and realize the process of decision making. Service providers, in turn, are motivated to permanent changes, updates and upgrades.

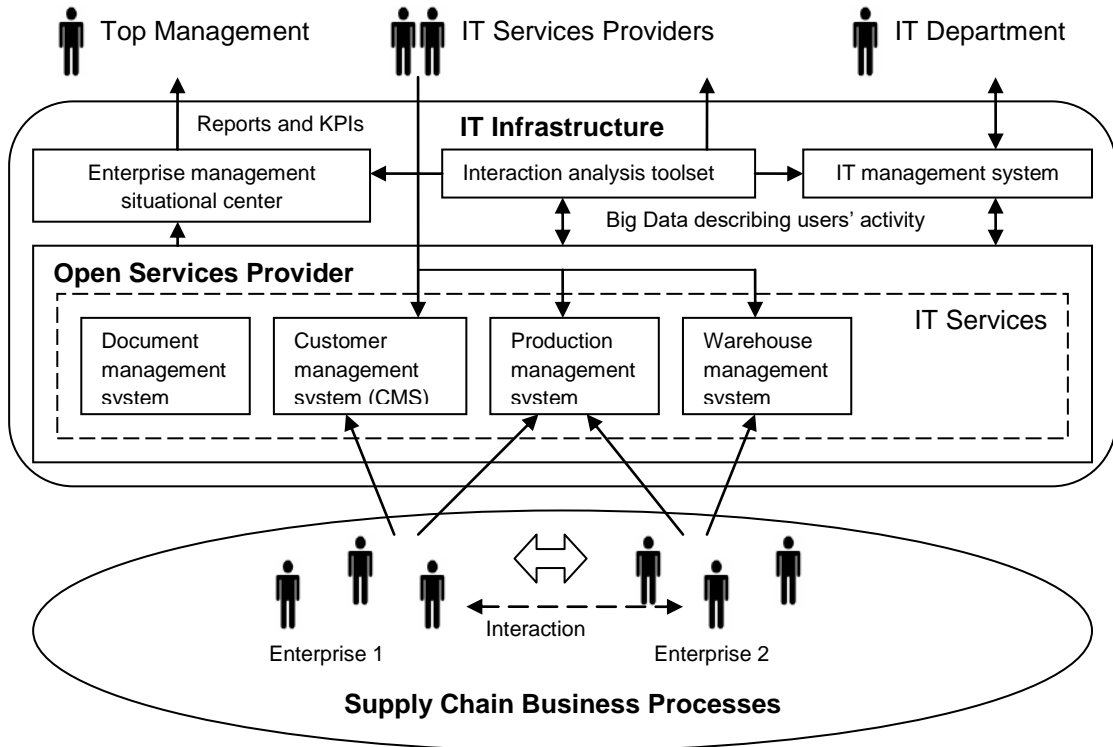


Fig. 1. An open service provider concept.

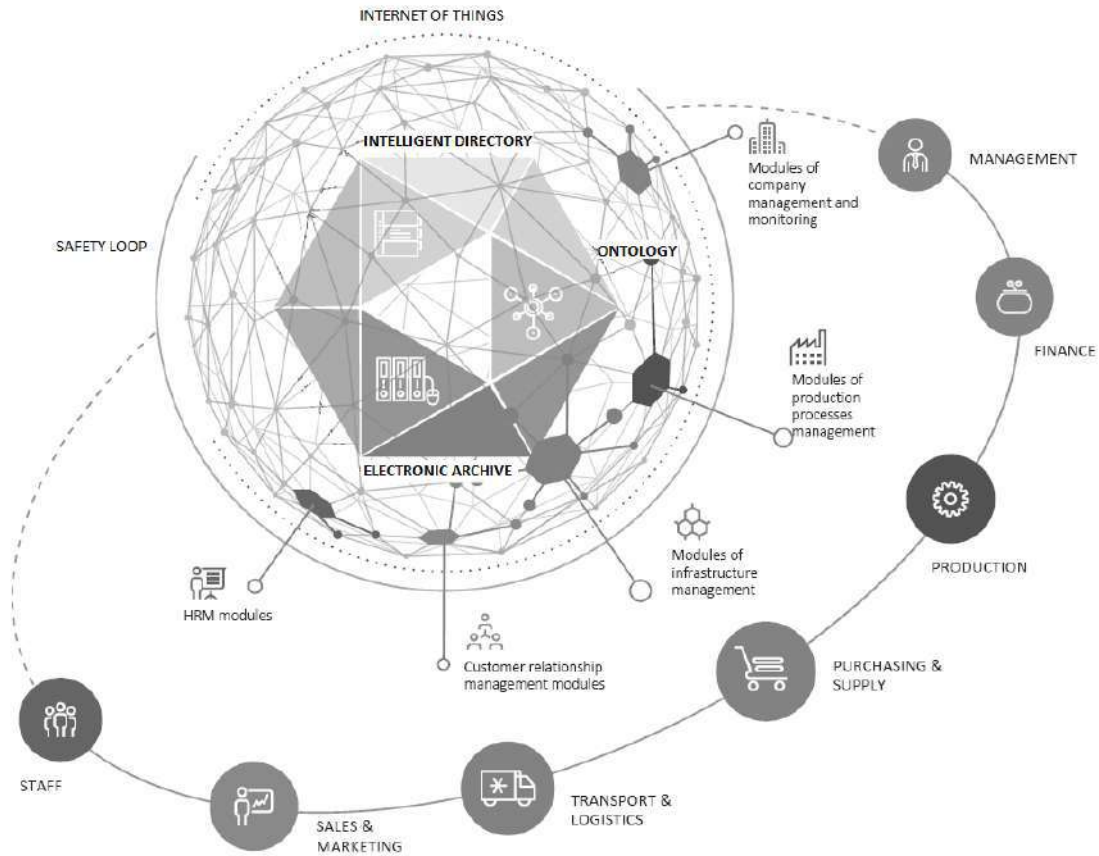


Fig. 2. An "Open Code" OSP solution.

## 5. OSP architecture

Implementation of the proposed concept and approach was performed by SEC “Open Code” for a number of IT solutions of complex automation of industrial enterprises and supply chains. A number of IT services were built on the basis of three components: knowledge base (ontology), electronic archive and intelligent directory. Open Service Provider was introduced to bring together these services. The resulting solution is presented in Fig. 2.

OSP becomes an open platform to provide enterprises different services based on implementation of the intermediate module of negotiation. In the process of OSP implementation, a number of technical problems were successfully solved. First of all, it was necessary to solve the problem of OSP scalable architecture development and componentization, to implement the functionality for configuring, adaptability and self-organization, to resolve issues related to the maintenance of the archives, to document management and event registration. Then, the enterprise information environment was revised so that users get convenient access to services, providers have access to the services registration and support and the management staff has been able to keep track of all these processes. Finally, enterprise business processes were reviewed and built in such way that users could understand the features of the services in the Internet instead of a software solutions with predefined fixed functionality.

## 6. Implementation

The example of OSP implementation for large production group of companies is given in Fig. 3. Organizational project management system represents set of the subdivisions or enterprises combined by a solid supply chain that are connected by relations and subordinations. In the case of management structure creation, it is necessary to consider specifics of enterprises' activities and features of their interactions with an external environment.

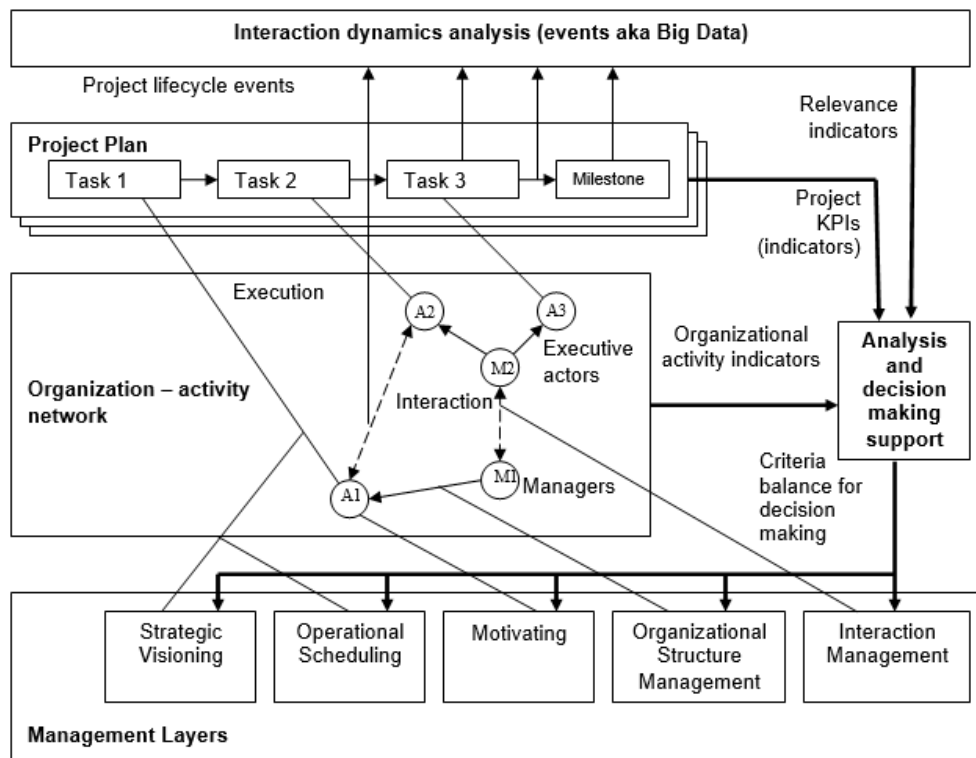


Fig. 3. Coordination of plans based on Big Data analysis.

The process the organization structure formation of project management usually includes three main stages: determination the type of the organization structure (direct subordination, functional, matrix, etc.); separation of structural subdivisions (administrative staff, independent subdivisions, applications programs, etc.); delegating/devolution of the authority and responsibility for parts of the project to the subordinate authority levels (governance relation – subordination, the centralization relation – decentralization, organizational mechanisms of coordination and monitoring, a regulation of subdivisions' activities, development of regulations in structural subdivisions and positions).

This architecture affects the organization structure of enterprises' functioning, project part, management system, units of the analysis and analytics, product life cycle events, the functional relations. Resource assignment is provided according to the performed project specification (tasks) in the form of the oriented organization – activity network. The nodes represent the staff of the enterprise (performers and their principals), and the links – the relations between the employees. Based on the proposed solution there can be introduced the following process of project management using Big Data analysis for knowledge engineering.

At the first stage, the enterprise management makes decision in implementation of a certain project. Then, it makes a decision about the decomposition of the project in a number of tasks. The project implementation (elaboration of each task) is followed

by the set of project life cycle events, and the efficiency of all projects' implementation depends on effective activity. It is worth mentioning that for large enterprises, project life cycle events form the Big Data.

Therefore, the processes of the overall performance analysis of the enterprise and the processes of finding the closest optimal decision are more complicated at each stage. Contingency planning involves identifying alternative courses of action that can be implemented if and when the original plan proves inadequate because of changing circumstances. Events beyond a manager's control may cause even the most carefully prepared alternative future scenarios to go awry. Unexpected problems and events frequently occur. When they do, managers may need to change their plans. Anticipating change during the planning process is best in case things don't go as expected. Management can then develop alternatives to the existing plan and ready them for use when and if circumstances make these alternatives appropriate.

Therefore, the processes of the overall performance analysis of the enterprise and the processes of finding the closest optimal decision are more complicated at each stage. Contingency planning involves identifying alternative courses of action that can be implemented if and when the original plan proves inadequate because of changing circumstances. Events beyond a manager's control may cause even the most carefully prepared alternative future scenarios to go awry. Unexpected problems and events frequently occur. When they do, managers may need to change their plans. Anticipating change during the planning process is best in case things don't go as expected. Management can then develop alternatives to the existing plan and ready them for use when and if circumstances make these alternatives appropriate.

## 7. Evaluation. OSP for Russian post office management

One of the successful examples of OSP implementation in practice is a software solution for Post office management, which was deployed and probated at Samara post office. This system was used to solve the problem of mail processing scheduling by Samara main mail sorting facility, which is a basic management unit of a posting supply chain. It requires effective allocation and scheduling of various resources that affect the post sorting procedures. One of the business features nowadays is the increasing outsourcing facilities: mail services extensively involve 3<sup>rd</sup> parties for e.g. transportation and delivery. Expected scheduling horizon is one month. Each transportation logistics unit has an own schedule that can be affected by unpredictable events, delays of other parties and failures of man force and equipment. In addition to this seasonal fluctuations and human factor has a strong implication over the business processes. All this information is big in volume changes in time, which makes it Big Data. The designed and delivered software solution is presented in Fig. 4 – 5.

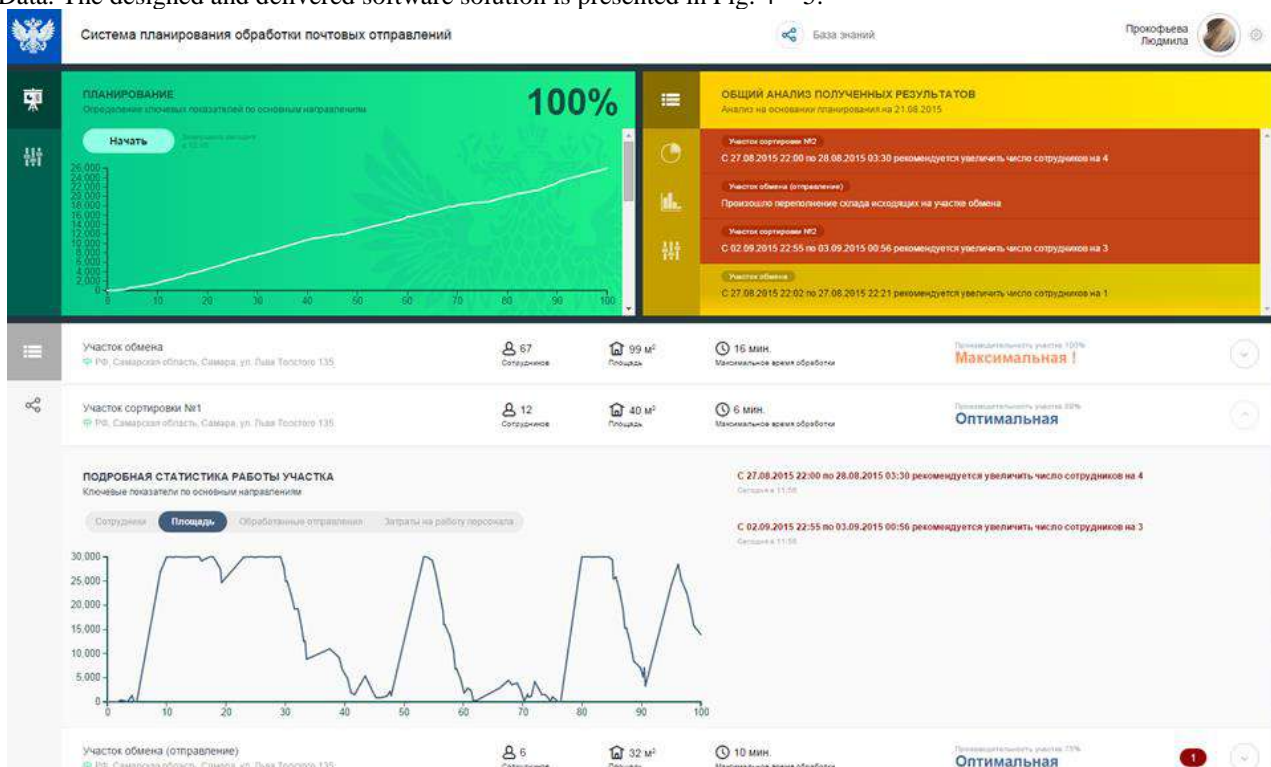


Fig. 4. Service providers statistics.

According to the introduced OSP model there were specified a number of services like transportation, sorting and delivery and service requirements on sorting personnel, space and time. This model allowed to state and solve an optimization problem of mail sorting processes optimization.

Software functionality is distributed between the widgets that present the enterprise KPIs for efficiency monitoring in real time, detailed information on mail sorting services, and automatically generated recommendation for decision making support. At the bottom there is presented a details statistics on resources utilization for a specified service. OSP concept allows to organize the scheduling system as an open platform for integrated services provided by various resources and therefore allow operator a toolset for their coordination and optimization in real time.

First results of the proposed solution probation in practice allowed to reduce the required man force by 20 employees per sorting center, reduce transportation costs by 720 000 RUB per year, and reduce the queues of mail by minimization of time needed for its sorting at the sorting center. This result has proven the benefits of OSP.



Fig. 5. System recommendations for decision making support.

## 8. Conclusion

The proposed concept for Open Service Provider allows enterprises to incorporate their data flows and resources and construct a distributed and flexible matrix management architecture. OSP benefits includes easy adaptation, configuration flexibility, an ability to expand, a possibility of constant updates in response to changing users' needs and a technical capability of constant updating and being in a permanent state of efficiency. One of the important advantages is the simplicity of support after the system implementation.

## 9. Acknowledgment

This work was partially supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world's leading scientific and educational centers for 2013-2020 years; by the Russian Foundation for Basic Research grants (# 15-29- 03823, # 15-29- 07077, # 16-41-630761; # 16-29- 11698); by the ONIT RAS program # 6 "Bioinformatics, modern information technologies and mathematical methods in medicine" 2017.

## References

- [1] Ford RC, Randolph WA. Cross-functional structures: A review and integration of matrix organization and project management. *Journal of management* 1992; 18(2): 267–294.
- [2] Minar N. Distributed systems topologies: Part 1 (online), 2001. URL: [http://www.openp2p.com/pub/a/p2p/2001/12/14/topologies\\_one.html](http://www.openp2p.com/pub/a/p2p/2001/12/14/topologies_one.html).
- [3] Leitao P. Holonic rationale and self-organization on design of complex evolvable systems. *HoloMAS, LNAI 5696*. Berlin, Heidelberg: Springer-Verlag, 2009; 1–12.
- [4] Gorodetskii VI. Self-organization and multiagent systems: I. Models of multiagent self-organization. *Journal of Computer and Systems Sciences International* 2012; 51(2): 256–281.
- [5] Schoder D, Fischbach K. Peer-to-peer prospects. *Communications of the ACM* 2003; 46(2): 27–29.
- [6] Ivaschenko A, Lednev A. Time-based regulation of auctions in P2P outsourcing. *Proc. IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*. USA, Atlanta, Georgia, 2013: 75–79.
- [7] Hickson A, Wirth B, Morales G. Supply chain intermediaries study. University of Manitoba Transport Institute, 2008; 56 p.
- [8] Ivaschenko A. Multi-agent solution for business processes management of 5PL transportation provider. *Lecture Notes in Business Information Processing* 2014; 170: 110–120.
- [9] Ivaschenko A, Dvoynina O, Sitnikov P, Syusin I. Intermediary service provider for supply chain. *Proceedings of the 18th FRUCT & ISPIT Conference*. Technopark of ITMO University, Saint-Petersburg, Russia 18-22 April, 2016: 480–485.
- [10] Bessis N, Dobre C. *Big Data and Internet of Things: A roadmap for smart environments*. Studies in computational intelligence. Springer, 2014; 450 p.
- [11] Fleischmann A, Kannengiesser U, Schmidt W, Stary C. Subject-oriented modeling and execution of multi-agent business processes. *Proc. IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)*. USA, Atlanta, Georgia, 2013; 138–145.
- [12] One Internet. Global commission on Internet Governance, 2016. Report. 138.

# Big Data Analysis for Demand Segmentation of Small Business Services by Activity in Region

V.M. Ramzaev<sup>1</sup>, I.N. Khaimovich<sup>1,2</sup>, V.G. Chumak<sup>1</sup>

<sup>1</sup>International Market Institute, Aksakova street, 21, 443030, Samara, Russia

<sup>2</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Annotation

The article suggests a tool for the efficiency improvement in using budget funds in the region in the sphere of small business. This is the most important task in the current economic conditions, in which solution there is a possibility of making effective management decisions. The suggested method of regulation based on the analysis of social networks using BIG DATA technology can be effective in managing various innovative processes of economic development in the region, which are characterized by a variety of forms and a wide range of components and factors, as well as dynamic development and active transformation of life. The use of modern software and hardware from BIG DATA technology allows real time evaluation and visualization of changes

*Keywords:* competitiveness; territory management; intensive data; mathematical models; BIG DATA technology

---

## 1. Introduction

Under modern social and economic conditions the vital task is the state regulation of market economy players, among which one of the most important in the region is small business (SB). The foreign experience shows that without this sector it is impossible to develop economy as far as the economic growth rate depends on it as well as the structure and the quality of up to 40-50% of gross national product.

## 2. Subject of research

The structure of small and medium-sized enterprises by types of economic activity is varying. As it can be seen in Figure 1, the largest number of enterprises are engaged in trade, repair of motor vehicles, motorcycles, household products and personal items, which is explained by lower barriers to entry these areas of activity.

The following features of SB development management can be distinguished. First, it is necessary to note a wide range of services provided by SB subjects, as well as a huge range of goods sold by them. Secondly, the SB differs significantly in being more mobile in comparison with the large one. By the mobility we mean a continuous change in the market conditions, the closure of old and the emergence of new economic entities, which is explained by the high variability in tastes and preferences of consumers of goods and services of SB entities, i.e. there is a quite active process where some types of activities are substituted by others, determined by a change in consumer demand, which is especially important under the modern conditions of import substitution. Thus, according to the statistics, up to 85% of the new entities of the SB is closed during the first year of its existence. 94 out of the 100 registered small businesses stops operations by the fourth year.

In this regard, the use of traditional methods of public administration, based on the data of monthly, quarterly and annual statistics, does not bring the expected result and does not allow us to identify trends for the development or closing up of certain activities, therefore, often making decisions about financial support and funds allocation for some projects is significantly behind the needs, and in some cases also contradict the changed real market situation by the time the financing begins.

For example, at the present time in the Samara region, state support for small and medium-sized enterprises is being implemented within the framework of the State Program “Development of Entrepreneurship, Trade and Tourism in the Samara Oblast” for 2014 - 2019, approved by the Government of the Samara Oblast Decree No. 699 dated November 29, 2013. Support to businessmen of the Samara Oblast is maintained in different directions and consists of information and consulting services, training, financial assistance, assistance in selling goods and services.

At the same time, it should be noted that, despite a number of measures used by the authorities in the region to manage the development of SB, effective methods for selecting priority directions for the development of SB have not been developed so far, which make it possible to direct budgetary funds to the development and support of entrepreneurs more appropriate [1-4]. The market of small and medium-sized businesses is a quite dynamically changing environment. It is necessary to take this into account in the medium and long term planning and regional authorities should take it as a basis for the support and stimulation of the development of the most high-demand areas of the SB activities and for monitoring the effectiveness of budgetary funds application for programs in this field of entrepreneurship under the changing market conditions.

## 3. The methods of applying the business intelligence at determining small business segments in the region

This task may be solved with the help of modern information technologies [5,6,7], to which BIG DATA technology refers, directly connected with business intelligence [8,10]. Along with this application of modern BIG DATA technologies provides an opportunity to distinguish the zones – territories of the most active consumption and demand for some or other products and services on the market in real time mode.



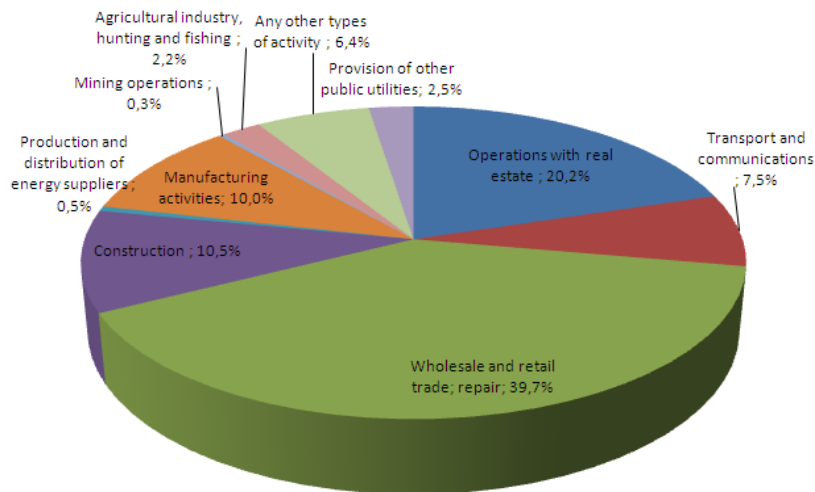


Fig. 1. The structure of small and medium-sized enterprises by types of economic activity at the end of 2014, %.

To manage the development of small and medium-sized businesses in the region the special methodology was worked out based on the BIG DATA technology [9], which consists of the following stages: identification of the role and place of small business in the region; identification of the main types of goods and services, offered by small businesses in the region; creation of consumer profile who uses the services of small business; creation of information model of small business consumer in the region; formation of small business zones in the region; development of guidelines for management decision making.

If the role and place of small business in the region, main types of goods and services, offered by entrepreneurs in the region were analyzed then to create a consumer profile and information model it is necessary to use BIG DATA technology. The method of applying the business intelligence is as follows:

1. Formation of a set of BIG DATA in hadoop from twitter using filter Samara Oblast, showing the hit count;
2. Division of formed set into different filters connected with basic factors of small business;
3. Carrying out monitoring of flow content analysis in filters;
4. Taking quick actions in cases of stable “burst” of hit count;
5. Program development in Scala language to work with filtration in the BIG Data area;
6. Program debugging and testing with a set of practical data;
7. Analysis of computational results.

To receive data we use social network «twitter», as it is “open” product, its application does not require any additional investment, and 50% of Internet users have profiles in this program. Twitter is the second in popularity network among the users in the entire world, come second only to Facebook. However unlike Facebook, which does not make accessible its data, Twitter provides such access, there are no limitations in access to the sets of data in the server. The users of this social network share mainly text messages, and this fact is absolute advantage while processing. Twitter is not a network with a specific focus and more broadly reflects public opinion in many points of interest, that is why the processing of data from this social network was the best possible to form small business zones in the region.

To work with BIG DATA in social networks we used the methods of collecting, processing and analyzing the data. Data collecting is carried out in a real time, within the certain geo location, or within the entire network according to the predefined patterns. Information of interest for analysis in the area of SM is: location, date and time, content, “author” of content (user), links with users. Data collecting may be fulfilled with the help of following tools: Apache Hadoop, BigInsights (IBM), Cloudera, Hortonworks, Storm. To carry out the research in the field of SB we chose Hortonworks. We used Twitter Application (apps.twitter.com), where the key parameters were defined using API key, API secret, Access token, Access token secret.

For data collecting with Hortonworks, Twitter App we used flume service configuration file in Hortonworks Virtual Machine Sandbox. System is ready to load data from twitter after Hortonworks Virtual Machine Sandbox version 2.3 is installed and flume service is configured. Navigate to HDFS folder in order to view downloaded files for data processing. HDFS view in Hortonworks virtual machine while solving tasks in the area of SB is shown in Fig. 2.

Collected data must be structured (i.e. processed) according to MapReduce paradigm. MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

MapReduce gave ability to structure the data flow from social networks using following criterion: font, text size, color, user profile hyperlink, location, date and others.

In order to define user profile for SB in our research we need data of following types: location, text, language and date. We used MapReduce to retrieve only required data in Hortonworks Sandbox tool. For data processing in Hadoop environment we chose Hive DB that gives ability to operate with the data and apply analysis via SQL-like queries. For this we created sql-script hivedll.sql for necessary tables creation. File contents is shown below:

```
// twitter table identifiers
CREATE EXTERNAL TABLE tweets_raw (
  id BIGINT,
  created_at STRING,
```

```

source STRING,
favorited BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text:STRING,
usr:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
usr STRUCT< screen_name:STRING, name:STRING, friends_count:INT, followers_count:INT, statuses_count:INT,
verified:BOOLEAN, utc_offset:STRING, -- was INT but nulls are strings time_zone:STRING>,
in_reply_to_screen_name STRING,
yearint,
monthint,
dayint,
hourint
)
CREATE EXTERNAL TABLE time_zone_map (
time_zone string,
country string,
notes string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION '/user/data/time_zone_map';
...
create table tweets_sentiment stored as orc as select
id,
case
when sum( polarity ) > 0 then 'positive'
when sum( polarity ) < 0 then 'negative'
else 'neutral' end as sentiment
from l3 group by id;
-- put everything back together and re-number sentiment
CREATE TABLE tweetsbi
STORED AS ORC
AS
SELECT
t.*,
cases.sentiment
when 'positive' then 2
when 'neutral' then 1
when 'negative' then 0
end as sentiment
FROM tweets_clean t LEFT OUTER JOIN tweets_sentiment s on t.id = s.id.

```

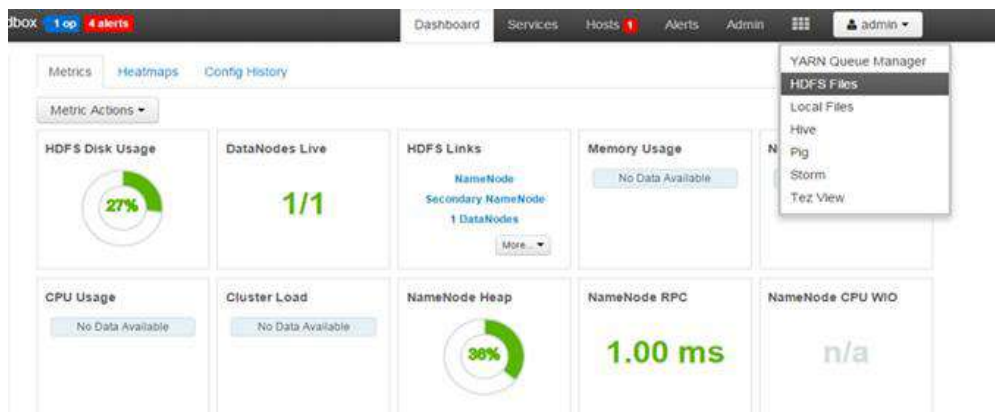


Fig. 2. HDFS view in Hortonworks when downloading files while solving tasks in the area of SB.

Execute script using command: Hive\_f hiveddl.sql. Structured data are placed in Table 1.



Table 1. Column headers for structured data analysis in tasks for SB.

A	B	C	D	E	F
Data/Time	Time/Zona	language	Text	location	Sentiments

The following metrics are used for data analysis. The total number of tweets ( $Kol_i$ ) for every location ( $R$ ) is defined:

$$Kol_R = \sum_{i=1}^N k_i, k_i \in R,$$

Where  $k_i$  is the every following tweet from processing thread.

The unique word frequency  $ch(m)$  is defined from total collection  $L$  of text data:

$$ch(m) = \sum_{i=1}^N m_i, m_i \in L.$$

Relationship of every tweet  $otn(m, rez)$  can be defined from thesaurus  $tez$ , where relationship to every word is set:

$$otn(m, rez) = \begin{cases} 0, m - negative\_value \\ 1, m - neutral\_value \\ 2, m - positive\_value. \end{cases}$$

For further work we created a dictionary with filters of SB domain in order to identify the number of tweets by location  $ch(m)$  and number of tweets by location with respect of relationship  $otn(m, rez)$  hereafter. We define thesaurus taking into account filter with base metrics of small and medium-sized businesses: food, clothes, entertainment and kids. In conclusion we got 4 base metrics of medium-sized business.

Metric «food»  $P_1$  is calculated as number of tweets in overall text data  $L$ :

$$Kol_{omP_1} = \frac{\sum_{i=1}^N S_i(S_i \in P_1)}{L} = 9\%.$$

Metric «clothes»  $P_2$  is calculated as number of tweets in overall text data  $L$ :

$$Kol_{omP_2} = \frac{\sum_{i=1}^N S_i(S_i \in P_2)}{L} = 8\%.$$

Metric «entertainment»  $P_3$  is calculated as number of tweets in overall text data  $L$ :

$$Kol_{omP_3} = \frac{\sum_{i=1}^N S_i(S_i \in P_3)}{L} = 6\%.$$

Metric «kids»  $P_4$  is calculated as number of tweets in overall text data  $L$ :

$$Kol_{omP_4} = \frac{\sum_{i=1}^N S_i(S_i \in P_4)}{L} = 12\%.$$

#### 4. Results and discussion

As a result it is possible to conclude what area of SB is especially in high demand in Samara Oblast. According to the Figure 3 it is apparent that the main strategy of SB promotion for authorities must be connected with opening of centers for children.

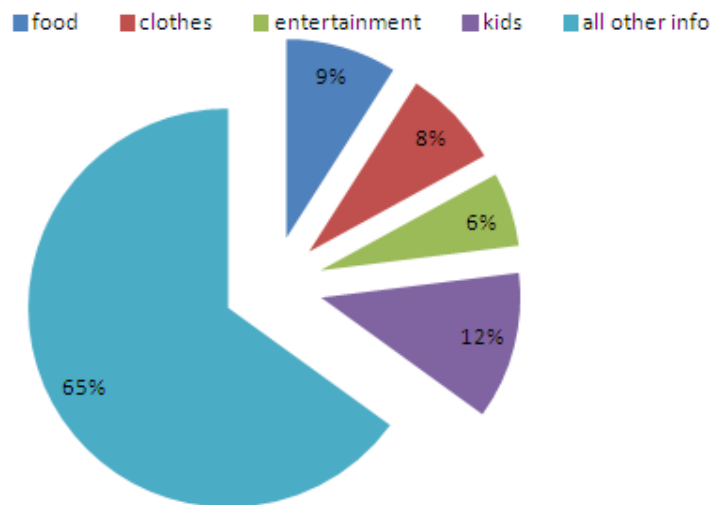


Fig. 3. Metrics of small business in Samara Oblast.

Due to BIG DATA technology it is possible to distribute and update data in «hadoop» file system using filter “Samara Oblast” (filter1= {Samara Oblast}). Then it is necessary to filter this area on base metrics of small and medium-sized businesses, setting up for example the following metrics: Filter2 (food) = {cafe, bar, restaurant, cuisine\*, beer\*, meat, fish, tavern}; Filter3 (clothes)= {coat, jacket, dres\*, skir\*, jacke\*, bra\*, stuf\*}; Filter4 (entertainment)= {night club, concert, session, hangout}; Filter5 (kids) = {kindergarten, baby-club, club}.

A set of descriptors for filtering the Internet discourse will be determined by the lexical representatives of the concept formed in the world building of the average Russian-speaking consumer of services. The main in the sphere of concepts "Food" is the micro-situation "Cooking", which includes the following cognitive and propositional structure: Subject - Predicate of cooking (how it is cooked) - Object of cooking - The property of the cooking object - Method of cooking – Premises - Kitchenware – Appliances - Devices - Affair- Substance - Food / Dish – Food quality / Dish quality. In the situation of Internet communication, only the structure elements relevant to the user are being explicated, the lexical interpretation of which let us draw a conclusion about the needs of the residents of a particular district of Samara city. Building –up of block of descriptors on Filter3 (clothes); Filter4 (entertainment); Filter5 (kids) may be fulfilled according to the lexical and semantic fields “clothes”, “fashion”, associative and semantic field “leisure”; concept “childhood”.

For making decision in the area of SB it is necessary to create multimodal clusterization of social networks. The clusterization is based on the method of Formal Concept Analysis (FCA). A large number of structured and unstructured data generate trivial data. For example, the data of social websites in the SB area may be submitted in the form of following three items (user, group, interest) (Fig. 4).

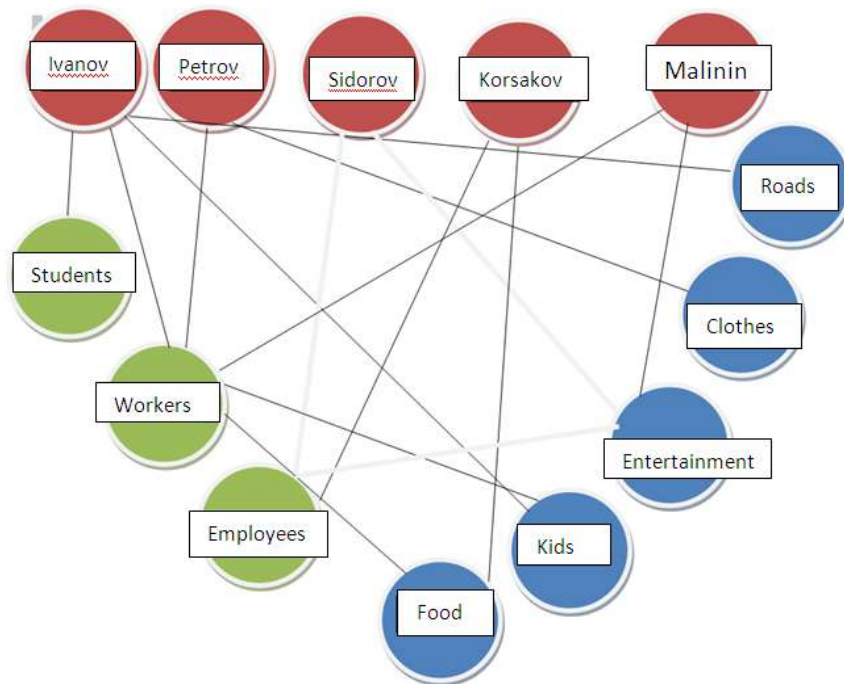


Fig. 4. SB data from social network «twitter» as a graph.

By the method of formal notions it is necessary to introduce the following definitions:  $G$  - set of objects,  $M$  – feature set, the dependence of  $I \subseteq G \times M$  such that  $(g, m) \in I$  when and only when the object  $g$  posses the feature  $m$ ;  $K := (G, M, I)$  is called formal context.

Galois operator is defined in the following manner: for  $A \subseteq G, B \subseteq M$   $A' \stackrel{def}{=} \{m \in M \mid g / m \forall g \in A\}$ ,  $B' \stackrel{def}{=} \{g \in G \mid g / m \forall m \in B\}$ , where  $A$  is the formal volume,  $B$  is the formal content.

Formal notion is the pair  $(A, B) : A \subseteq G, B \subseteq M, A' = B$  and  $B' = A$ .

Notions ordered by ratio  $(A_1, B_1) \geq (A_2, B_2) \Leftrightarrow A_1 \supseteq A_2 (B_2 \supseteq B_1)$ , from the complete lattice, called a context lattice  $\underline{\beta}(G, M, I)$ .

The example of social network context in the SB area and their context lattice are shown in Table 2 and in Figure 5.

Table 2. The example of SB data context from social network (a is the attributes of “food” filter, b is the attributes of “kids” filter, c is the attributes of “entertainment” filter, d is the attributes of “clothes” filter).

G/M		a	b	c	d
1	Pensioners	x			x
2	Employees	x		x	
3	Workers		x	x	
4	Students		x	x	x

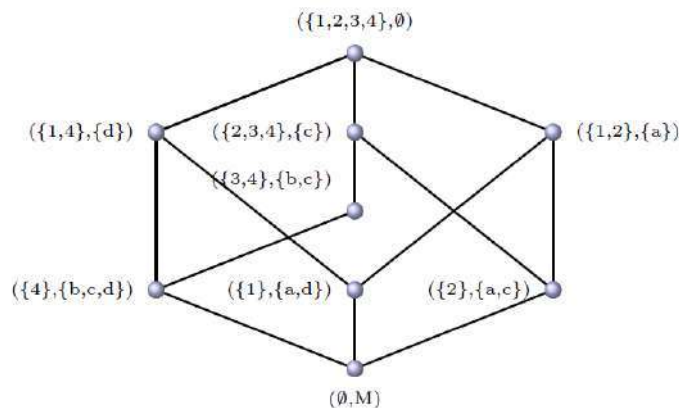


Fig. 5. Context lattice for social network.

The use of this clusterization method permits to define the groups of interest, with increase of connections where it is required to make managerial decisions. But this tool has restrictions of use. Users who work with social network Twitter are in the group of “students” and partly in groups of “employees” and “workers” and slightly in group of “pensioners”, that is why it is necessary to add field marketing research in these groups in order to take management decision.

There is ability to get correlation between number of user requests with respect to filters and date and time of data collecting [9]. Time of data collecting from Internet using Big Data is not limited.

As a result we get dynamic change of information in real time from Internet, which allows conduct monitoring of continuous analysis of unstructured information by filters with minimal investments (In-Memory Data Processing and Stream technology). For the purpose of this method implementation we coded a program using Scala language:

```
val file = spark.textFile("hdfs://... ")
val errors=file.filter(line=>line.contains("Samara Oblast"))
//count all the data
errors.count()
//count data mentioning Filter
errors.filter(line=>line.contains("meat")).count()
//Fetch the filter as an array of string
errors.filter(line=>line.contains("food")).collect()
```

After program execution we got dynamic change of parameters in BIG DATA environment, that allow to identify zone of SM business in geo region taking into account unstructured information. In case of consistent “peaks” detected in hit counts in accordance with forms of business there should be supporting investment program take place for development of small and medium sized businesses with a focus on certain business activity in target area.

In conclusion we suggested a tool for increasing of budget funds usage effectiveness in geo region. This is the most important challenge in modern economic reality, the solution for which is based on opportunity to take management decision in most optimal way. Suggested approach of regulation can be efficient in innovative process management in developing of region economy typical of lots of forms and wide range of factors, as well as dynamic progression and active transformation of daily living.

Using of modern software and hardware allows conducting evaluation and visualization of changes in almost real time that can be useful not only to local region governments but also to businesses in a way of design and implementation of investment projects.

## References

- [1] Drovyanikov VI, Khaymovich IN. Development of control pattern to manage the competitive improvement of social cluster of the region. *Fundamental Studies* 2015; 7(4): 822–827.
- [2] Drovyanikov VI, Khaymovich IN. Simulation modelling of social cluster administration in Any Logic system. *Fundamental Studies* 2015; 8(2): 361–366.
- [3] Ramzaev VM, Kukolnikova EA, Khaymovich IN. Development of a model for the functioning of production active elements in regional management. *Bulletin of SSEU* 2014; 12: 87–99.
- [4] Ramzaev VM, Khaymovich IN. Integrated model of control over economic development of the region based on competitiveness improvement of the enterprises. *Modern Issues of Science and Education* 2014; 6: 136 p.
- [5] Ramzaev VM, Khaymovich IN, Chumak VG. Forecasting model of competitive growth for enterprises with energy modernization. *Forecasting problems* 2015; 1: 67–75.
- [6] Bonacich P. Power and Centrality: A Family of Measures. *American Journal of Sociology* 2007; 92(5): 1170–1182.
- [7] Chumak PV, Ramzaev VM, Khaimovich IN. Models for forecasting the competitive growth of enterprises due to energy modernization. *Studies on Russian Economic Development* 2015; 26(1): 49–54.
- [8] Chumak VG, Ramzaev VM, Khaimovich IN. Challenges of Data Access in Economic Research based on Big Data Technology. *CEUR Workshop Proceedings* 2015; 1490: 327–337.
- [9] Chumak VG, Ramzaev VM, Khaimovich IN. Use of Big Data technology in public and municipal management. *CEUR Workshop Proceedings* 2016; 1638: 864–872.
- [10] Grechnikov FV, Khaimovich AI. Development of the requirements template for the information support system in the context of developing new materials involving Big Data. *CEUR Workshop Proceedings* 2015; 1490: 364–375.

# Matrix model of data and knowledge presentation to revealing regularities of the fluid flow regime in a pipeline based on hydrodynamics parameters

A. Yankovskaya<sup>1,2,3,4</sup>, A. Travkov<sup>2</sup>

<sup>1</sup>*Tomsk State University of Architecture and Building, 634003, Tomsk, Russia*

<sup>2</sup>*National Research Tomsk State University, 634050, Tomsk, Russia*

<sup>3</sup>*National Research Tomsk Polytechnic University, 634050, Tomsk, Russia*

<sup>4</sup>*Tomsk State University of Control Systems and Radioelectronics, 634050, Tomsk, Russia*

---

## Abstract

The study offers an original solution to one of the problems of hydrodynamics, namely revealing the regularities in the flow regime of fluid in a pipeline depending on the hydrodynamic parameters. The solution is based on using the intelligent system of the regularities revealing and decision-making. For the first time, a matrix model of data and knowledge representation (MM) is used for these purposes in the form of two matrices: descriptions of the fluid state in the space of characteristic features of hydrodynamics (pressure, velocity, temperature, and others); its rows are associated with various combinations of characteristic features values, and distinguishing of the diagnostic type, whose rows are associated with the corresponding rows of the description matrix, and its columns are associated with the two classifying features. The first classifying features takes four values corresponding to four fluid flow regimes, and the second classifying features takes three values only for the turbulent flow regime value from the first classifying features.

*Keywords:* matrix model; description matrix; distinguishing matrix; data and knowledge representation; regularities; fluid flow regimes; hydrodynamics; intelligent system; regularities revealing; decision-making

---

## 1. Introduction

Development of computer systems for current and precise determination of the fluid flow regime in a pipeline [1] is an extremely urgent issue in exploiting pipelines [2-3]. If a flow regime is determined inaccurately, it may interfere with or stop production, lead to breakdowns or other undesirable consequences. Such consequences may result in considerable expenses of finance and time. However, studies of the existing methods for calculating the fluid flow regime in a pipeline [3-7] have shown that these methods do not fully include all the parameters (features), which affect the fluid flow in a pipeline. This issue is of great importance to control fluidic flow inside oil pipes [8] because transportation of hydrocarbons is very dangerous process that requires continuous monitoring.

It is well known fact that every flow regime can be described with a set of parameters (features) determining the behavior of the fluidic flow. The exact type of the fluid flow depends on a number of features, such as velocity, viscosity, density, pressure, and more.

Taking into consideration all these factors, especially a large number of parameters influencing the type of fluidic flow in pipeline, it becomes clear that modern computer systems should be used to study regularities between these hydraulic parameters and support diagnostic decisions on the fluidic flow regimes. For these purposes, it is obviously rational to create intelligent systems (IS) for diagnostics of the fluid flow regimes in a pipeline (IS DFFRP) to determine various regularities, and also decision-making and their justification for such diagnostics.

Unlike the modeling methods described in the papers [10-11], proposed IS DFFRP is being developed specifically for oil transportation pipelines.

The Intelligent Systems Laboratory of Tomsk State University of Architecture and Building (TSUAB) under A.E. Yankovskaya's supervision has achieved significant results in the field. The researchers have developed 3 original intelligent instrumental software (IIS) for constructing applied intelligent systems. More than 30 applied intelligent systems are based on them. These systems are intended for different areas, such as geology, geoecology, medicine, ecology, electronics, psychology and other, more than 30 applications overall [12–25]. These IIS and applied intelligent systems are based on test methods of pattern recognition intended for various regularities revealing and decision-making, and also for making and justification decisions using cognitive tools. Thus, it is logical to apply the developed IIS to create an IS DFFRP.

The next section describes the matrix model to represent data and knowledge of hydraulics, their structuring, fragment of the description and distinguishing matrices as well as directions of the future research.

## 2. Matrix model of data and knowledge representation. Regularities

The IS DFFRP currently being developed is based on the matrix model of data and knowledge representation [25].

The matrix model includes an integer matrix of descriptions  $\mathbf{Q}$  and matrix of distinguishing  $\mathbf{R}$  [25]. They represent a learning sample with objects belonging to the known (determined by experts) patterns.

The rows of the  $\mathbf{Q}$  matrix are corresponded with the  $s_i$  learning objects ( $i = \overline{1, N}$ , where  $N$  is the quantity of objects), columns are corresponded with the  $z_j$  characteristic features ( $j = \overline{1, M}$ , where  $M$  is the quantity of features, which together represent the description of each object). The element  $q_{ij}$  of the  $\mathbf{Q}$  matrix sets the value of the  $j^{\text{th}}$  feature for the  $i^{\text{th}}$  object.

The data and knowledge base is formed on the base of the matrix model of data and knowledge representation [25], which includes the description integer matrix ( $\mathbf{Q}$ ) setting descriptions of objects in the space of  $k$ -valued  $z_1, \dots, z_m$  characteristic features, and the distinguishing integer matrix ( $\mathbf{R}$ ), which sets the partition of objects into equivalence classes according to each mechanisms of the classification. If the value of a feature is not essential for the object, it is marked with a dash ("-") in the respective element of the  $\mathbf{Q}$  matrix. For each  $z_j$  ( $j \in \{1, 2, \dots, m\}$ ) feature, either an interval of its value change, or an integer value is set.

Rows of the  $\mathbf{R}$  matrix are associated with the rows of the  $\mathbf{Q}$  matrix, columns are associated with the  $k_j$  classification features ( $j = \overline{1, L}$ , where  $L$  is the quantity of classification mechanisms partitioning learning objects into equivalence classes). An  $r_{i,j}$  element of the  $\mathbf{R}$  matrix sets belonging of an  $i^{\text{th}}$  object to a certain class (by designation its number) based on the  $j^{\text{th}}$  classification mechanism.

Objects with same combination of the classification features  $k_j$ , corresponding to a certain final solution, belong to the same pattern. This means that a number of patterns are equal to a number of non-repeating rows of the matrix  $\mathbf{R}$  and equal as well to the subset of the rows from the matrix  $\mathbf{Q}$  assigned to the same rows from the matrix  $\mathbf{R}$ . These mutually assigned rows describe patterns.

It should be noted that proposed model allows us to represent not only data but also experts' knowledge, because one row of the matrix  $\mathbf{Q}$  describes subset of the similar objects in the form of an interval (using a dash "-"). All these objects have the same final solution set by the corresponding row from the matrix  $\mathbf{R}$ .

Let us consider that the objects from the learning sample do not contain of measuring and entry errors. Otherwise it is necessary to reveal data and knowledge inconsistencies using a special subsystem realized in IS DFFPR.

Figure 1 shows an example of matrix representation of data and knowledge.

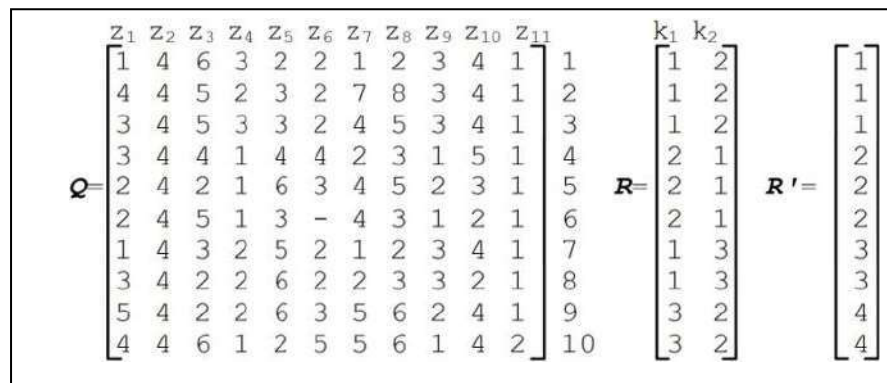


Fig. 1. Example of a description  $\mathbf{Q}$  matrix and a distinguishing  $\mathbf{R}$  and  $\mathbf{R}'$  matrix.

One of the important means of the data and knowledge analysis [25] are diagnostic tests, i.e. tests distinguishing objects from different patterns [25] constructed during of the regularities revealing in the data and knowledge base, which are represented by the  $\mathbf{Q}$ ,  $\mathbf{R}$  matrices. These tests are used for decision-making in the IS based on the methods of test pattern recognition.

Regularities are subsets of features with particular, easy-to-interpret properties that affect the distinguishing ability of objects from different patterns that are stably observed for objects from the learning sample and are exhibited in other objects of the same nature and weight coefficients of features that characterize their individual contribution [25] to the distinguish ability of objects and the information weight given, unlike, on the subset of tests used for a final decision-making.

These subsets include constant (taking the same value for all patterns), stable (constant inside a pattern, but nonconstant), non-informative (not distinguishing any pair of objects), alternative (in the sense of their inclusion in DT), dependant (in the sense of the inclusion of subsets of distinguishable pairs of objects), unessential (not included in any irredundant DT), obligatory (included in all irredundant DT), and pseudo-obligatory (which are not obligatory, but included in all IUDT involved in decision-making) features, as well as all minimal and all (or part, for a large feature space) irredundant distinguishing subsets of features that are essentially minimal and irredundant DTs, respectively. The weight coefficients of characteristic features are also included in regularities [25], as well as the information weight of characteristic features.

Regularities of the fluid flow regime in a pipeline include the described regularities used for decision-making and their justification. The revealed regularities will allow a significant decrease the number of measurements to determine the fluid flow regime in a pipeline.

### 3. Data and knowledge structuring. An illustrative example of presenting data and knowledge on the fluid flow regimes in a pipeline

#### 3.1. Data and knowledge structuring

According to the described matrix model of data and knowledge representation, for the purpose of to determine the fluid flow regime in a pipeline, data and knowledge in hydraulics were structured.

Since the limits of the paper, it is impossible to describe the whole feature space, whose dimension exceeds 30, so only a part of it and a fragment of the description and distinguishing matrices are introduced.

Based on the analysis of a range of papers [3-11, 26-29], 9 real characteristic features and their values used in forming the **Q** matrix were formulated. As mentioned above, real characteristic features are represented by intervals of their values. Characteristic features and partitioning intervals for each of the 9 characteristic features represented by integers, are listed in the Table 1.

Table 1. The list of the characteristic features and their partitioning intervals.

Characteristic features	Code	Value intervals
Flow velocity (m/s)	$z_1$	1 – up to 0.5 inclusive; 2 – from 0.5 to 1; 3 – from 1 to 1.5; 4 – from 1.5 to 2; 5 – from 2 to 2.5; 6 – from 2.5 to 3; 7 – from 3 to 3.5 ; 8 – from 3.5 to 4; 9 – from 4 to 4.5; 10 – from 4.5 to 5; 11 – from 5 to 5.5; 12 – from 5.5 to 6; 13 – from 6 to 6.5; 14 – from 6.5 to 7;
Viscosity (mPa·s)	$z_2$	1 – from 0.2 to 0.5; 2 – from 0.5 to 0.8; 3 – from 0.8 to 1; 4 – from 1 to 1.2; 5 – from 1.2 to 1.5; 6 – from 1.5 to 1.8; 7 – from 1.8 to 2 ; 8 – from 2 to 2.5; 9 – from 2.5 to 3; 10 – from 3 to 4; 11 – from 4 to 5; 12 – from 5 to 8; 13 – from 8 to 10; 14 – from 10 to 20; 15 – over 20;
Density (kg/m <sup>3</sup> )	$z_3$	1 – from 550 to 580; 2 – from 580 to 610; 3 – from 610 to 640; 4 – from 640 to 670; 5 – from 670 to 700; 6 – from 700 to 730; 7 – from 730 to 760 ; 8 – from 760 to 790; 9 – from 790 to 820; 10 – from 820 to 850; 11 – from 850 to 880; 12 – from 880 to 910; 13 – from 910 to 940; 14 – from 940 to 970; 15 – from 970 to 1000;
Cross-sectional area (m <sup>2</sup> )	$z_4$	1 – up to 0.3 inclusive; 2 – from 0.3 to 0.5; 3 – from 0.5 to 0.7; 4 – from 0.7 to 0.9; 5 – from 1 to 1.2; 6 – from 1.2 to 1.4;
Element of hydraulic power unit	$z_5$	1 – circular pipe (smooth); 2 – flexible tubing; 3 – smooth concentric annulus; 4 – tap valve; 5 – dispensable slide-valve port; 6 – plate and poppet valves; 7 – strainer;
Temperature (°C)	$z_6$	1 – from 0 to 10; 2 – from 10 to 20; 3 – from 20 to 30; 4 – from 30 to 40; 5 – from 40 to 50; 6 – from 50 to 60; 7 – from 60 to 70; 8 – from 70 to 80; 9 – from 80 to 90; 10 – from 90 to 100;
Type of fluid	$z_7$	1 – water; 2 – sea water; 3 – oil; 4 – ether; 5 – alcohol; 6 – gasoline; 7 – kerosene;
Roughness (mm)	$z_8$	1 – 0.0001; 2 – 0.001; 3 – 0.006; 4 – 0.015; 5 – 0.017; 6 – 0.02; 7 – 0.025; 8 – 0.1; 9 – 0.15; 10 – 0.25;
Pressure (MPa)	$z_9$	1 – from 0.25 to 0.75; 2 – from 0.75 to 2.5; 3 – from 2.5 to 5; 4 – from 5 to 6.4;

It should be noted, that the lines of the **Q** matrix are associated with the fluid flow regimes and represent only a part of various combinations of characteristic features values.

Table 2 contains the classification features and their values for the **R** matrix.

Table 2. The list of classification features and their values.

Characteristic features	Code	Values
Fluid flow regime	$k_1$	1 – ideal; 2 – laminar; 3 – transient; 4 – turbulent;
Zones of turbulent fluid flow regime	$k_2$	1 – zone of hydraulically smooth pipes; 2 – zone of mixed friction; 3 – zone of square-law resistance.

### 3.2. An illustrating example of representing data and knowledge about the fluid flow regimes in a pipeline

Figure 2 gives an illustrating example of data and knowledge representing on the fluid flow regimes in a pipeline. The illustrating example is a fragment of matrix description of data and knowledge in hydraulics.

The description matrix (Fig. 2) contains 9 columns associated with the mentioned above characteristic features, and 15 rows filled out by us.

The rows of the **R** matrix are associated with the rows of the **Q** matrix, the columns are associated with the aforementioned characteristic features  $k_j$  ( $j \in \{1,2\}$ ). The  $r_{ij}$  element of the distinguishing matrix sets the belonging of the  $i^{\text{th}}$  object (fluid flow) to a class based on the  $j^{\text{th}}$  classification mechanism (fluid flow regime) by way of indicated the class number. A row of the **R** matrix sets the fluid flow regime and the zone of the turbulent regime for the studied field of hydraulics.

As mentioned above, the set of all non-repeating rows of the **R** matrix is associated with the set of selected patterns represented by the single-column **R'** matrix, whose elements are numbers of the patterns. This model does not permit intersecting objects from different patterns. Presence of such intersections is revealed via analysis using IS DFFRP.

Taking into consideration that with the limits of the paper do not allow us to fully present the matrix description of data and knowledge, Fig. 2 contains only a fragment of matrix data and knowledge representation on fluid flow regimes in a pipeline (**Q**, **R**, and **R'** matrices). The aforementioned fragment, which represents partial (only a part of characteristic feature space and its



values are used to determine the fluid flow regime) description of knowledge represented in the  $\mathbf{Q}$  matrix containing 9 columns and 15 rows, while the  $\mathbf{R}$  matrix – 2 columns.

									$k_1 \ k_2$						
$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	$z_6$	$z_7$	$z_8$	$z_9$							
$Q =$	1	5	15	2	1	5	1	1	1	$R =$	4	1	$R' =$	4	1
	1	11	1	2	1	7	3	1	1		2	–		2	2
	1	13	1	2	1	8	3	1	1		2	–		2	3
	2	12	1	2	1	6	3	1	1		3	–		3	4
	2	4	1	4	1	4	3	6	5		4	3		6	5
	2	2	2	4	1	4	3	6	5		4	3		6	6
	2	2	6	1	1	4	3	6	1		4	3		6	7
	2	2	5	1	1	4	3	1	1		4	1		4	8
	3	2	5	1	1	3	3	1	1		4	2		5	9
	3	1	5	1	1	6	3	1	1		4	2		5	10
	2	12	1	2	1	7	3	1	1		3	–		3	11
	1	13	1	2	1	8	3	1	1		3	–		3	12
	2	12	1	2	1	4	3	1	1		2	–		2	13
	3	3	1	3	1	6	3	1	1		4	1		4	14
	2	12	1	2	1	5	3	1	1		3	–		3	15

Fig. 2. Fragment of the description  $\mathbf{Q}$  matrix and the distinguishing  $\mathbf{R}$  and  $\mathbf{R}'$  matrix.

The fragment of the description and distinguishing matrices includes only 3 types of fluid flow regimes in the pipeline: laminar, transient, turbulent, and 3 aforementioned zones only for the turbulent regime.

The complete description and distinguishing matrices will be represented in the data and knowledge base of the IS DFFRP based on the fluid flow regimes in the pipeline, and the IS DFFRP is destined for prompt determination of various combinations of fluid flow regimes with the zones of the turbulent regime: 1) ideal flow regime; 2) laminar flow regime; 3) transient fluid flow regime; 4) turbulent flow regime, zone of hydraulically smooth pipes; 5) turbulent flow regime, zone of mixed friction; 6) turbulent flow regime, zone of square-law resistance. These 6 regimes are represented in the  $\mathbf{R}'$  matrix.

#### 4. Conclusion

Based on the performed analysis of modern-day state of research in determination of fluid flow regimes in a pipeline, for the first time it is proposed to create an intelligent system for diagnostics of the fluid flow regime in a pipeline designed to determine various regularities in parameters of hydrodynamics, which affect the fluid flow regimes, and also for decision-making and its justification on diagnostics of the fluid flow regime in a pipeline.

We also showed advisability of applying the matrix model for data and knowledge representation in the intelligent system for diagnostics of the fluid flow regime in a pipeline. The system is developed by us. For the first time, in accordance with the suggested model of representing data and knowledge, the feature space was formed; data and knowledge in the studied field were structured; characteristic and classification features were determined; real values of features were recoded into integer ones; the illustrating example representing partial description of data and knowledge in a matrix form was given. Herewith, in order to determine the fluid flow regime, we used only a part of the characteristic feature space and its values.

The IS DFFRP, which is being developed, will enable prompt and cheaper determination of the fluid flow regime in a pipeline, based on the hydrodynamic parameters under study: ideal; laminar; transient; turbulent, and also the zones of the turbulent regime: hydraulically smooth pipes; mixed friction; square-law resistance. A company serving the pipeline will obtain the ability to operative reaction to any changes in a pipeline and take the appropriate measures of managing the flow and eliminating the possibility of breakdowns.

The IS DFFRP, which we are developing, may be applied on direct appointment at various types of enterprises, for scientific purposes, as well as to train specialists in hydrodynamics.

#### Acknowledgment

The research was funded by RFBR grant (project No. 16-07-00859a).

#### Reference

- [1] SNiP 2.01.07-85 «Loads and Impacts». M.: Gosstroy of the USSR, 1985; 48 p. (in Russian)
- [2] SNiP 2.05.06-85 «Trunk pipelines». M.: Gosstroy of Russia, 2001; 86 p. (in Russian)
- [3] Brandt I. Multiphase Flow. Euroil, 1991: 62–63
- [4] Roache PJ. Computational Fluid Dynamics. M.: Mir, 1980; 616 p. (in Russian)
- [5] Yakovlev PV, Hodzhamuradova NB, Guba OE. Reynolds's analogy in heatmass exchange model at the turbulent mode of the convective movement of liquid in limited volume. Vestnik ASTU 2008; 6(47): 75–77. (in Russian)
- [6] Nikonova VT, Sautkina TN. Determination of coefficient of hydraulic friction upon transition from the laminar mode of the movement of liquid to turbulent. Improvement of methods of hydraulic calculations of water throughput and treatment facilities 2009; 1(35): 65–68. (in Russian)

- [7] Sargsyan NM. About the modes of the movement of viscous liquid in horizontally located round pipe under isothermal conditions. *Chemical technology* 2013; 2: 104–112. (in Russian)
- [8] Kudinov VI. Bases of oil and gas business. M.: IKI, 2005; 720 p. (in Russian)
- [9] Shammazov AM, Bakhtizin RN, Mastobayeva BN, Soshchenko AE. Pipeline transport of Russia. *Pipeline transport of oil* 2001; 2: 406 p. (in Russian)
- [10] Edwards DA, Gunasekera D, Morris J, Shaw G, Shaw K, Walsh D, Fjerstad PA, Kikani J, Franco J, Hoang V, Quettier L. Reservoir simulation: Keeping Pace with Oilfield Complexity. *Oilfield Review* 2012; 5(23): 4–15.
- [11] Aziz I, Brandt I, Gunasekera D, Hatveit D, Havre K, Weisz G, Xu ZG, Nas S, Spilling KE, Yokote R, Song S. Multiphase flow simulation-optimization field productivity. *Oilfield Review* 2015; 1(27): 26–37.
- [12] Yankovskaya AE, Gedike AI, Ametov RV, Bleikher AM. IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition. *Pattern Recognition and Image Analysis* 2003; 13: 4: 650–657.
- [13] Yankovskaya AE, Il'inskikh NN. On the Question of the Development and Application of Intelligent Biomedical Systems. *Pattern Recognition and Image Analysis* 1998; 8(3): 470–472.
- [14] Yankovskaya AE. An Automaton Model, Fuzzy Logic, and Means of Cognitive Graphics in the Solution of Forecast Problems. *Pattern Recognition and Image Analysis* 1998; 8(2): 154–156.
- [15] Rocher G, Brattstrum A, Gho S, Francenson F, Hintricus H, Mauser R, Packianather M, Pogrzeba G, Yankovskaya A, Zvegintsev V. Real-time Recognition of ECG by Using Powerful Information and Communication Technology for Intelligent Monitoring of Risk Patients. *Application, Trends, Visions, VDE World Microtechnologies (MICRO.tech)*. Proceedings of International Congress. Hannover, Germany, 2000; 2: 759–762.
- [16] Ryumkin A, Yankovskaya A. Intelligent Expansion of the Geoinformation System. The 6th German-Russian Workshop "Pattern Recognition and Image Understanding" OGRW-6-2003. Workshop proceedings. Novosibirsk, Russia, 2003; 202–205.
- [17] Yankovskaya AE, Chernogoryuk GE, Muratova EA. Intelligent Test Recognizing Biomedical System. The 6th German-Russian Workshop "Pattern Recognition and Image Understanding" OGRW-6-2003. Workshop proceedings. Novosibirsk, Russia, 2003; 248–251.
- [18] Yankovskaya AE, Semenov ME. Intelligent system for knowledge estimation on the base of mixed diagnostic tests and elements of fuzzy logic. Proceedings of the IASTED International Conference Technology for Education. December 14–16, Dallas, USA, 2011; 108–113.
- [19] Yankovskaya AE, Kitler S. Mental Disorder Diagnostic System Based on Logical-Combinatorial Methods of Pattern Recognition. *Computer Science Journal of Moldova* 2013; 21(3-63): 391–400.
- [20] Yankovskaya AE, Yamshanov A. Bases of intelligent system creation of decision-making support on road-climatic zoning. *Pattern Recognition and Information Processing (PRIP'2014)*: Proceedings of the 12th International Conference (28–30 May 2014, Minsk, Belarus). Minsk : UIIP NASB, 2014; 311–315.
- [21] Yankovskaya AE, Efimenko S, Cherepanov D. Structurization of data and knowledge for the information technology of road-climatic zoning. *Applied Mechanics and Materials* 2014; 682: 561–568.
- [22] Yankovskaya AE, Demytyev Y, Lyapunov D, Yamshanov A. Intelligent Information Technology in Education. *Information Technologies in Science, Management, Social Sphere and Medicine (ITSMSSM-2016)*. Atlantis Press Publishing, 2016: 17–21.
- [23] Yankovskaya AE, Gorbunov I, Hodashinsky I, Chernogoryuk G. On a Question of the Information Technology Construction Based on Self-learning Medicine Intelligent System. *Information Technologies in Science, Management, Social Sphere and Medicine (ITSMSSM-2016)*. Atlantis Press Publishing, 2016: 22–28.
- [24] Yankovskaya AE, Shelupanov AA, Mironova VG. Construction of Hybrid Intelligent System of Express-Diagnostics of Information Security Attackers Based on the Synergy of Several Sciences and Scientific Directions. *Pattern Recognition and Image Analysis* 2016; 26(3): 524–532.
- [25] Yankovskaya AE. Logical Tests and Cognitive Graphics Means in Intelligent System. Proc. 3<sup>rd</sup> All-Russian Conf. with International Participation "New Information Technologies in Discrete Structures Research" (Izd. SO RAN, Tomsk), 2000: 163–168. (in Russian)
- [26] Birkhoff G. *Hydrodynamics: A study in logic, fact, and similitude*. M.: Foreign literature, 1954; 184 p. (in Russian)
- [27] Loytsyansky LG. *Mechanics of liquid and gas*. M.-L.: Gostekhizdat, 1950; 479 p. (in Russian)
- [28] Ayala LF, Adewumi MA. Low-Liquid Loading Multiphase Flow in Natural Gas Pipelines. *Journal of Energy Resources and Technology* 2003; 4(125): 284–293.
- [29] Li C, Liu E, Yang Y, Najafi M, Ma B. The simulation of steady Flow in Condensate Gas Pipeline. *Advances and Experiences with Pipelines and Trenchless Technology for Water, Sewer, Gas, and Oil Applications*. Reston, Virginia, USA, American Society of Civil Engineers, 2009: 733–743.



# Prediction of Cluster System Load Using Artificial Neural Networks

Y.S. Artamonov<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

Currently, a wide range of high-performance environments is available for a researcher to perform computations. It is a really difficult task to select an environment in which the computations will be completed as soon as possible. To solve this, you need to analyze the load of the environment computing resources, and also to predict their availability in the future.

In this paper we describe a solution for prediction of computing resources load in a cluster environment using neural network models. We considered a process of configuring the neural network architecture: selection of activation functions, algorithms of initialization and updating of the weights of neurons. Training and testing was performed on a set of data for the load of the cluster "Sergey Korolev" for the period from November 2013 to December 2016.

*Keywords:* load prediction; cluster; neural network; model

---

## 1. Introduction

Recently, many studies have been devoted to forecasting the load of various computational resources, such as CPU cores [1], individual nodes of a cluster or clouds [2]. Load prediction in cloud and cluster environments is a critical problem that needs to be solved to achieve high performance, since a lot of processes depend on its effective solution, such as resources planning, maintenance periods and modernization of machines and even whole data centers. For instance, without prediction of the availability of shared resources it is impossible to effectively use classic cluster environments where users use shared nodes with different performance and features taking them partially or completely by their computations.

In the previous paper [3] we solved the task of forecasting the load of computational resources using the EMMSP model and examined the applicability of this model. As we noticed, the model is well suited for predictions only on specific data and load history points, but also we showed that it can be effectively used as a component of a simple model mixture: adaptive selection and adaptive composition. This paper considers the use of neural network models to predict the load of cluster resources and compares this approach with that demonstrated previously.

## 2. Neural network prediction models

Neural network prediction models are based on the use of neural networks that can be trained in regression problems and produce the output value, based on some input parameters, approximating the unknown functional dependencies of the output data on the input.

Neural network models were used in papers [4] and [5] to predict the load of resources with different nature: CPU servers and electrical networks. In both problems, neural network models showed good results and were recognized as effective and adequate to the prediction problem. Given these results, let's look at how well neural network models are suitable for predicting the number of loaded cluster nodes.

Neural network models were chosen for this study because of peculiarities of the task and historical data collected by us. We took into account the following aspects:

- time series of resources load are non-stationary,
- there are templates and periodic components in historical data, as well as segments with low and high load, corresponding to weekends, holidays and work days,
- time series have known minimal and maximum value.

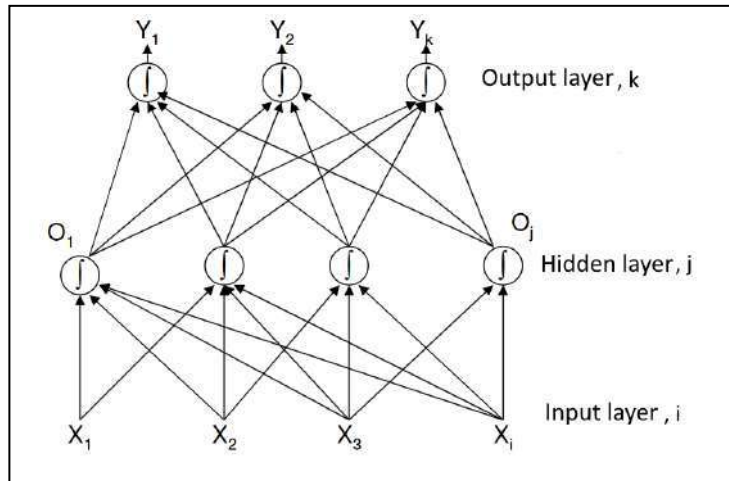
To predict values of time series of this nature we can use neural networks, in fact solving the approximation problem of an unknown function. Taking into account the papers [6] and [7] that apply neural networks for solving forecast problems of similar in nature time series (load of computational resources of cluster / cloud environments), we chose to study the model of a multilayer perceptron (MLP) with single (SL MLP) and two hidden layers (DL MLP).

MLP consist of neurons and connections between them (fig. 1). Neurons have a special transformation function – activation function, each connection has characteristic called weight. Output signal of a neuron in a layer  $Z$  of MLP is determined using equation 1 [8]:

$$z_j = f\left(\sum_{i=1}^N w_{ij}u_i\right) \quad (1)$$

where  $u_i$  – output signals of the layer  $Z$ ,  $w_{ij}$  – weights of connections between  $i$  neuron of the previous layer and  $j$  neuron of the layer  $Z$ ,  $f$  – activation function,  $z_j$  – output signal of a neuron. In this paper, we used hyperbolic tangent function as an activation function for neurons of hidden layers.

The training of a neural network is a process of changing the weights of the neuron connections. The main goal of a learning algorithm is to find a configuration of the weights of all the links where the error function is minimized. In the task that we solve we use MSE (Mean Squared Error) as a criterion for model training using gradient descent method, as a final benchmark of a model we use MAE (Mean Average Error) since time series contains a lot of segments with zero value or sequential equal



values, that is why we cannot use MASE (Mean Average Scaled Error) and MAPE (Mean Absolute Percentage Error).

Fig. 1. Structure of MLP with one hidden layer.

We use DeepLearning4j library for training and testing models based on MLP. This library provides battle proven tools and algorithms for training and usage of various artificial neural networks, including the most popular neural network architectures, learning and optimization algorithms. The library is written in Java and uses native extensions for computations on the CPU and GPU to provide the required performance [9]. The DeepLearning4j library is licensed under the Apache License 2.0, this enables us to use it in any applications including commercial, the open source development approach attracts a large number of researchers and improves the quality of the library.

### 3. Configuring network architecture and learning parameters

To train neural MLP networks the method of back propagation of the error is used with various modifications. The method is an iterative gradient algorithm that is used to minimize the MLP error and to obtain the desired output values. The essence of the method consists in propagation of error signals from the outputs of the network to its inputs, back to direct propagation of signals in the usual mode of operation [10].

Primary parameters of the method and its modifications are:

- learning epochs count,
- learning rate,
- weights initialization algorithm,
- weights update algorithm,
- optimization algorithm,
- learning momentum.

In the task, we need to predict the number of occupied cluster nodes in several of the most intensively used groups of nodes. Target prediction interval – 12 hours, we need to predict 12 points of a time series, one mean value of cluster group load per one hour. We chose qdr\_tmp and ddr\_tmp cluster groups for training and prediction of a group load. Their load is of the greatest interest because of a large regular load.

To compare the learning methods with different modifications we chose the training parameters presented in Table 1. We compared the training of SL MLP and DL MLP models in the prediction task with 12 points of cluster load (each point – mean load of a cluster group for 1 hour). In training and forecasting, only time series data were considered and passed to neural network inputs. The optimal number of inputs, selected experimentally, is 6.

In the process of training, we fed to the input of the neural network various sets of consecutive 6 values of the series; we used random order of data sets. For each test set, 12 values were generated at the output of the neural network, which were compared with 12 values from the test set. Parameters  $i = 6$ ,  $k = 12$ .

Table 1. Experimental learning parameters.

Model	Learning epochs	Learning rate	Momentum	Inputs count	Hidden layer neurons count
SL MLP	300	0.01	0.9	6	15
DL MLP	400	0.01	0.9	6	1st: 20 2nd: 10

The backward propagation of errors method is subject to the following problems:

- slow convergence,
- convergence to local minima,
- overfitting.

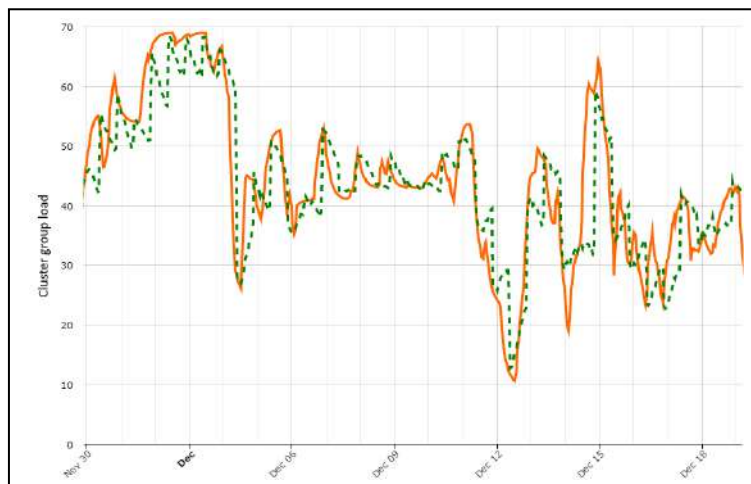


Fig. 2. The forecast of resources load of the cluster "Sergey Korolev" using SL MLP model.

Modifications to the method of back propagation of errors with momentum and various updating algorithms of the link weights, such as Adadelta, enable us to fix or partially fix the above problems, accelerate training and reduce the error of MLP based prediction models.

In the study of neural network models we considered various configurations of training neural networks by the method of back propagation of errors. As parameters of the configuration in the training we used: the algorithm for initializing the weights of neurons, the algorithm for updating the weights of neurons and the optimization algorithm.

We tested 2 options for initializing the balance: uniform distribution (Uniform) and using the Xavier method. The following algorithms for updating the weights of neurons were tested: Nesterov Accelerated Gradient (Nesterovs), adaptive gradient descent (Adagrad), Adaptive learning rate (Adadelta), adaptive momentum estimation (Adam). Two optimization algorithms were tested: linear gradient descent (LGD) and stochastic gradient descent (SGD). These optimizations and the parameters of the gradient descent method and the back propagation of errors algorithm are described in paper [11].

The test used a training sample of length 6000 points and a test sample with a length of 1000 points, the sample data were obtained for the period from January 1, 2015 to January 1, 2016. The results of testing models with different learning parameters for solving the task of forecasting the cluster load are presented in Table 2, the RMSE (Root Mean Square Error) error values are given to estimate the dispersion of the forecast values.

Table 2. RMSE and MAE error values depending on neural network training configuration.

Weights initialization	Weights updater	Optimization algorithm	SL MAE	DL MAE	SL RMSE	DL RMSE
UNIFORM	NESTEROVS	LGD	<u>8,03</u>	7,99	9,28	9,24
XAVIER	NESTEROVS	LGD	8,05	8,20	9,30	9,38
UNIFORM	NESTEROVS	SGD	<u>8,02</u>	<u>7,64</u>	9,28	8,94
XAVIER	NESTEROVS	SGD	8,06	7,70	9,32	8,97
UNIFORM	ADADELTA	LGD	9,71	11,15	10,99	12,11
XAVIER	ADADELTA	LGD	9,62	11,77	10,78	12,75
UNIFORM	ADADELTA	SGD	15,09	8,33	16,08	9,86
XAVIER	ADADELTA	SGD	17,44	12,52	18,61	14,64
UNIFORM	ADAGRAD	LGD	13,24	11,52	13,61	12,64
XAVIER	ADAGRAD	LGD	15,24	9,70	16,36	12,13
UNIFORM	ADAGRAD	SGD	19,04	10,60	21,36	19,23
XAVIER	ADAGRAD	SGD	17,21	11,21	18,12	17,22
UNIFORM	ADAM	LGD	8,10	7,83	9,34	9,13
XAVIER	ADAM	LGD	8,14	7,91	9,38	9,18
UNIFORM	ADAM	SGD	<u>7,99</u>	<u>7,54</u>	9,26	8,84
XAVIER	ADAM	SGD	8,09	<u>7,56</u>	9,34	8,85

Table 2 shows the results of testing the modifications of the method of back propagation of errors, the 3 best results for each model are highlighted with underscores. From these results, we can conclude that the most effective modifications of the back propagation of errors method for the task of forecasting the cluster load are:

1. stochastic gradient descent with initialization of weights using uniform distribution and updating of weights using the ADAM algorithm – for SL MLP and DL MLP models,
2. linear gradient descent with initialization of weights using uniform distribution and updating of weights using the Nesterov method with momentum – for SL MLP model,
3. stochastic gradient descent with initialization of weights using the Xavier method and updating of weights using the ADAM algorithm – for DL MLP model.

The results of DL and SL MLP models differ slightly, which is probably due to the peculiarity of the test data.

#### 4. Comparison of model errors

An example of forecasting cluster load data for a neural network with a single hidden layer is shown in fig. 2, a neural network with two hidden layers - in fig. 3. The dashed line shows the forecast values of the series. The graphs of the forecast values were obtained by calculating the forecast every 12 points.

As the final error metric, the mean absolute error (MAE) is selected, because the relative forecast error (MAPE) can not be used in series that include values close to or equal to zero. The distribution of MAE errors in the SL MLP and DL MLP models is shown in fig. 4, the distribution of errors of both models is close to normal.

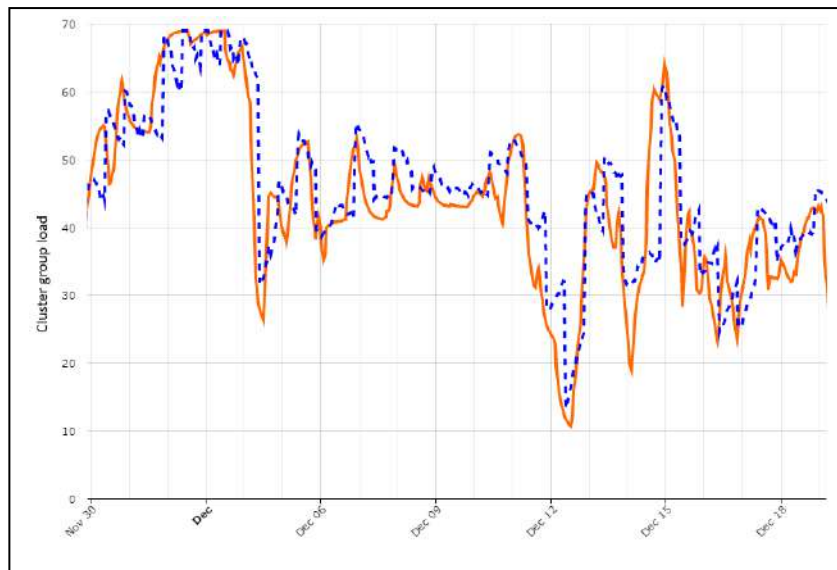


Fig. 3. The forecast of resources load of the cluster "Sergey Korolev" using DL MLP model.

Previously, the task of forecasting 12 cluster load points was solved by the time series prediction method using the maximum resemblance sample (EMMSP) [3]. The MAE prediction errors for method comparison are given in Table 3.

In addition to direct comparison of models, we tried to use all three models (EMMSP, SL MLP, DL MLP) together. In order to do this, we put forward a hypothesis: Each of the models is the best (shows the smallest MAE error) in a certain length of data  $L > M$ , where  $M$  is the number of prediction points. We tested this hypothesis for the data on which MLP models were tested. Each of the models retains its leadership at the average on a section of 24 to 36 points in length, which corresponds to a time interval of 1 to 1.5 days.

Table 3. MAE errors of different prediction models.

Model	EMMSP	SL MLP	DL MLP	Simple adaptive selection
MAE	8.7	8.02	7.54	6.8

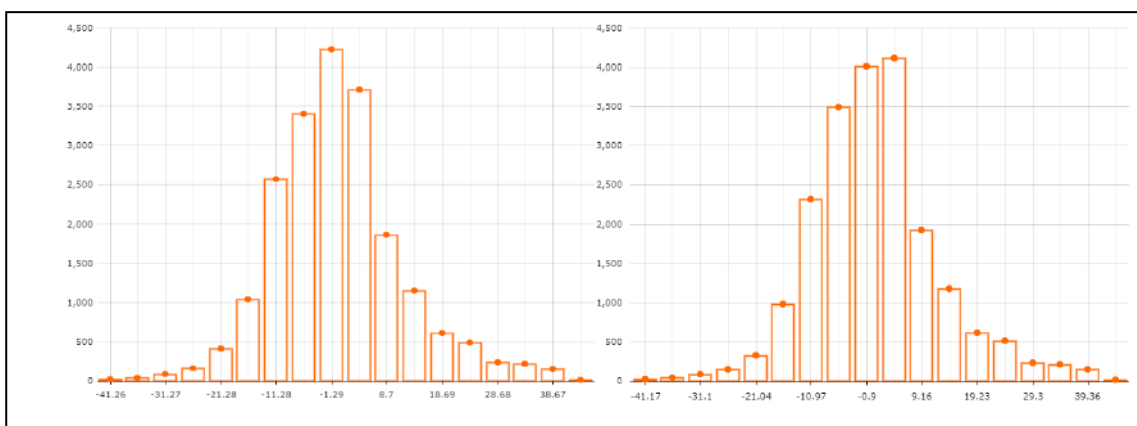


Fig. 4. The distribution of the mean absolute error of models with one hidden layer (on the left) and two hidden layers (right).

The error value for a simple adaptive selective model [12] was obtained for a model that selects the best model for predicting future values by a simple heuristic rule: If one of the models was better in the previous section of the data, then it should be used to predict again. Data for testing were collected between November 2013 and December 2016. The open load monitoring data of the "Sergey Korolev" cluster is available in JSON machine-readable format at: [http://templet.ssau.ru/wiki/открытые\\_данные](http://templet.ssau.ru/wiki/открытые_данные).

## 5. Conclusion

The prediction algorithms based on neural network models with one and two hidden layers are integrated into the Templet Web service, which enables users to estimate the task launch time. The forecast graphs and cluster load history are available to registered users of the system. In the future, we plan to provide users with an interactive hint about the number of available resources and the estimated time to start the task based on the task requirements (nodes, groups, software licenses) specified at the time of adding task to a batch queue.

The results of the cluster load forecasting can be applied to solve several types of tasks:

- increase the efficiency of cluster use (energy efficiency, load efficiency),
- selection of optimal environments and parameters for computations,
- planning of cluster growth and maintenance periods.

Methods of forecasting the loading of computing resources are most in demand now in cloud environments where they can enable commercial companies to reduce server maintenance costs or, on the contrary, to effectively adapt to the growing demands of customers.

## Acknowledgements

This work is partially supported by the Russian Foundation for Basic Research (RFBR#15-08-05934-A), and by the Ministry of Education and Science of the Russian Federation within the framework of the State Assignments program (№ 9.1616.2017/ПЧ).

## References

- [1] Naseera S, Rajini GK, Sunil Kumar Reddy P. Host CPU Load Prediction Using Statistical Algorithms a comparative study. *International Journal of Computer Technology and Applications* 2016; 9(12): 5577–5582.
- [2] Di S, Kondo D, Cirne W. Host load prediction in a Google compute cloud with a Bayesian model. *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press, 2012; 21 p.
- [3] Artamonov YS. Application of the EMMSP model to predict available computing resources in cluster systems. *Bulleten of the Samara Scientific Center RAS* 2016; 18(4): 681–687. (in Russian)
- [4] Naseera S, Rajini GK, Amutha Prabha N, Abhishek G. A comparative study on CPU load predictions in a computational grid using artificial neural network algorithms. *Indian Journal of Science and Technology* 2015; 8(35).
- [5] Kalaitzakis K, Stavrakakis G, Anagnostakis EM. Short-term load forecasting based on artificial neural networks parallel implementation. *Electric Power Systems Research* 2002; 63(3): 185–196.
- [6] Chandini M, Pushpalatha R, Boraia R. A Brief study on Prediction of load in Cloud Environment. *International Journal of Advanced Research in Computer and Communication Engineering* 2016; 5(5): 157–162.
- [7] Engelbrecht HA, van Greunen M. Forecasting methods for cloud hosted resources, a comparison. *Network and Service Management (CNSM)*. 11th International Conference on IEEE 2015; 29–35.
- [8] Hajkin S. *Nejronnye seti*. M.: Vil'jams, 2006; 1104 p.
- [9] DeepLearning4j: Open-source distributed deep learning for the JVM. URL: <http://deeplearning4j.org> (01.01.2017).
- [10] Osovskij S. *Nejronnye seti dlja obrabotki informacii*. M.: Finansy i statistika, 2002; 344 p.
- [11] Ruder S. An overview of gradient descent optimization algorithms, 2016. ArXiv preprint arXiv: 1609.04747.
- [12] Lukashin JuP. Adaptive methods of short-term forecasting of time series. M.: Finansy i statistika, 2003; 415 p.

# Network disruption prediction based on neural networks

D.S. Taimanov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

Network disruptions cause significant financial losses and discomfort of customers. However, communication systems provide various data about equipment condition. This information can be used to predict network disruptions. Purpose of research is applying neural networks to network disruption prediction. Introduced approach has good feature extraction ability and data corruption resilience. Representation for network disruption dataset has been developed. Chosen network type is deep belief networks. Net structure variations have been proposed for chosen data representation and neural network type. Experimental research of proposed methods gives meager results for selected dataset. Results can be improved by net structure complication and by increasing dataset volume.

*Keywords:* data mining; neural networks; prediction; network disruptions; big data

## 1. Introduction

Network disruptions cause financial losses and inconveniences. They are increasing with the growth of telecommunication influence on all spheres of life. Communication systems amplification and data science advancement allow predicting network disruptions. Communication systems provide various data about equipment condition. Data science gives different methods of prediction. Network disruption prediction involves continuous network condition monitoring and disruption hazard evaluation.

It is difficult to find present task in previously published works. So tasks with similar data types have been observed. Network disruptions data values mostly belong to enumerated unordered type. Similar data types appear in work [1]. Authors used deep belief networks to identify risk factors and predict bone disease progression.

There is an interesting approach to use neural networks for telecommunication disruptions prediction. “Telstra Network Disruptions” dataset [2] is destined for disruption prediction.

## 2. Network equipment condition data

“Telstra Network Disruptions” dataset consists of records about events. Each event described by attributes: *location*, *event\_type*, *resource\_type*, *severity\_type*, *log\_feature*. The goal of competition [2] is distinguish each event between different kinds of *fault\_severity*. There are three kinds (values) of *fault\_severity*: 0 – normal operation, 1- momentary glitch, 2 – total interruption of connectivity. Each attribute described in table 1.

Table 1. “Telstra Network Disruptions” dataset structure.

Attribute name	Number of different values, $N_i$	Type	Repetition, values/event	Description
id	7381	integer, unordered	-	Identifier, unique. Intended for data union.
<b>fault_severity</b>	3	integer, ordered	-	Class of event, <b>forecast objective</b> .
location	929	enumerated, unordered	-	Network equipment location.
event_type	49	enumerated, unordered	12468/7381	Type of event.
resource_type	10	enumerated, unordered	8460/7381	Type of resource that caused event.
severity_type	5	enumerated, unordered	-	Type of log message
log_feature	331	tuple: <i>type</i> – <i>volume</i> <i>type</i> : enumeration, unordered <i>volume</i> : integer, ordered	23851/7381	Features extracted from logs with its volume

As we can see, all significant attributes belong to enumerated (categorical) unordered type. This attributes shows record affiliation with some of class by some of parameter. For example, “location 1”, “event\_type 15”, “resource\_type 8”, “severity\_type 2”. *log\_feature* attribute is a tuple “type of value, volume of value” (e.g. “feature 35, 17”).

Attributes are kept in separate tables distributed by different CSV-files. This manner of data storage is used because of attribute values repetition. For example, single event can be of “event\_type 15” and “event\_type 11” at the same time. *id* attribute is used for table matching. Table structure is represented in figure 1 a.

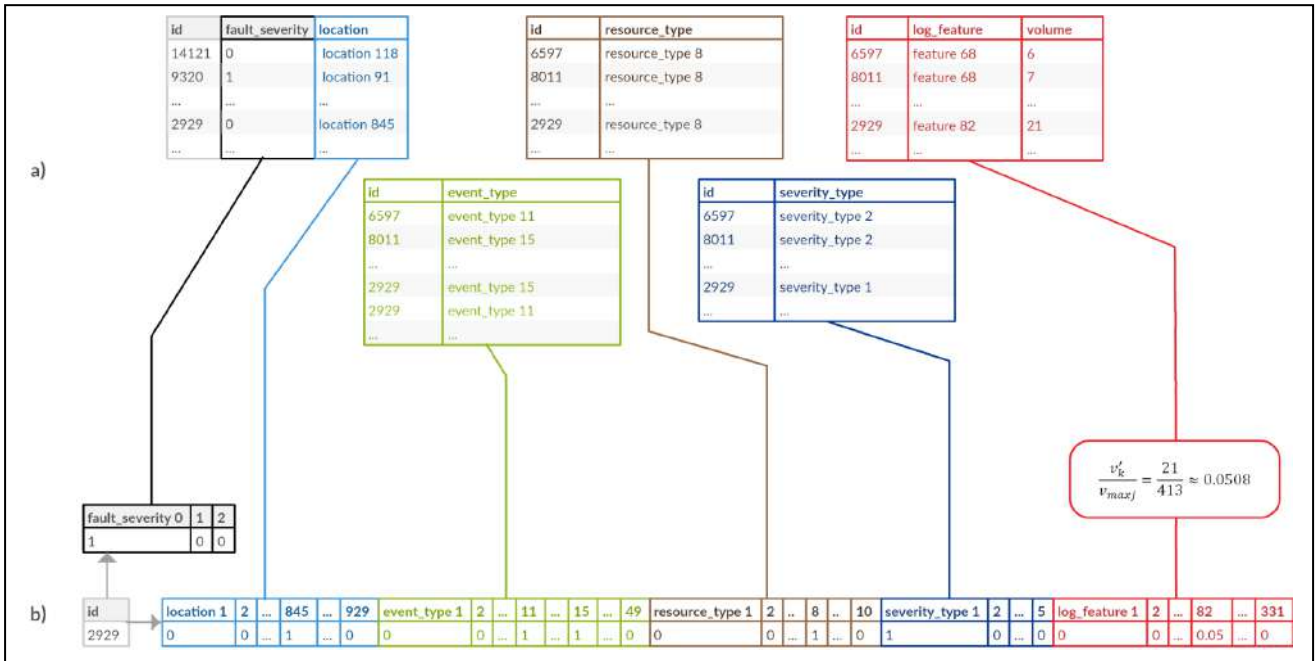


Fig. 1. Dataset structure (a) and input vector representation (b).

We have to make representation of event to construct input vector. Each attribute translates into a sparse vector. This vector has a corresponding index for each possible attribute value. Sparse vector has “1” at corresponding index if attribute possess this value for current event and “0” otherwise (figure 1 b). This is the simplest way to construct input vector for unordered categorical types.

Thus, input vector parts for *location*, *event\_type*, *resource\_type* and *severity\_type* are as follows:

$$v_i = (x_1, \dots, x_{N_i}).$$

Here  $N_i$  is the amount of  $i$ -th attribute values;  $i = \overline{1, M}$ ,  $M$  – amount of categorical attributes;  $x_j$  is as follows:

$$x_j = \begin{cases} 1, & \text{if } X_{ij} \in f_i \\ 0, & \text{otherwise} \end{cases}$$

Here  $f_i$  are values of  $i$ -th attribute for current event;  $X_{ij}$  – value of  $i$ -th attribute corresponding to  $j$ -th index.

*fault\_severity* input vector part constructs in the same way.

Values for *log\_feature* are different from values for other attributes. They not only appeared or not appeared for current event but have volume. Corresponding elements of input vector possess this volume (normed) instead of “1” if present. Thus, last equation is as follows for *log\_feature* attribute:

$$x_j = \begin{cases} \frac{v'_k}{v_{maxj}}, & \text{if } Y_j = f_k \\ 0, & \text{otherwise} \end{cases}$$

Here  $v'_k$  is volume of  $k$ -th appeared feature;  $f_k$  –  $k$ -th appeared feature («type» in tuple);  $Y_j$  – value of *log\_feature* corresponding to  $j$ -th index,  $j = \overline{1, N_Y}$ ;  $N_Y$  – possible *log\_feature* values amount; где  $v_{maxj}$  – maximum value for  $j$ -th feature.

Thus, input vector  $V$  is a sequence of attribute vector parts  $v_i$ , normed and has length  $N$ :

$$N = \sum_{i=1}^M N_i + N_Y.$$

$N$  is 1324 for “Telstra Network Disruptions” dataset. This is a large value. So, deep neural networks have been used.

### 3. Deep neural networks

Deep belief networks [3] have been chosen. This type of deep neural networks has good hidden feature extraction ability and can be fast trained and fine-tuned for high-dimensional datasets. Deep belief networks (DBN) is a composition of restricted Boltzmann machines (RBM) [4]. RBMs stacks layer by layer to construct DBN (figure 2).



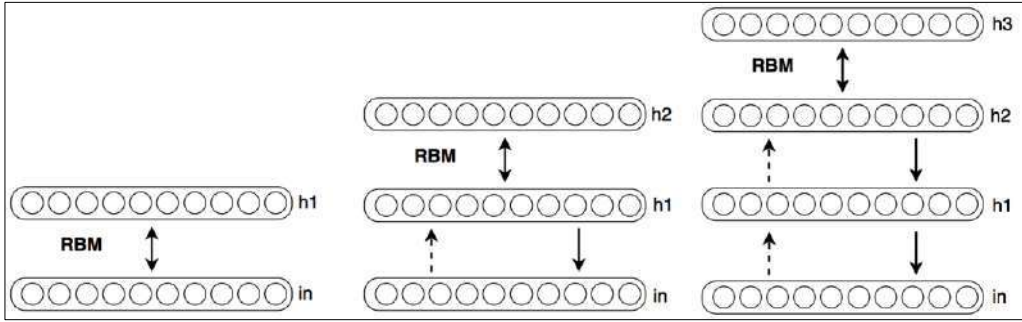


Fig. 2. Deep belief network construction.

Boltzmann machine (figure 3 a) is a stochastic machine formed by stochastic neurons [5]. This net fully connected. Neurons are divided into visible and hidden. Visible neurons are clumped onto values from dataset during the training. Hidden neurons always operate freely. These neurons can capture high order statistical correlations in the clumped values [5].

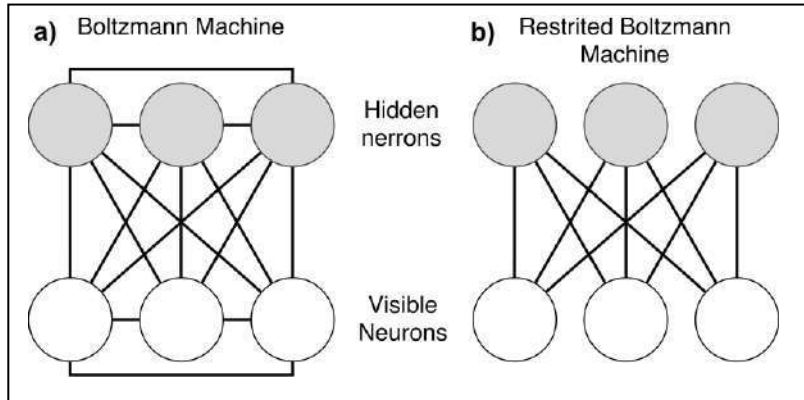


Fig. 3. Example of a figure caption.

Boltzmann machine has one significant disadvantage: too long time of training. It caused by multiple neuron activation. Restricted Boltzmann machine avoid this problem by removing connections between neurons with the same type (figure 3 b). RBMs can be fast trained at practice [4].

Algorithm [6] of training DBNs are shown below (figure 2):

- 1) Train first two layers (h1, h2) as RBM using dataset.
- 2) Pass dataset through trained RBM and get values from layer “h2”. Values got from “h2”layer are a new dataset.
- 3) Train next pair of layers (h2, h3) using new dataset, then repeat step 2 and 3 for next layers and so on until last hidden layer trained. Previous steps are called “pretraining phase”.
- 4) Train whole net using backpropogation or another common algorithm to fine-tune weights. This step is called “training phase”

Set of possible architecture variations forms during research:

- Various types of nets (DBN, Perceptron)
- Different layers count
- Various layer sizes

Neural nets have 1324 neurons at input layer and 3 neurons at output layer for “Telstra Network Disruptions” dataset. This dataset can be divided into train and test parts using three different ways:

- 1) Random division.
- 2) Division for each *location*. Events for each location are divided into train and test parts separately.
- 3) Division by *location*. All events for certain location occur in one of sets entirely.

#### 4. Results and Discussion

Regular accuracy ranking function has been used. It is a right-to-total ratio:

$$A = \frac{P}{N}$$

Here  $A$  is accuracy;  $P$  is a count of correct predictions;  $N$  – total predictions count.

Two loss functions have been tried: mean square error and weighted mean square error.

$$f(\mathbf{y}) = \frac{1}{I} \sum_{i=1}^I (\hat{y}_i - y_i)^2.$$

Here  $f(\mathbf{y})$  – mean square loss function;  $\mathbf{y}$  – vector of predictions;  $y_i$  – likelihood of  $i$ -th class,  $y_i \in \overline{0,1}$ ;  $\hat{y}_i$  – observed value for  $i$ -th class;  $I$  – number of classes.



Mean square function has one significant disadvantage for “Telstra Network Disruptions” dataset: considerable class imbalance often causes degenerate models construction. These models classify any input vector into one and the same class, the largest class.

Described problem can be solved via using a weighted mean square loss function:

$$f_w(\mathbf{y}) = \frac{1}{I} \sum_{i=1}^I ((\hat{y}_i - y_i))^2 \times c_i.$$

Here  $f_w(\mathbf{y})$  – weighted mean square loss function,  $c_i$  – significance of prediction error for  $i$ -th class.

Setting  $c_i$  inversely proportional to class occurrences number helps to solve the problem.

Best accuracy values for different networks consist of two and three hidden layers are listed in table 2.

Table 2. Best accuracy for different neural networks.

Network type	Error measure dataset	Pretraining	Hidden layers count	
			2	3
Deep Belief Network	Train	Usual	0.901	0.913
	Test	Usual	0.745	0.756
Perceptron	Test	Extended	0.749	0.757
	Test	-	0.749	0.751

Two types of networks have been considered: deep belief networks and perceptron (as a comparison). Accuracy and loss function have been measured via train or test selection. Additional data from test part of source dataset has been used in one of experiments.

Deep belief networks reach better results in comparison with perceptron for selected dataset and train vector representation.

Accuracy is increasing with the growth of hidden layers count. Therefore, it is reasonable to try networks that are more complex.

Additional data usage causes accuracy growth. It shows that better results can be reached for bigger datasets.

## 5. Conclusion

There is an interesting approach to use neural networks for telecommunication disruptions prediction. It is reasonable because of high dimension of input data. Deep belief networks have been selected. This type of deep neural networks has good hidden feature extraction ability and can be fast trained and fine-tuned for high- dimensional datasets.

Best reached accuracy is 75.7 % of correct predictions. This result can be improved by net structure complication. Bigger dataset usage also can help.

## References

- [1] Li H, Li XY, Ramanathan M. Identifying informative risk factors and predicting bone disease progression via deep belief networks. *Methods* 2014; 69(3): 257–265.
- [2] “Telstra Network Disruptions” competition. URL: <https://www.kaggle.com/c/telstra-recruiting-network> (01.02.2017).
- [3] Hinton GE. Deep belief networks. URL: [http://www.scholarpedia.org/article/Deep\\_belief\\_networks](http://www.scholarpedia.org/article/Deep_belief_networks) (01.02.2017).
- [4] Hinton GE. Boltzmann\_machine. URL: [http://www.scholarpedia.org/article/Boltzmann\\_machine](http://www.scholarpedia.org/article/Boltzmann_machine) (01.02.2017).
- [5] Haykin SS. Boltzmann machine. *Neural Networks: A Comprehensive Foundation* 1999; 11: 584–491.
- [6] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Computation* 2006; 18: 1527–1554.

# Automated system for modeling traffic of multiservice networks

B.Ya. Likhtsinder<sup>1</sup>, A.V. Kharkovsky<sup>1</sup>, S.Yu. Antsinov<sup>1</sup>

<sup>1</sup>Povolzhsky State University of Telecommunications and Informatics, LevTolstoy street, 23, 443010, Samara, Russia

---

## Abstract

The paper deals with the system for modeling parameters of traffic of multiservice networks, developed by the authors on the basis of Visual Studio, in C#. The system is based on the principles of interval method of flow analysis general systems of mass service. The results of comparison with similar products are presented. Software structural scheme and its performance are analyzed with examples of video and Poisson traffic stream. The possibility of approximation coefficients determination is given which are coefficients of generalized Khinchin-Pollaczek formula. As a result, for real traffic, average size of the queues were determined by using approximation coefficients applications at different load factors, as well as a number of other characteristics of multiservice traffic, such as dispersion, correlation, probability distribution.

*Key words:* modeling; multi-service network; generalized Khinchin-Pollaczek formula; traffic analysis; packets

---

## 1. Introduction

The traffic of multiservice communication networks with packet switching is rather nonuniform [1], [2], [4]. The incoming packets are grouped at one time and virtually absent in others. The features of distribution function of the number of applications on the time intervals for the flow of multiservice networks were examined previously [7], [9]. It was shown that there are periods with different activity, which alternate time after time with different appearance probabilities, due to the fact that the number of claims in multiservice networks is often irregular. In each period there is only one flow. The lack of claims corresponds to the period with zero activity.

All the results given in [5] were obtained with the corresponding program, written in MatLab. The software, developed in the MatLab system, is well suited for scientific research; it is also suitable for analysis of traffic in real networks. Using the experience gained previously, the authors developed a system that allows you to quickly and accurately determine the main characteristics of traffic of multiservice communication networks, suitable for analysis of queues.

## 2. Software structure

The structural scheme of Fig. 1 shows the algorithm of the program.

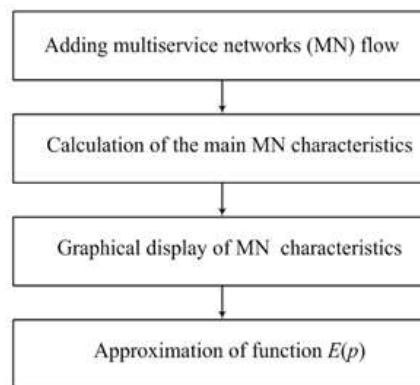


Fig.1. Software structure.

The first step is a user input file that contains information about the arrival of the packets meanwhile the operation of a multiservice network (obtained, for example, using the WireShark program [10]).

The second step is processing the file. The main traffic characteristics (time delays between packets, the expected number of packets in the queue at different load factors, etc.) are calculated. The next step is to build selected characteristics in the graphics area of the software window

At the last stage, approximating the numerator function of the generalized Khinchin-Pollyaczek formula, a user can determine coefficients that characterize this traffic, examined in [5].

## 3. Analysis of software correctness

The basis of this software lay down the methods of analysis and processing algorithms of traffic, given in [3], [8], [9]. To assess the correct operation of software, a comparison was made between previously defined video streaming characteristics and characteristics obtained from the analysis using the developed system. Below are the graphs obtained using developed system analysis.

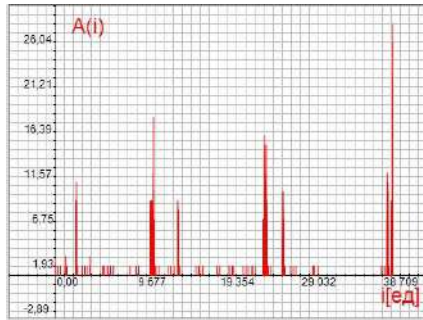


Fig.2. The number of packets at the intervals  $\tau$  when the load factor  $\rho = 0.1$ .

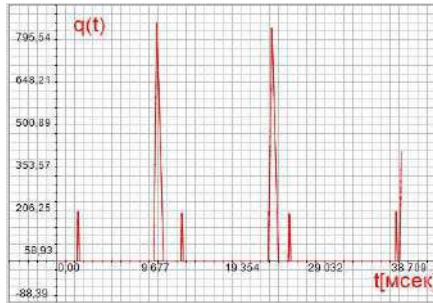


Fig.3. The number of packets in the queue at the intervals  $\tau$  when the load factor  $\rho = 0.1$ .

The resulting graphs Fig. 2 and Fig. 3 are completely analogous to those examined in [6].

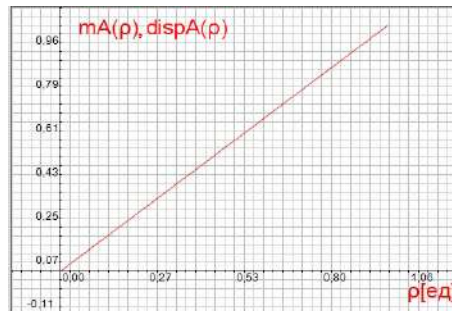


Fig.4. Dispersion and mathematical expectation of the number of packets on intervals  $\tau$  for Poisson traffic stream.

The developed program was tested on Poisson traffic stream. In Fig. 4 demonstrated the graphs of dependencies of average value and dispersion of the packets at intervals corresponding to different load factors of the system  $\rho$ . Both graphs are linear and completely coincide, which is typical for Poisson traffic stream.

#### 4. Main features of the program

Interface of the main program window consists of three blocks (Fig. 5):

1. Panel with flows.
2. Graph.
3. Setting bar of graph display.

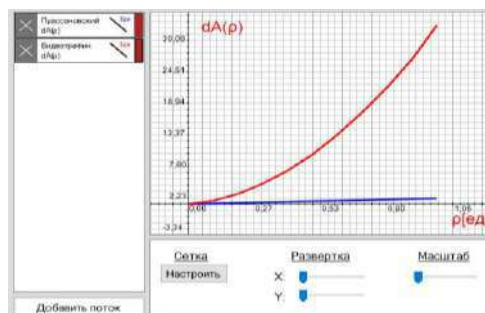


Fig. 5. The interface of the MSS traffic analysis program.

Examine each block in more details. The system provides the addition to a flow, using the "Add flow" button in the user interface (Fig. 6).

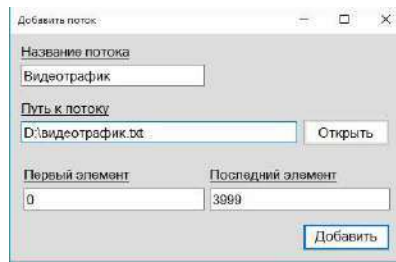


Fig.6.Interface adding flow.

The user has the ability to specify a name of the flow and limit the number of packets of input traffic. The flow file must contain information about the arrival times of packets or the intervals between packets.

The program is not limited to the number of flows and packets therein. You can add multiple flows and compare their characteristics, as shown in Fig. 5. After addition, the title of the flow is displayed on the panel (Fig. 7).

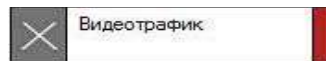


Fig.7. Graphical representation of flow.

The graphical representation of flow consists of three "buttons":

1. Remove the flow from the panel.
2. Display setting.
3. Shading the flow from the general graph.

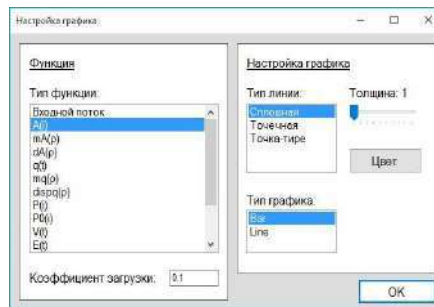


Fig. 8. Setting interface for flow display.

In the display settings dialog box of the graph (Fig. 8), a user can select traffic characteristic to build, tune in the graph (color, thickness and line type). When you click "OK", graphical representation of the flow changes, and selected characteristic is built.

The graph displayed in the Cartesian coordinate system is automatically scaled by the size of the window. The user has the ability to scale the graph, shift the plane graphics using the mouse, and reset these values and determine the coordinates of the point on the graph using the context menu (right mouse button). You can adjust the grid, sweep along the axes in the panel for setting the graph display.

## 5. Determination of approximation coefficients

The main practical benefit of this software is the ability to determine the coefficients of the approximation of the numerator of generalized Khinchin-Pollaczek formula [5], [6], [8]. It is necessary to build a characteristic of numerator dependence  $mE(\rho)$  of Khinchin-Pollaczek formula on the load factor  $\rho$  and to approximate it by a function  $F(\rho) = \alpha(\rho - \rho_0)^2 + \beta(\rho - \rho_0)$ .

By changing the coefficients  $\alpha$ ,  $\beta$  and  $\rho_0$ , an approximation is made automatically, the results of which are shown in the diagrams Fig. 9.

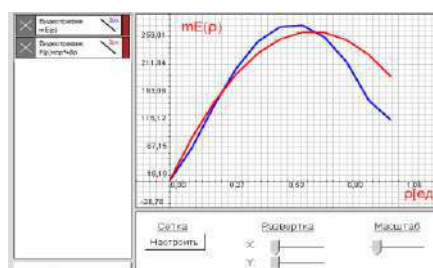


Fig.9. Approximation of characteristics  $mE(\rho)$  of video stream.

The obtained coefficients determine the generalized formula of Khinchin-Pollaczek:

$$\overline{q(\rho)} = \frac{mE(\rho)}{2(1-\rho)} = \frac{\alpha(\rho - \rho_0)^2 + \beta(\rho - \rho_0)}{2(1-\rho)}.$$

Using the generalized Khinchin-Pollaczek formula, the average values of the queues are determined for different load factors. In Fig. 10 shows the real values of the queues and the values obtained as a result of approximation.

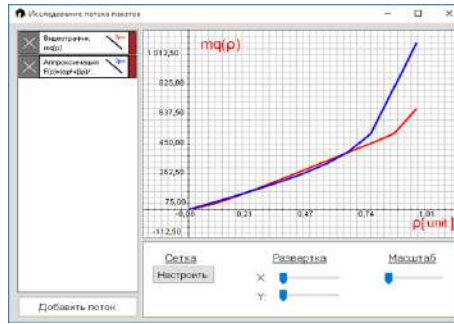


Fig.10. The numbers of packets in the queue, obtained using generalized Khinchin-Pollaczek formula and as a result of approximation.

## Conclusion

The developed software can be used in the analysis of traffic of multiservice telecommunication networks. Further development involves automatic collection of information about the flow and processing in real time, allowing to obtain average values of coefficients of e generalized Khinchin-Pollaczek formula.

## References

- [1] Stepanov SN. Teletraffic theory. Concepts, models, applications. Moscow: Goryachaya liniya-Telecom Publ., 2015; 808 p. (in Russian)
- [2] Kleinrock L. Communication Nets; Stochastic Message Flow and Delay. New York: McGraw-Hill, 1964.
- [3] Sveshnikov AA. Applied methods of a random function theory. Moscow: Nauka Publ., 1968; 460 p.
- [4] Martin Jh. A System analysis of data transmission. Englewood Cliffs, NJ: Prentice-Hall, 1972.
- [5] Likhtsinder BY. Interval method of traffic analysis in multiservice communication systems. A supplement to journal Infokommunikacionnye tehnologii 2013; 8: 104–152. (in Russian)
- [6] Likhtsinder BY. Interval analysis method of access networks. Samara: PSUTI Publ., 2015; 121 p. (in Russian)
- [7] Likhtsinder BY. Correlation features of queue sizes in queueing systems with common type flows. Infokommunikacionnye tehnologii 2015; 13(3): 276–280. (in Russian)
- [8] Likhtsinder BY. About some generalizations of Pollaczek-Khinchin formula. Infokommunikacionnye tehnologii 2007; 5(4): 253–258. (in Russian)
- [9] Likhtsinder BY. Correlation connections in batch flows of queueing systems. Telecommunications 2015; 9: 8–12.
- [10] Sanders C. Practical Packet Analysis: Using Wireshark to Solve Real-World Network Problems. No Stach Press, 2017.

# Semantic Analysis of Text Data with Automated System

O. Chernenko<sup>1</sup>, O. Gordeeva<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

This paper describes application of the basic methods of semantic analysis of text data – Porter stemming, frequency semantic analysis, latent semantic analysis and syntactic semantic analysis using an automated system. The system allows analyzing the text using these methods. The characteristics and features of the methods' implementation as well as the obtained results of their applying to texts of small complexity are considered. The research allows to reveal features of usage of the methods according to the text analysis purposes.

*Keywords:* text analysis; frequency semantic analysis; Porter stemmer; latent semantic analysis; core words

---

## 1. Introduction

At present days, it is difficult to imagine an effective work with text data without using computer processing. One of the most relevant and ever-evolving types of text processing is the semantic analysis. Depending on the criteria which are set in the automated system, the most appropriate type of semantic analysis can be selected. For example, in the case of search audit of a site, the criteria for choosing a method of semantic analysis are the speed of processing and the minimal dictionary volume. In the case of literal piece of art with complex speech turns, the main criterion of selecting an analysis method is the quality of processing. Consequently, the algorithm of semantic analysis should achieve the results that are as close to natural human as possible, so the parameters such as speed and volume of the dictionaries are not decisive.

## 2. Task Formulation

The object of the research is a text in Russian, no longer than 20 sentences, with the topic which is unambiguously understandable for human. The purposes of the research are to evaluate the execution of the four selected methods of semantic analysis the developed system is based upon and to compare efficiency and speed of analysis for the methods.

## 3. Methods of Text Semantic Analysis

All variety of text analysis methods can be divided into two groups:

- linguistic analysis – is based on extracting the meaning of the text from its semantic structure;
- statistical analysis – is based on extracting the meaning of the text and core words by the frequency distribution of words in the text.

The division into two groups is conditional, since in real problems a combination of methods is always used to achieve the result [1, 2].

In this paper, algorithms of semantic analysis from both groups, which are most often used for tackling practical problems, are realized.

### 3.1. Frequency Semantic Analysis

Method of Frequency Semantic Analysis (FSA) is based on calculating of frequency of word occurrence in the text. There are several refinements for the correct operation of the algorithm [3]:

- since not every word in the text can be the core word, only nouns are counted;
- to distinguish nouns in the text, a dictionary should be used.

The steps of algorithm work: firstly, all words of the text are compared with the dictionary; secondly, the matches are entered into the array, and then they are compared by the number of occurrences. Finally, the words with the largest number of occurrences are the core words of the text.

### 3.2. Porter Stemming

“Stemming” is a clipping the ending and suffix of the word so that the rest part becomes the basis for all grammatical forms of the word. “Porter Stemmer” or “Porter Stemming” is the algorithm of stemming that determines the basic part of a word. The stemmer can only work with languages that implement word modification through affixes, for example, Russian or English. The main advantage of this algorithm is that it does not need any dictionary or library.

First of all, there are several notations introduced for the parts of the word for stemming process:

- RV – the part of the word after the first vowel. It can be empty if there are no vowels in the word;
- R1 – the part of the word after the first "vowel-consonant" combination;
- R2 – the subpart of R1 after the first "vowel-consonant" combination.

In the article [4], the Porter's algorithm for a word's basic part (stem) determining is described. The algorithm includes the deleting of prefixes, endings and suffixes:

- if there is a gerund ending in the word, it must be removed. Otherwise, if the endings "sia" or "sj" are in the word they must be removed. Next, an adjective/verb/noun ending are looked for. If one of them is found – it must be removed;
- the ending "i" should be found and removed if it is there;
- the endings "ost" or "ostj" must be found and removed if one of them is there;
- if the word ends with "nn" – the last letter "n" must be removed;
- if the word ends with "eyesh" or "eishe" – this part must be removed and then, the last letter "n" must be removed if the word ends with "nn" again;
- if the word ends with "ь" – it must be deleted;

To determine the theme of the text using an algorithm based on Porter Stemming, it is necessary to carry out the stemming process for all words of the text being analyzed. As a result, an array of stems is obtained. The words in the text that are derived from the stem with the most frequent number of occurrences are marked as the theme of the text [5].

### 3.3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method of processing data in a natural language. The method analyzes the relationship between the set of documents and the terms in them and juxtaposes some factors (themes) to all documents and terms. The LSA method is based on the principles of factor analysis. As an input, the LSA uses a term-to-document matrix (terms – words or phrases) [6].

Elements of this matrix contain coefficients (weights) taking into account the frequency of occurrences of each term in each document. The most common version of LSA is based on the using of the singular decomposition of the diagonal matrix (SVD - Singular Value Decomposition). Using the SVD-decomposition, any matrix decomposes into a set of orthogonal matrices, the linear combination of which is a quite accurate approximation to the original matrix.

More formally, according to the singular decomposition theorem, any real rectangular matrix can be decomposed into a product of three matrices:

$$A=USV^T,$$

where matrixes U and V are orthogonal, and the matrix S is diagonal, values in diagonal of the matrix S are called "singular values" of matrix A. Letter "T" means transpose for matrix V.

Such decomposition has a significant feature: if in the matrix S retain only "k" largest singular values, and in the matrices U and V retain only columns corresponding to these values, then the product of the resulting matrices S, V and U is the best approximation of the original matrix A to the matrix  $\hat{A}$  of "k" rank:

$$\hat{A} \approx A=USV^T.$$

The main idea of the LSA is that if the matrix A is the term-to-document matrix, then the matrix  $\hat{A}$  containing only the first "k" linearly independent components of the matrix A reflects the basic structure of the dependencies presenting in the original matrix. Proceeding from this decomposition, the dependence between terms and documents is analyzed and the theme of the text is determined [7].

### 3.4. Syntactic Semantic Analysis

Syntactic Semantic Analysis is a method of processing textual information, which creates templates for comparison with words of text. As a result of the method a list of pairs is created for each sentence [8]. Pair includes:

- the type of word in the sentence;
- the position of the main word for this dependent word.

It is assumed that the basic templates are formed from the most important and often used semantic relations in the text. The basic semantic template is the rule by which the semantic relation is determined in the text being analyzed. The basic semantic template consists of 4 main parts:

- a sequence of words or indivisible semantic units for which their morphological features are indicated;
- the name of the semantic relation that should be formed if the sequence described in the previous item is found in the text;
- a sequence of numbers that determines the positions in a sequence whose elements should be added to the queue with priority. According to the queue the words from the sentence being analyzed are deleted;
- a number indicating the value of the priority, the group of semantic dependencies to which this semantic relation relates.

Using the basic semantic templates, the priority queue is composed. This queue is used to store words that are the argument on the right side of a semantic collocation found in the sentence being analyzed.

To determine the theme of the text from each sentence, according to the priority queue, the word with the biggest number of dependencies is selected and the number of its occurrences in the text is calculated. The word with the maximum number of occurrences is the theme of the text.

## 4. System Operation

To conduct the research on the methods of text analysis, an automated system was developed. The system operation includes several steps.

At the first step the system splits the text into elements (words or sentences, it depends on the algorithm chosen by the user), and sends them for processing. The second step is handling the elements and selection the core words.

If the FSA has been chosen by the user, the system compares words from the text with words from the dictionary and finds among them words with the maximum number of occurrences in the text. Next, it displays the finding result – core words of the text. Additionally, system displays the list of words has not been found in the dictionary. It is possible to add that words to the dictionary and run the algorithm anew.

If the algorithm based on Porter Stemming has been chosen, the system defines the basics of original words in the text and looks for the most frequent among them. In this way the core words of the text are found by this algorithm.

In the case of LSA the system constructs a word-on-sentence matrix using the sentences of the text and carries out the SVD decomposition. Then only the first two columns of the resulting matrices are used. From the first two columns of the matrix  $V^T$  corresponding to the sentences, a maximum and a minimum are selected. These values correspond to the maximum and minimum coordinates  $x$  and  $y$  on the coordinate plane. In this way, the area indicated, the entry into which for points from the first two columns of the matrix  $U$  corresponding to words means the inclusion into core words.

If the SSA has been chosen by the user, in each sentence words are checked for matching patterns. After that the weight value sets for every word according to the pattern. The more word dependent words, the weight is less and priority is higher. Next, the word with a minimum weight is determined in each sentence, the words with the most number of occurrences form the core of the text.

## 5. Results

As objects for research, texts for the essays of the Unified State Examination in the Russian language were chosen. These texts were chosen because of their simplicity and small size, and also because their themes are clearly defined.

In tables 1-5 core words for text examples are presented. In addition, time of processing for each method of analysis is given.

Table 1. «Send your head on vacation!» (P. Izmaylov).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	vacation
Porter Stemming	1	feelings each other another series head heads vacation
Latent Semantic	210	head feelings love series passion time rhythm rubles vacation
Syntactic Semantic	720	series publishing vacation

The main idea of the first text is “the influence of mass literature on the human intellectual development”. No methods produced similar theme, but the most suitable core words were given by the latent-semantic method and Porter stemming method.

Table 2. «Things and books, books and things...» (L. Lickhodeev).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	locomotive light book thing time interlocutor
Porter Stemming	2	book
Latent Semantic	240	think idea interlocutor light book thing things time another each other
Syntactic Semantic	840	think things time interlocutor

The main idea of the text could be defined as “relationship between book and time”. The most appropriate core words got the latent semantic method of analysis.

Table 3. « Earth is a cosmic body, and we are cosmonauts...» (V. Solouckhin).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	life support system spaceship cosmonaut Earth communication possibility sides human disease
Porter Stemming	1	life support cosmic cosmonaut cosmonauts spaceship human
Latent Semantic	120	Solar life support cosmic cosmonaut cosmonauts small spaceship Earth river nature disease outside human inner life
Syntactic Semantic	540	cosmonauts cosmonaut spaceship human



The text's main idea is "relations between human and nature". As in previous example, the latent semantic analysis gave the most similar core words.

Table 4. «Books...» (A. Yetoyev).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	life people human childhood book friend
Porter Stemming	1	book people
Latent Semantic	120	measure meet human people similar similarly book books enable
Syntactic Semantic	540	human space life population people book

Core words given by syntactic semantic analysis are the most similar to theme of the fourth text "the role of book in human life"

Table 5. «About the soul...» (M. Prishvin).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	soul raincoat
Porter Stemming	1	soul
Latent Semantic	90	soul raincoat
Syntactic Semantic	600	soul year raincoat

The topic of the fifth text is "soul of human". All algorithms gave the satisfactory results, the most accurate of which gave the Porter stemming.

## 6. Conclusion

In the article methods of classification of texts, such as Porter stemming, syntactic semantic, frequency semantic and latent semantic analysis, are considered. The results of the analysis of little complexity texts are given. Based on these results it can be concluded that the usage of methods for determining the topic of a text depends on the complexity of the text – the more accurate analysis for the more complex text should be.

The same applies to trivial texts. The using of complex methods for simple texts leads to unnecessary waste of time and resources, and the result is superfluous in comparison with simple algorithms. Thus, the research shows that for short texts the most effective method is the latent semantic analysis, the fastest method is the Porter stemming. Finally, it should be mentioned that the combination of text analysis methods, for example, combining the Porter method of stamping and frequency-semantic analysis, can be appropriate for effective and accurate core words determination.

## References

- [1] Velichkevich AG, Cherepackhina AA. Latent semantic analysis of text using Porter algorithm. Youth scientific and technical herald 2015; 10: 38 p. (in Russian)
- [2] Mikhaylov DV, Kozlov AP, Emelyanov GM. An approach based on tf-idf metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets. Computer Optics 2015; 39(3): 429–435. DOI: 10.18287/0134-2452-2015-39-3-429-438.
- [3] Understanding and synthesizing text by computer. URL: <http://compuling.narod.ru/index2.html> (11.12.16). (in Russian)
- [4] Russian stemming algorithm. URL: <http://snowball.tartarus.org/algorithms/russian/stemmer.html> (11.12.16). (in Russian)
- [5] Silva G, Oliveira C. A lexicon-based stemming procedure. Lecture Notes in Computer Science 2003; 2721: 159–166.
- [6] Zaboлева-Zotova AV. Latent semantic analysis and new solutions in Internet. Moscow: Information Technologies, 2001; 22 p. (in Russian)
- [7] Kuralenok I, Nekrest'yanov I. Automatic document classification based on latent semantic analysis. Programming and Computer Software 2000; 26(4): 199–206.
- [8] Rabchevsky EA. Automatic construction of ontologies based on lexical-syntactic templates for information search. Petrozavodsk, 2009; 107 p. (in Russian)

# Modeling of online social networks for automated monitoring system

Yu.B. Savva<sup>1</sup>, Yu.V. Davydova<sup>1</sup>

<sup>1</sup>Orel State University, 95, Komsomol'skaya, 302026, Orel, Russia

---

## Abstract

Monitoring using keywords is necessary step in solving the problem of detection of users' illegal behavior such as drug use, extremist propaganda in online social networks. Analysis of text posts is difficult because of using jargon and making mistakes in communications. In paper model of online social networks for automated monitoring system is presented. This model focuses not on communications between users but on text posts. Features of Russian text posts are given. Problem of text posts obfuscation by users involved in illicit fields of activities is discussed.

*Keywords:* online social networks; monitoring; text analysis; information retrieval; fuzzy search

---

## 1. Introduction

Advantages of online social networks (OSNs) such as high speed of information dissemination which can be compared with a virus and ease of use make them a convenient tool for information influence and propaganda of deviant and illegal actions. Threats of OSNs, such as extremist and terrorist groups, were discussed in [1]. For providing information and psychological security of users, automated monitoring system of OSNs is required. Most existing monitoring systems [2] are used for business goals and find out users' attitude to brands. Sharing their opinions about products, service or social events users try to use hashtags – correctly written mentions with special label or metadata. It makes it easier for users to find messages with a specific theme or content. As for illegal activity, it isn't advertised or advertised for closed groups of users though propaganda can be an exception. It is more difficult to find posts connected with searching topic without hashtags especially if they are written with mistakes. Spelling errors and typos are common for informal writing on the whole and for text posts in OSNs in particular. Also informal writing is often characterized by using slang and different abbreviations. As for illicit fields of activity, communications often contain specialized jargon. We consider jargon as a highly specialized slang, which is often used in closed communities and is hard to understand. Taking these features into consideration it can be said without doubt that monitoring process is difficult and requires special methods for analysis of text posts.

Process of monitoring involves a kind of information retrieval, text posts from OSNs are gathered and fuzzy search by keywords is organized. Keywords represent lexics, which is used in communications in illegal fields of activity. This paper describes model of online social networks used in automated monitoring system. According to system's goal, the emphasis of the model is made on text posts.

## 2. The object of the study

Usually online social network is defined as a graph  $G(N, E)$ , where  $N = \{1, 2, \dots, n\}$  is a set of vertices (agents - users, communities) and  $E$  is a set of edges which represents interaction of agents [3]. Tasks of users' behavior modeling, users' interaction modeling, analyzing features of subgraphs of friendship are popular. The main goal of current automated monitoring system is decision support in detection of illegal behavior in OSNs, wherein information retrieval and analysis of text posts play a big role. Thus, text posts should be included in model.

Let us denote  $I = \{i_1, i_2, \dots, i_{ic}\}$  is a set of identifiers of OSNs users or communities, where  $ic$  is the number of identifiers.  $M = \{m_1, m_2, \dots, m_{mc}\}$  is a set of posts,  $mc$  is the number of posts. Posts are gathered into groups:  $M = \sum_{k=1}^{ic} M_k$ . Every post can be represented as follows:

$$m_j = \langle i_k, \text{text}_j, t_j, \text{type}_h, \text{parent}_j \rangle, i_k \in I, j = \overline{1..mc}, h = \overline{1..3},$$

$$\text{parent}_j = \begin{cases} \emptyset, & \text{type}_j = \text{type}_1 \\ i_n \in I, n = \overline{1..ic} & \end{cases},$$

where: –  $i_k$  is identifier of user who posted the message  $m_j$ ;

–  $\text{text}_j$  is text of the post  $m_j$ ,  $\text{text}_j = \langle w_{j1}, w_{j2}, \dots, w_{jg} \rangle$ ,  $w_{ji}$  is the  $i$ -th word in text;

–  $t_j$  – date and time of the posted message  $m_j$ ;

–  $\text{type}_h \in \text{Type}$ ,  $\text{Type} = \{\text{type}_1, \text{type}_2, \text{type}_3\}$  is a set of post types, where  $\text{type}_1$  is original post (which means that user who posted message is its author),  $\text{type}_2$  is reposted message (which means that user posted somebody's message),  $\text{type}_3$  is a comment to original or reposted message;

–  $parent_j$  is a user's or community's identifier. If type of current post is a repost or a comment then  $parent$  contains identifier of author who posted original message.

$P = \{p_1, p_2, \dots, p_{ic}\}$  is a set of pages of OSNs, number of pages is equal to number of users' and communities' identifiers as every page belongs to user or community. Page is defined as follows:

$$p_k = \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \mid z = \overline{1..x} \right\} \right\rangle, i_k \in I, x < mc, q = \overline{1..2},$$

$$c_k = \begin{cases} \emptyset, tt_k = tt_1 \\ \{i_n\} \subset I, n = \overline{1..ic} \end{cases},$$

where: –  $i_k$  is the identifier of user or community of current page  $p_k$ ;

–  $tt_k \in TT$ ,  $TT = \{tt_1, tt_2\}$  is a set of pages type.  $tt_1$  is a personal page and  $tt_2$  is a community page;

–  $c_k$  is a set of user's identifiers. If current page  $p_k$  is a personal page then  $c_k$  is an empty set as page  $p_k$  belongs to one user. If  $p_k$  is a community page then  $c_k$  keeps user's identifiers who are owners or managers of community (it can be one user, so  $c_k$  keeps one element);

–  $M_k$  is a group of posts which are posted on the page  $p_k$ . It can be empty  $M_k = \emptyset$ , that means OSNs page doesn't contain any posts at the moment.

Set of keywords is given  $L = \{l_1, l_2, \dots, l_{lc}\}$ , where  $lc$  – the number keywords. Every keyword is represented by its grammatical, semantic information and word forms (according to inflection rules in Russian language)  $l_s = \{GR_s, SM_s, WF_s\}$ . This keywords storage model was described in [4]. In this work we are focused on word forms of keywords. They was defined as a language  $WF$  over the alphabet  $A$ .  $WF \in A^+$ .

The goal of automated monitoring system of OSNs is to find set of pages  $PF \subset P$  which contains required amount of keywords, therefore these pages are indicators of potential illegal actions of their owners. Conceptually it can be presented as follows:

$$PF = \left\{ \left\langle i_k, tt_q, c_k, M_k = \left\{ m_{kz} \mid z = \overline{1..x} \right\} \right\rangle \mid \left( \sum_{z=1}^x \sum_{q=1}^{lc} f(m_{kz}, l_q) \geq \delta \right) \wedge (k = \overline{1..ic}) \right\},$$

where: –  $f(m_{kz}, l_q)$  is a function which is defined as  $f(m_{kz}, l_q) = y(\text{text}_{kz}, WF_q)$ ;

–  $\delta$  is a threshold of presence of keywords in text posts of current user. It can be defined by decision maker.

$y(\text{text}_{kz}, WF_q)$  is a function of fuzzy search matching, conceptually it can be presented as follows:

$$y(\text{text}_{kz}, WF_q) = \sum_{i=1}^g \sum_{j=1}^r d(w_{zi}^k, wf_{qj}), d(w_{zi}^k, wf_{qj}) \leq \varphi,$$

where: –  $d(w_{zi}^k, wf_{qj})$  is a distance measure, which shows similarity between two words  $w_{zi}^k$  and  $wf_{qj}$ . Initial states are:  $d(0, wf_{qj}) = wf_{qj}$  and  $d(w_{zi}^k, 0) = w_{zi}^k$ ,

–  $\varphi$  is a threshold of distance measure, it can be defined by decision maker. The less is the value of distance measure, the higher is similarity between words. That means that current word in text post is a keyword written with mistakes with great probability. By choosing the value of threshold of distance measure, decision maker can manage the levels of precision and recall of information retrieval. The less is the value the more precise search is, but in this case, some relevant text posts will be lost and recall will be lower.

As the result of Russian text posts analysis from OSNs it was revealed that users use informal style of writing and often neglect the language rules.

### 3. Features of Russian text posts of OSNs users

Text posts in OSNs have the following features:

- use of conversational style in writing, slang and jargon use, abbreviations use;
- short length of average text post with weak formal syntactic relations;
- use of smileys, different special symbols;
- intentional and unintentional garble of words, including spelling errors and typos;
- borrowings from English language, like "4u" (For You).

These features characterize modern informal communications, where there is a high speed of information exchange and additional expression. Thus, text posts of OSNs users can be considered as unstructured sequences of letters symbols and images combining with each other. This fact should be taken into account in text analysis. As for communications in illegal fields of activity, additional features should be noted. To avoid detecting by law-enforcement agencies people use jargon. The main difficulty of text analysis in case of jargon use consists in high degree of homonymy. Words of common used lexics may be organized in collocations and thus get new semantics as a result. There is a constant appearance of new jargon. The most glaring example is jargon in the field of illicit traffic of narcotic drugs as new substances appear rather quickly. Also it should be considered that OSNs users involved into illegal activities obfuscate posts with the same aim to prevent their detection.

#### 4. Problem of text posts obfuscation and methods dealing with it

First mention of obfuscation method appeared in [5]. Authors suggested confusing of program code by adding extra variables and constructions with aim to prevent algorithm analysis and deter reverse engineering. Also obfuscation can be used to optimize code. Analysis of obfuscation methods of computer program is given in [6], deobfuscation methods are presented in [7]. Later obfuscation was applied to creating spam emails, spam messages on different web sites. In this case obfuscation allows to pass through content filtering. Obfuscated words can't be found during exact matching between words from message and words from dictionary. Dictionary contains words, which are indicators of spam messages.

Text posts in OSNs can be obfuscated by users involved in illegal fields of activity, for instance terrorist and extremist propaganda, illicit drug sales. In this case as it was discussed in [8] users obfuscate their posts to prevent effective linguistic analysis of texts and avoid detection of their destructive actions and influence on other OSNs users. For text obfuscation generally the following methods are used:

- intentional garble of words, including spelling errors, typos, wrong word boundaries (space insertions and deletions);
- letter substitution by digits, symbols which look like substituted letters;
- insertion extra not meaningful symbols;
- transliteration use.

Thus, text posts deobfuscation is the actual and difficult issue. Solution by computer means is not a trivial task as there are many ways of obfuscation of even one word. Thereby such methods as spell checking, deleting non-alphabetic symbols and constructing variants by possible substitutions are not so effective. Applying Hidden Markov Model to the task of spam emails deobfuscation showed good results [9]. Also, statistical models can be useful, for instance, model based on Bayesian rule, n-gram model [10, 11].

#### 5. Using model of OSNs in automated monitoring system

Automated monitoring system includes the following main subsystems:

- data collection;
- fuzzy text search which includes linguistic knowledge base, keywords database, algorithmic search and deobfuscation modules;
- results processing and report generation modules;
- database of text posts and database of search index.

According to model, data collection subsystem gathers identifiers and text posts with additional attributes like type of messages, time and date of posting. This information is stored in database of text messages. Decision maker can specify settings of OSNs crawl strategy.

Subsystem of fuzzy text search takes information from database of text posts and implements the goal of automated monitoring system, trying to detect illegal behavior by using linguistic knowledge base and keywords database. At first stage tokenization is held, text is deobfuscated if it requires. The second stage is fuzzy text search using keywords. The use of linguistic knowledge base helps to make information retrieval not so sensitive to mistakes. Linguistic knowledge base contains information about inflection paradigms, models of mistakes, typos. Keywords database stores grammatical, semantic information and word forms of keywords lexemes. In case some text post contains threshold amount of keywords, it is indexed and is sent to database of search index. Processes of gathering information by data collection subsystem and searching by subsystem of fuzzy text search are parallel. Report generation modules show different slices of results to user of monitoring system such as topic distribution, age and location distribution of OSNs users and some others. Results are grouped according to threshold of similarity distance measure.

#### 6. Results and discussions

At the present time automated monitoring system is to be used in detection of drugs use propaganda and illicit drug sales in OSNs [12], though system can be used in different fields, it depends on keywords database. Linguistic database of keywords used in the field of illicit traffic of narcotic drugs and psychotropic substances was developed [13]. It allows to store not only word forms but semantics of jargon. Deobfuscation method using Hidden Markov Model was developed [14], example of algorithm is presented at 0

Corpus of text posts is gathered from OSN Vkontakte. Currently algorithms of fuzzy search using keywords from developed linguistic database and models for linguistic knowledge base are developed. Features of algorithms and default values for distance measure should be tested on text corpus and corrected in case of need as they are a kind of empirical data because natural language is not a good formalized object [10, 11].

#### 7. Conclusion

For providing information and psychological security of users, it is necessary to organize online social networks monitoring. Monitoring process has many difficulties like short messages in OSNs, informal communications using jargon, text posts obfuscation. To detect users' illicit activities and destructive influence effective text analysis and search by keywords should be organized. Thus, in OSNs modeling emphasis should be on text posts, corresponding model was presented in this paper. Main

subsystems of automated monitoring system using aspects of the model were described. The results of the work done were discussed and features of future work were given.

```

===== RESTART: C:/Python34/deobf.py =====
>>> D=Deobf(connection)
>>> D.deobfuscate(tree, """"спано
так кт о х пм, на конец? уа - счастье той силы, чтоо вежно хочет зла и вежно осв
ершает бллар""")
Товариш, друг, не скупись, купи немножко коноплии
Ой и@ла мы|ла ггазу лала вода водафон
я-пришёл-я-тебе-с-приветом
это мыло давно и неправда н е п р а в д а
ч
у
ш
ь
весь котрый соабка пришёл дмой анольд коова кроова
жизнь - 50/\ъ и прочие радости страдание исчо""")
шано так что мы наконец я часть той силы чтоо вежно хочется и вежно совершает бл
аго товариш друг не скупись купи немножко коноплией и мыла разу вода вода он я
пришёл тебе с приветом это мыло давно и неправда не правда чушь весь котрый со
бака решил домой анольдкоова крова жинь больше и прочие радости страдание исче
>>> |

```

Fig. 1. Example of text deobfuscation.

## References

- [1] Davydova YuV. To the issue of need for automation of threats search process in virtual social networks and communities. Actual problems in modern science in XXI century: proceedings of the 6<sup>th</sup> international scientific-practical conference. Makhachkala: "Aprobaciya" Publisher, 2014; 25–26. (in Russian)
- [2] The top 25 social media monitoring tools. URL: <http://keyhole.co/blog/the-top-25-social-media-monitoring-tools/> (19.01.2017).
- [3] Gubanov DA, Novikov DA, Chhartishvili AG. Online social networks: models of information influence, control and confrontation. Moscow: "Fizmatlit" Publisher, 2010; 228 p. (in Russian)
- [4] Savva YuB, DavydovaYuV. Linguistic database for monitoring system of online social networks in providing information and psychological security. European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. Belgrade: "Criminalistic-Police Academy" Publisher, 2016; 1: 145–154.
- [5] Diffie W, Hellman M. New directions in cryptography. IEEE Transactions on Information Theory 1976; IT-22(6): 644–654.
- [6] Korobejnikov AG, Kutuzov IM, Kolesnikov PYu. Analysis of obfuscation methods. Cybernetics and programming 2012; 1: 31–37. (in Russian)
- [7] Kasperski K, Rokko E. The art of disassembling. SPb: BHV-Peterburg, 2008; 892 p. (in Russian)
- [8] Savva YuB, Eryomenko VT, Davydova YuV. About the problem of the linguistic analysis of the slang in the problem of the automated search of threats of spread of drug addiction on virtual social networks. Information systems and Technologies 2015; 6(92): 68–75. (in Russian)
- [9] Honglak L, Andrew YNg. Spam Deobfuscation using Hidden Markov Model. Proceedings of the Second Conference on Email and Anti-Spam, 2005. URL: <http://ai.stanford.edu/~ang/papers/ceas05-spamdeobfuscation.pdf> (11.07.2016).
- [10] Ingersoll GS, Morton TS, Farris AL. Taming text. How to find, organize and manipulate it. NY: Manning Publications Co., 2013; 320 p.
- [11] Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008; 496 p.
- [12] Savva YuB, Eryomenko VT, Davydova YuV. Design of information system identification of persons which participate illicit in field of narcotic drugs and psychotropic substances in the virtual social networks using the database jargon. Information systems and Technologies 2016; 1(93): 68–75. (in Russian)
- [13] Savva YuB, Davydova YuV. Certificate of state registration database no. 2016620197. Jargon in the field of illicit traffic of narcotic drugs and psychotropic substances. Registered 10 February 2016.
- [14] Nikol'skaya AN, Savva YuB. About the problem of opening of obfuscated Russian-language texts of participants of online social networks. Information systems and Technologies 2016; 6(98): 44–55. (in Russian)

# Development and research of algorithms for clustering data of super-large volume

I.A. Rytsarev<sup>1</sup>, A.V. Blagov<sup>1</sup>, M.I. Khotilin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The work is devoted to the research of text data clustering algorithms. As the object of research, the social network Twitter was selected. At the same time, text data was collected, processed and analyzed. To solve the problem of obtaining the necessary information, studies in the field of optimizing the data collection of the social network Twitter were carried out. A software tool that provides the collection of necessary data from specified geolocation has been developed. The existing algorithms for clustering data, mainly of large volume were explored.

*Keywords:* data clustering algorithms; superhigh volume data; text analysis; k-means; tf-idf metric; lda; collective decision-making method

---

## 1. Introduction

The aim of the paper is to explore the algorithms of clustering text data of social networks collected on certain geolocations. As the object of research data from the social network Twitter was used. To achieve the goal, the following tasks were set:

- collection of social network data,
- processing of the received data with extraction of the necessary information,
- research, approbation and modernization of data clustering algorithms.

During the research work the following algorithms were studied and tested:

- The k-means algorithm,
- LDA algorithm;
- algorithm of data classification by the judge method.

In addition to the algorithms, the following measures were tested:

- TF-IDF,
- Word2Vec.

A software product to collect data from the social network Twitter was developed. A software product for cluster analysis of collected data is also being developed.

## 2. Text data clustering

Clustering (or cluster analysis) is the task of dividing a set of objects into groups, called clusters [1]. Within each group there should be "similar" objects, and the objects of different groups should be as different as possible. At the same time, some measure must be defined. Unlike the classification for clustering, the list of groups is not clearly defined and is determined during the operation of the algorithm. The main goal of clustering is the search for existing structures [2-6].

The most popular approach to solving the classification problem is the classification of information through machine learning.

Machine learning is the process by which a machine (computer) is able to display behavior that has not been explicitly programmed into it. There are two types of training: inductive and deductive.

In the works of researchers engaged in cluster analysis of textual information in various types of search engines, there is often an inductive measure of Word2vec [7-8]. The most popular deductive approach can be considered Dirichlet's Latent Placement (LDA).

For a more detailed analysis, it is best to combine different approaches and methods depending on the amount of processed data.

## 3. Data collection from social network Twitter

To investigate the operation of the TF-IDF algorithm, a software tool that allows data to be collected directly from Twitter servers was developed. The implementation is built on the open interface Twitter API 2.0. The object of the study was a message from a twitter (tweet) of the Samara and Moscow regions. The main criterion for the selection of messages was the presence of a certain geolocation (including all settlements of the region).

To perform the collection, a request to the Twitter server containing the consumer key and the consumer secret key is sent. In response to the request, `oauth.accessToken` and `oauth.accessTokenSecret` were obtained, which allowed receiving data from the servers of the social network.

The second step in the implementation of data collection is the sending of a query, in response to which a set of tweets is returned.

## 4. Results and Discussion

Data for analysis and subsequent clustering were collected within 24 hours, according to two query-requests: Samara and Moscow regions. 1.5 GB of information was collected (> 40,000 messages). After that, the following algorithms were applied to this information: modified TF-IDF, LDA [9-10], data classification algorithm with the help of graphs.

### 4.1. Processing with the modified TF-IDF algorithm

By applying the modified TF-IDF metric:

$$tfidf(t, d, D) = k * tf(t, d) \times idf(t, D), \quad (1)$$

where  $tf(t, d) = n_i / (\sum k * n_k)$ ,  $idf(t, D) = \log |D| / |(d_i \ni t_i)|$ ,  $k$  – correction factor, for words that are hashtags; and the k-means algorithm, 22 clusters were obtained. On the example of one of the obtained clusters (figure 1) it is clear that the messages are close in meaning, but among them there are messages with "foreign" subjects.

Such an inaccurate result was most likely obtained due to the fact that the researched messages on Twitter have a 140 character limit.

```

17.417364700059597: пайросс тридцать два потому что люблю бтс lovebts
17.412488867909754: когда ты будешь танце lovebts #ttrr t со 5wgentbrst
17.377890466760327: lovebts пожалуйста пожалуйста приходи к себе
17.385013020434603: just posted a photo #ttrr t со rtavuz4ort
17.33785524961285: why did i wake up iam
17.26244280531484: сердобно in самары самарская #ttrr t со kmployay7
17.256147680915472: ну не могу я думать #ttrr t со amr1qstyd
17.24574812632282: lovebts но все же если однажды увидишь улыбку
17.242810395943955: пайросс тридцать два моя фантазия закончилась lovebts
17.18060503437332: #хадивали это я так у мамы хоршо поднал
17.165552773189763: lovebts я пом в одиночестве всё ну же песню
17.15126146428252: чепуха пайросс чепуха пайросс чепуха lovebts
17.143256397595636: обожаю танки вчарахасисб ван #ttrr t со K13y9rnx1
17.13887528934295: старый мост река самара #ttrr t со Ikvsochfnu
17.122425825760548: с н доната хвалю #ttrr t со zvn1656yop
17.100052153754354: только версия мумбад lovebts с калитаном ливин
17.087077012033237: акура65651 как версия будет засну в ателье
17.052100143709136: zavaia in и сколько вполонных дел пайросс
17.02051064681348: мой утр после пайросс #ttrr t со s0s011rvbf
17.006630468592117: r bar terrace #ttrr t со 114637kuy1
16.9993484458442: обожаю лисы которм спазивают #ttrr t со Inm1zkaun
16.937749179184305: lovebts смелая в бесшумностях ко всем кроме себя
16.935328573424083: какой же день ужасный а мог быть лучше
16.91188651734057: #xuzhalk ego second box 2
16.911526256447324: ведь есть см на сета джидильми спасибо
16.88121521655293: пайросс тридцать два режиссёр танков перед сном lovebts
16.87953404614601: так не кончатся никому идти но надо lovebts
16.869827024667654: lovebts милая просто скажи что хочешь расстаться
16.25709517099365: уже просто не выскоу ситуацию которая происходит
15.84083916617625: anastasya2614 блин я думала на 4 пришла

```

Fig. 1. Example of one of the obtained clusters.

In addition, the high density of clusters (figure 2) indicates a low accuracy of the metric.

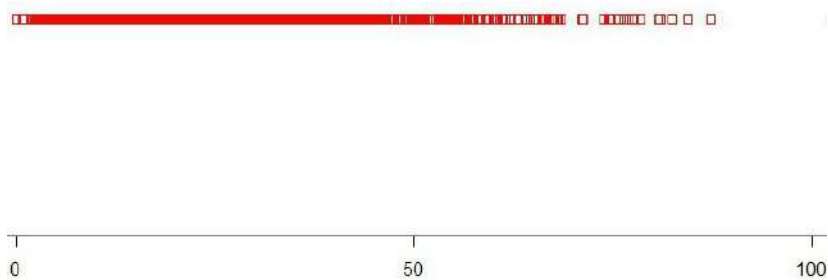


Fig. 2. Distribution of the values of the TF-IDF metric of the processed data on the number line.

### 4.2. LDA algorithm processing

LDA algorithm is based on the definition of the most used topics (themes) that can form clusters.

The LDA model solves the classical problem of text analysis: create a probabilistic model of a large collection of texts (for example, for information retrieval or classification).

- Obviously, one document can have several topics; Approaches that cluster documents on topics do not take this into account. LDA is a hierarchical Bayesian model consisting of two levels:
  - on the first level - a mixture, the components of which correspond to "themes";
  - at the second level, a multinomial variable with a priori Dirichlet distribution, which specifies the "distribution of topics" in the document.

Complex models are often the easiest to understand so - let's see how the model will generate a new document:

- choose the length of the document  $N$  (this is not drawn on the graph - it's not that part of the model);
- select a vector  $\theta \sim (\alpha)$  — the vector of the "degree of expression" of each topic in this document;



- for each of the N words w:
  - select a topic  $z_n$  by distribution ;  $\text{Mult}(\theta)$
  - Select a word  $w_n \sim p(w_n | z_n, \beta)$  with probabilities given in  $\beta$ .

For simplicity, we fix the number of topics k and assume that  $\beta$  is simply a set of parameters  $\beta_{ij} = p(w^j = 1 | z^i = 1)$ , Which need to be evaluated, and we will not worry about the distribution on N. The joint distribution then looks like this:

$$p(\theta, \dots, N | \alpha, \beta) = p(N | \xi) p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta). \tag{2}$$

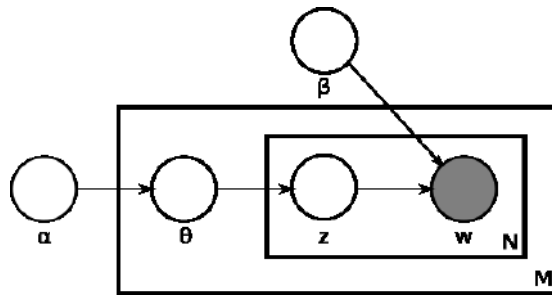


Fig 3. Graph of the model.

Unlike the usual clustering with the a priori Dirichlet distribution or the usual naive Bayesian, we do not select the cluster once, and then we insert words from this cluster, and for each word we first select the topic by the distribution of  $\theta$ , and then we sketch this word on this topic [11].

In the course of the work, it was revealed by expert means that the optimal number of initial clusters is six. The result of the algorithm can be seen in figure 4.

11.460942247791175	29.335382219962934	33.08138871682837	31.019046746471062	17.12769412794323	27.564212413341927
13.049531463180466	5.807253042654407	7.784337503305447	33.168623222689696	12.623453355849703	27.423509270276053
23.11225055594479	33.637297993034906	10.533192856647895	4.718768901268544	13.466039636433987	20.14923601340606
32.59713480426936	34.618140545200355	5.064047503331063	11.96549090693515	24.892465348230548	14.070350859706746
22.494536241486763	11.094202517033393	2.252791232746202	5.890999742299181	13.17795629911592	28.925242051447416
32.7846048240028	34.82922773238242	0.22356478712059102	33.80520371858417	28.40775214640371	14.602455713069292
19.19775718222764	7.329140887879075	35.51978788736594	2.09014471536552	29.681828849955853	2.4623585290876813
1.7600781755941732	5.355063964398675	5.138483776356878	7.473737421990786	2.5067380999779045	9.669343266991447
20.52925977425199	18.06148058511132	8.177967830963656	32.26262588328617	22.329141018220017	7.982835345897748
14.629077447636615	10.080926175674712	7.069979958572707	31.961259354637168	24.937467532713338	8.44733625747021
23.310952838889104	35.63766266479014	9.650058873488849	17.149514134602022	9.982123974097092	5.856771791668895
9.13114106376548	19.14392549146147	3.376295946268222	6.635326119765376	31.99333678264394	9.763100116902278
24.674674110388427	4.745730949581696	16.720354779081394	9.039702698119775	27.783475851905262	35.96695533435996
27.497431040189205	31.520007699604733	19.431465949779813	5.70387744541283	10.953834728290259	7.729405804718484
19.932501452966115	20.85694334625088	32.991541254927846	5.27385971193919	26.752520961069166	26.53592044587592
4.125152150536014	13.413765237717488	15.037014409417731	8.454689459164918	11.790367429629647	33.830463319036596
34.55286172113333	6.6184341194506136	12.891793178326186	2.220687083098924	5.077529455883766	14.220250875300106
21.4555930623101	19.300517834627577	35.12896658379044	28.501115830449958	36.62846957436838	8.777962129484473
36.07114222349534	26.695309184858758	36.421687233813714	32.86205170907326	20.053019352366267	28.082812593111854
16.537326211518387	28.916407130472674	2.942860286517592	33.6008581059411	35.33368181837114	8.936494502038368
3.40511125470972	13.428721666736125	1.4877018441541456	28.632344402842435	4.53371432944942	14.453018618454102
13.194579823414719	24.03229066612039	10.003549596770473	6.545171787561844	6.808099410090769	5.688326030005968
31.53080506742084	17.13838789273517	16.670051539214448	36.75996074957919	24.54343274729088	14.678391223880253
6.503021971766602	12.481008928918346	4.354442614419621	12.140328000083082	4.148074541871825	24.772516941338516
10.80364293374654	7.953688889307579	20.61733069227769	18.05508710198097	32.287534297249955	1.3576009474568327
18.40117359693428	4.635138603656412	29.332417618526126	29.421314767768372	1.720780932638921	28.392187294893265
19.38237734898474	4.706756105491227	11.948917140466982	30.87801445486595	27.271652901289496	33.825825922366484
11.735392992509071	25.353634136120732	14.93001256177041	31.949676296418346	9.46832615387672	28.736257457365102
26.230096388382425	36.66466422993565	14.578313454098147	17.905991373814395	13.527295235888436	35.43768890161636
34.849852086972916	29.330798708178623	20.36558541846089	22.353052389118307	9.342823435427412	28.5082849505737
3.0268397060981345	3.7136480095057403	29.7465797126157	8.294166588436644	3.086486845255883	0.5611939488837487

Fig. 4. The result of the algorithm.

Figure 4 shows the probabilities of the text belonging to each of the 6 clusters.

### 4.3. Algorithm of classification of data by the collective decision-making method

The algorithm for classifying data by the collective decision-making method is based on the idea that each word relates to one or another category (class). Then, as a result of processing, the text will be a set of "voices" of the affiliation of each word in the text to one or another class. Analyzing the resulting vector, we can decide which class the text belongs to.

Currently, the algorithm is being developed. The results will be presented for comparison later.

## 5. Conclusion

As a result of research work, a software package that allows to collect data from the social network Twitter for certain geolocations was written. With the help of this complex, data collection was carried out in the Samara and Moscow regions.



It was found that using algorithms based on the use of the TF-IDF metric, it is difficult to obtain a qualitative clustering of the textual information contained in short messages of the social network Twitter. From this we can conclude that the TF-IDF metric is not suitable for short text messages, or about the necessary modernization of this metric.

Algorithms based on "machine learning", in turn, demonstrated good results - six clusters of messages were identified: "study", "emotions", "photo sharing", "urban environment", "city news", "politics". This suggests "rejuvenating" the audience of the social network.

The data classification algorithm by the judge's method (currently) is under development.

Questions on clustering and further classification of text data are relevant in connection with the enormous spread of social networks and Internet services around the world.

In the course of further work, it is planned to compare the implemented algorithm for classifying text data and the LDA algorithm, as well as studying the issue in the direction of output and optimization of parallel clustering algorithms.

## Acknowledgements

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian Federation within the framework of implementation of the Program for Improving the Samara university Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.

## References

- [1] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Communications of the ACM* 2008; 51(1): 107–113.
- [2] Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. *IEEE Internet Computing* 2013; 5: 62–69.
- [3] Chubukova I. Tasks of Data Mining. Classification and clusterization. URL: <http://www.intuit.ru>.
- [4] Belim SV, Kutlunin PE. Boundary extraction in images using a clustering algorithm. *Computer Optics* 2015; 39(1): 119–124. DOI: 10.18287/0134-2452-2015-39-1-119-124.
- [5] Protsenko VI, Kazanskiy NL, Serafimovich PG. Real-time analysis of parameters of multiple object detection systems. *Computer Optics* 2015; 39(4): 582–591. DOI: 10.18287/0134-2452-2015-39-4-582-591.
- [6] Protsenko VI, Serafimovich PG, Popov SB, Kazanskiy NL. Software and hardware infrastructure for data stream processing. *CEUR Workshop Proceedings* 2016; 1638: 782–787. DOI: 10.18287/1613-0073-2016-1638-782-787.
- [7] Wang H. Introduction to Word2vec and its application to find predominant word senses, 2014.
- [8] Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. *Association for Computational Linguistics (ACL)* 2014; 545–550.
- [9] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of machine Learning research* 2003; 3: 993–1022.
- [10] Gong S. et al. Linear Discriminant Analysis (LDA).
- [11] Reference systems: LDA. Surfingbird Blog. Habrahabr. URL: <https://habrahabr.ru/company/surfingbird/blog/150607/> (23.11.2016).

# Research and analysis of links in social networks

M.I. Khotilin<sup>1</sup>, A.V. Blagov<sup>1</sup>, I.A. Rytsarev<sup>1</sup>

<sup>1</sup>Samara National Research University, Moskovskoye shosse, 34, 443086, Samara, Russia

---

## Abstract

This paper is devoted to the analysis of data and links in social networks. The approach of representation of a social network in the form of a graph is considered. The algorithms for finding communities and main nodes ("hubs"), which are the accounts that have the greatest impact on communities, have been explored and planned for finalization. Existing software environments for visualizing social network data are explored, a software package is developed.

*Keywords:* social networks; big data; graph; adjacency matrix; SCAN-algorithm; Gephi

---

## 1. Introduction

Over the past decade, social networks have played a huge role in the life of society. They, being the subject of socialization of people, occupy one of the leading positions in the production of "big data". The ability to spread and share messages, photos, music, videos with friends, and create and conduct various events, including for the purpose of business promotion - all this represents a huge amount of constantly generated, aging, updated data. Large amounts of data, including social networks data, as well as the relationships (links) between them must be presented in the form convenient for perception [1-6].

Often, when it comes to objects representing a network, for example, a social one, the notion of data visualization is closely related to the notion of graphs. The network represented as a graph is simple for perception and further analysis. An important task is to represent links in social networks to identify various kinds of dependencies.

## 2. Collecting data from a social network

To represent a social network in the form of a graph, many different tools and tools can be used. In the framework of this work, the following was used to solve this problem: an application developed in C # that allows obtaining the necessary data and performing their analysis; a tool for visualizing Gephi data for graphically representing the graph of dependencies (the so-called graph of the user's friends).

The software itself is visually an authorization form on which the user's login and password of the user account are entered (figure 1).

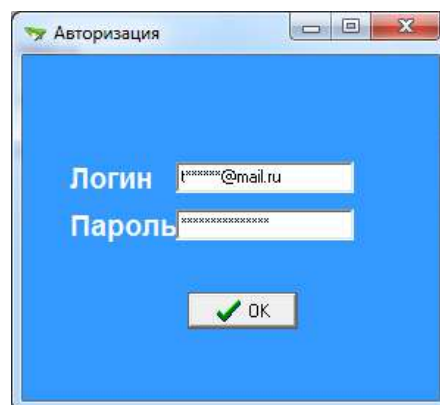


Fig. 1. Software interface.

After entering the login and password, the authorization of the user in the social network and access to the necessary information, namely: to the list of the user's friends, the list of communities, photos, messages, etc., takes place via the open OAuth authentication protocol version 2.0. In the framework of this work, we were interested in the possibility of extracting the user's friends from the social network, so the remaining items were ignored.

Each user of the social network has a unique identifier, or else an ID, which allows you to uniquely identify the user. Using the properties of the built-in API of the social network "Vkontakte", you can, knowing the user ID, extract information about his friends, up to nesting level N. In other words, you can extract a list of friends (N = 1), friends of friends (N = 2), etc. We were interested in the list of friends up to the level of nesting N = 2.

The list of friends extracted from the social network and converted, takes the form of a text file containing the user ID, authorized in the social network and then in the tabular form of the user ID and his name (full name). Knowing the user's ID, you can also find out the list and his friends, which is similarly recorded in the file. An example of the output file part by user and each of the user-friends is shown in figure 2.

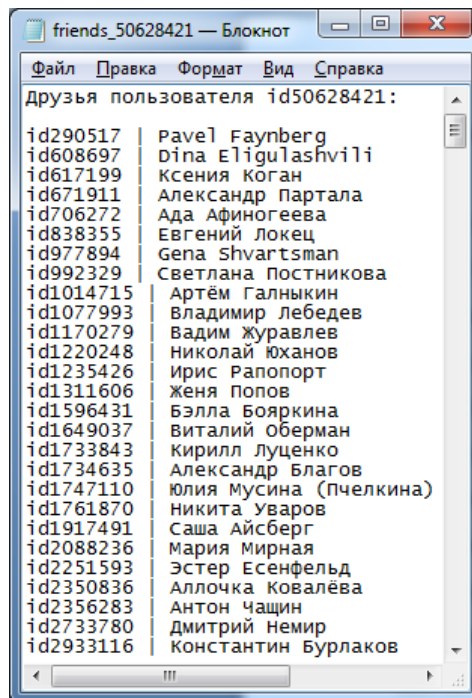


Fig. 2. An example of the file, which contains the information about the friends of the user.

Further, by concatenating the files, a common list of all friends is obtained, along which a list of all friends (dimension  $K$ ) is constructed, from which an adjacency matrix (of dimension  $K \times K$ ) is constructed, on which the dependency graph is subsequently built, by the Gephi software.

An adjacency matrix is a  $K \times K$  dimension matrix containing a list of friends horizontally and vertically, and a row of 0 or 1 at the intersection of the row and column. The contents of the cell of the table matrix is 0 if the users are not familiar (not included in the list of common friends between the user and the friend of the user) and 1 otherwise if there is a "friendship" relationship between the specified users. After construction, this matrix is saved in the .csv format for further loading into Gephi. An example of an adjacency matrix is shown in figure 3.

	Air Sola	Ildar Khalitov	Igor Rytsarev	Svetlana S	Alexey Sa	Anastasiya	Andrey M	Maksim R	Alexander	Yuri Nagulov
Air Sola	0	0	0	1	0	0	0	0	0	1
Ildar Khalitov	0	0	0	0	0	0	0	0	0	0
Igor Rytsarev	0	0	0	0	0	1	1	1	1	1
Svetlana Sukhanova	1	0	0	0	0	0	0	0	0	0
Alexey Satonin	0	0	0	0	0	0	0	1	0	0
Anastasiya Kireeva	0	0	1	0	0	0	1	1	1	1
Andrey Mukhataev	0	0	1	0	0	1	0	1	1	1
Maksim Raguzin	0	0	1	0	1	1	1	0	1	1
Alexsander Nagulov	0	0	1	0	0	1	1	1	0	1
Yuri Nagulov	1	0	1	0	0	1	1	1	1	0

Fig. 3. Matrix of adjacency of the graph of friends.

### 3. The construction of a graph, classification of vertices, finding the most significant vertices

Constructed on the basis of the list of all friends, the contiguity matrix of the future graph is loaded into the Gephi software tool, in order to further visualize the graph of dependencies. Gephi is a software product for network analysis and visualization of data, written in the high-level Java language [7]. The constructed Gephi graph looks like this (figure 4):

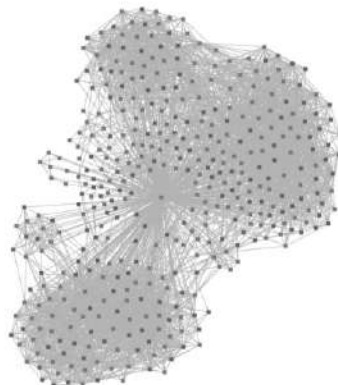


Fig. 4. The graph of users dependencies.

In this graph, the vertices are users of the social network, and the edges are the relation "friendship" between users.

It is worth noting that friends of friends of the user who do not have common relations with the user did not interest us in the framework of this work, so these tops-friends of friends were deleted from the graph.

The next step is to classify the vertices of the graph. The following classification is proposed in the paper:

- core is a vertex containing at least  $\mu$  vertices in an  $\varepsilon$ -neighborhood
- hub is a separate node whose neighbors belong to two or more different clusters;
- outlier - this is a separate vertex, all neighbors of which belong to the same cluster, or do not belong to any cluster [8].

To implement such a classification, the SCAN algorithm is used [9].

The principle of the SCAN algorithm is described below.

Search begins with an initial visit of each vertex once [8], in order to find structurally-connected clusters, and then visit isolated vertices to identify them (hub or outlier).

SCAN performs one network pass and finds all structurally-related clusters for a given parameter. In the beginning all vertices are marked as unclassified. The SCAN algorithm classifies each vertex as either a member of a cluster, or as not being a member [5]. For each vertex that is not yet classified, SCAN checks whether this vertex is a kernel. If the vertex is the core, the new cluster expands from this vertex. Otherwise, the vertex is marked as not being a member of the cluster.

To find a new cluster, SCAN starts with an arbitrary kernel  $V$  and looks for all vertices that are structurally reachable from  $V$ . This is quite enough to find a complete cluster containing the vertex  $V$ . A new cluster ID is generated, which will be assigned to all found vertices.

SCAN begins by setting all the vertices in the  $\varepsilon$ -neighborhood of the vertex  $V$  in the queue. For each vertex in the queue, all directly reachable vertices are calculated, and those vertices that have not yet been classified are inserted into the queue. This is repeated until the queue is empty [8].

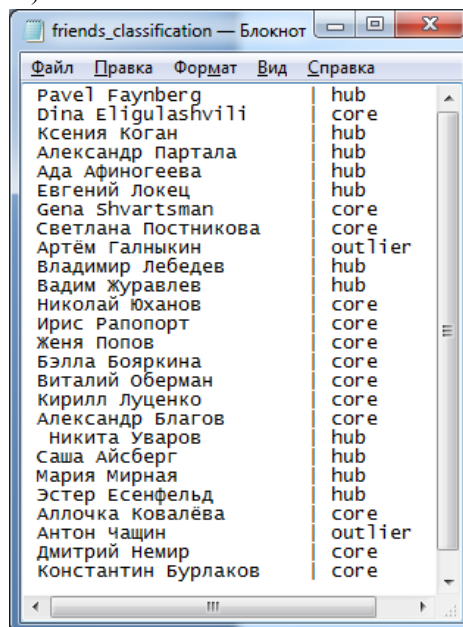
Vertexes that are not members of clusters can be additionally classified as hubs or extraneous. If a single vertex has edges on two or more clusters, it can be classified as a hub. Otherwise, it's an outsider.

A distinctive feature is the presence of parameters and that can be set by the user or expert. In this case, the optimal value of these parameters can be found by machine learning the system using certain network segments.

Since Gephi is an opensource platform [7], one of its great advantages is the ability to write its own modules that implement various algorithms. Thus, using the algorithm from [9], a module that implements the SCAN algorithm was written.

#### 4. Results and Discussion

The result of the SCAN algorithm work on the graph, constructed in Gephi, is a text file containing a list of the user's friends and typing it as the top of the graph (figure 5).



Имя	Классификация
Pavel Faynberg	hub
Dina Eligulashvili	core
Ксения Коган	hub
Александр Партала	hub
Ада Афиногеева	hub
Евгений Локец	hub
Gena Shvartsman	core
Светлана Постникова	core
Артём Галныкин	outlier
Владимир Лебедев	hub
Вадим Журавлев	hub
Николай Юханов	core
Ирис Рапопорт	core
Женя Попов	core
Бэлла Бояркина	core
Виталий Оберман	core
Кирилл Луценко	core
Александр Благов	core
Никита Уваров	hub
Саша Айсберг	hub
Мария Мирная	hub
Эстер Есенфельд	hub
Аллочка Ковалёва	core
Антон Чащин	outlier
Дмитрий Немир	core
Константин Бурлаков	core

Fig. 5. The results of SCAN-algorithm.

It is worth noting that the SCAN algorithm has a certain limitation on the dimension of the graph used. On graphs with high dimensionality ( $N > 500$ ) there is an error in the work.

One way to solve this problem is to modify the algorithm for parallel computations [10]. The idea of this modification is to split the sets of communities into subsets between processors. However, one should take into account that for balancing these subsets should have roughly the same sum of squares of community sizes.

The authors set the task of creating a distributed modification of the SCAN algorithm for ultrahigh-dimensional graphs.

## 5. Conclusion

Social networks and connections in them are the subject of research in this research work. The approach based on the representation of social networks in the form of graphs, makes it possible to apply algorithms for clustering graphs of high dimension. The algorithms described in the paper make it possible to classify segments of a social network, and also to find elements of the greatest interest, for example, users that affect all elements of the same community. These algorithms are planned to be finalized for subsequent application in solving practical problems of finding communities in the segment of social networks in the Samara region.

The Gephi tool, which makes it possible to implement visualization of social networks, was explored, and a software tool that allows to present data in the form required for research was developed.

## Acknowledgements

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian Federation within the framework of implementation of the Program for Improving the Samara university Competitiveness among the World's Leading Research and Educational Centers for the Period of 2013-2020s.

## References

- [1] Tan W, Blake MW, Saleh I, Dustdar S. Social-network-sourced big data analytics. *IEEE Internet Computing* 2013; 5: 62–69.
- [2] Blagov A, Rytcarev I, Strelkov K, Khotilin M. Big Data Instruments for Social Media Analysis. *Proceedings of the 5th International Workshop on Computer Science and Engineering* 2015; 179–184.
- [3] Ivanov PD, Lopukhovskiy AG. BigData technologies and various methods of representing big data. *Engineering Journal: Science and Innovation*, 2014; 9.
- [4] Agafonov AA, Myasnikov VV. Method for the reliable shortest path search in time- dependent stochastic networks and its application to GIS-based traffic control. *Computer Optics* 2016; 40(2): 275–283. DOI: 10.18287/2412-6179-2016-40-2-275-283.
- [5] Popov SB. The Big Data methodology in computer vision systems. *CEUR Workshop Proceedings* 2015; 1490: 420–425. DOI: 10.18287/1613-0073-2015-1490-420-425.
- [6] Ilyasova NYu, Kupriyanov AV, Paringer RA. Formation of features for improving the quality of medical diagnosis based on discriminant analysis methods. *Computer Optics* 2014; 38(4): 851–855.
- [7] An open platform for presenting data in the form of graphs GEPHI. URL: <https://gephi.org> (13.02.2017).
- [8] Khotilin MI, Blagov AV. Visual presentation and cluster analysis of social networks. *Information Technologies and Nanotechnologies*, 2016; 1067–1072.
- [9] Xu X, et al. Scan: a structural clustering algorithm for networks. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007; 824–833.
- [10] Drobyshevskiy MD, Korshunov AV, Turdakov DY. Parallel modularity computation for directed weighted graphs with overlapping communities. *Proceedings of the Institute of System Programming RAS* 2016; 28(6): 153–170.

# The analysis of profiles on social networks

V.A. Bakayev<sup>1</sup>, A.V. Blagov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The article is devoted to the analysis of data and communications on various social networks. The method of search of the profiles belonging to the same users, based on the analysis of the communications and communities which are available for a profile is offered. The program complex realizing this method is created.

*Keywords:* data mining; social networks; graphs; Label Propagation algorithm; Apache Spark

## 1. Introduction

Now one of the most urgent directions in information technologies is the analysis of data or Data Mining. The analysis of data represents process of detection of data, suitable for use, in large data sets, often diverse. Usually such data can't be found at traditional viewing and search as communications are too difficult, or because of the excessive volume.

On social networks big data flows are generated (profiles, communications, content are created). Analyzing these data it is possible to obtain a lot of useful information as on various groups, communities and discussions, and on each user separately [1-4].

The great interest in social networks is shown by various commercial organizations using them as the instrument of interaction with audience. Applying specialized services, the companies analyze information on users, their activities and personalize offers for separately taken segments of the target audience, thereby increasing conversion and reducing costs of advertising campaign.

In the article the method of increase in efficiency of this sort of tools and services which is based on psychology and patterns of human behavior is offered.

The offered method is based on the following facts:

- many Internet users have accounts on several popular social networks at once (VKontakte, Facebook, Instagram and Twitter);
- many users of social networks hide information on themselves from strangers (including information on existence of accounts in other social networks);
- as social networks are a subject of socialization of people, for each user it is possible to allocate at least one community of people it that users of it community are in pairs familiar with each other;
- the person has accounts in different social networks and contacts to the same people.

Is developed the program complex which for realization of a method:

- a) analyzes all profiles of target social networks and on the basis of public data finds the accounts belonging to the owner of an initial profile;
- b) for users on whose pages there is no information on existence at them of accounts in other social networks finds communications with other social networks on the basis of communities in which the user consists.

## 2. The object of the study (Model, Process, Device, Sample preparation etc.)

At the first stage the system analyzes all users of social networks VKontakte, Twitter and Instagram and groups them in the following rules:

- in each group there are no more than one profile from each social network;
- all profiles in one group belong to one person.

This problem is solved by means of a program framework of Apache Spark (in particular, superstructures of Spark Streaming intended for stream data processing see fig. 1) and the broker of messages RabbitMQ realizing delivery of basic data in Spark Streaming [5].

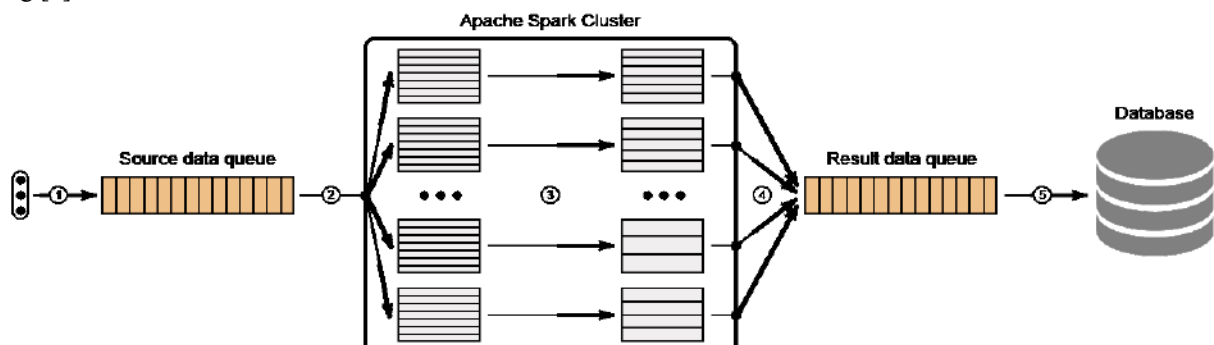


Fig. 1. Architecture of the aggregator of users of social networks (pipeline).

Description of steps:

1. Adding of data from different sources in queue for later processing. Data represent a set of couples (network\_id, user\_id) containing information on profiles which are required to be analyzed.
2. RDD (Resilient Distributed Dataset) formation by a packing of the basic data which are in queue for increase in productivity.
3. RDD (mapping) conversion. For each couple (network\_id, user\_id) the algorithm finds and groups profiles on other social networks, and also additional information on the person to whom belongs the initial account. The algorithm is restarted for each found profile until all available information on the user is found. As sources can be: the public information specified on the page of the user (the status, contact information, entries in the film, etc.).
4. Export of data retrieved from RDD in queue for the subsequent saving.
5. Saving results in NoSQL to the MongoDB database in the form of documents with structure, the reflected in table 1.

The speed of data processing makes about 120-130 profiles a second. For work the Microsoft Azure A2 v2 virtual computer was used (2 kernels, 4 GB of RAM, 20 GB of SSD). Casual users of social network VKontakte (1.000.000 profiles) were analyzed.

Thus, if to assume that speeds of processing of profiles of VKontakte, Instagram and Twitter are equal, we will receive an approximate assessment of time which will be required for the analysis of all users of target social networks:

$$T(n) = \frac{4 * 10^8 + 6 * 10^8 + 13 * 10^8}{120n} \approx 5324 \text{ hours} \approx 221 \text{ day}, \quad (1)$$

where n - the number of servers in a cluster with a similar configuration.

In case of horizontal scaling of a cluster the linear dependence between the number of servers and processing rate of profiles is watched.

Example of the reference to the table: results of an experiment are reflected in table 1.

Table 1. Data storage structure of profiles.

Name of the field	Type	Description
_id	ObjectId	the document identifier in a collection
vk_id	Int32	the profile identifier in VKontakte
facebook_id	Int64	the profile identifier in Facebook network
instagram_id	Int64	the profile identifier in Instagram network
twitter_id	Int64	the profile identifier in Twitter network
other	Object	The additional information (phone number, e-mail address, skype, etc.)

The further task comes down to expansion of the received base by association of profiles by the rules described earlier on which pages bek-links aren't specified other social networks.

### 3. Methods

Based on a hypothesis that the person consists in the same communities on all social networks which uses, we can establish connection between profiles of the different social networks which are in one community and with a certain probability to assume that they belong to one person.

The purpose of this stage is separation of communities among the people who are in the database received from the previous step. Extension of Apache Spark GraphX which is intended for distributed processing of graphs is for this purpose used.

The algorithm of preliminary data handling:

1. Generation of a graph on the basis of the available data. Peaks represent an entity of "people" and store identifiers of the profiles belonging to the specific user. Connection between peaks is established by the following principle: two peaks are adjacent if profiles of matching social networks are coherent. We will determine edge weight between peaks as:

$$w(A, B) = |\{i: i \in [0, n - 1] \wedge \exists A_i, B_i \wedge rel(A_i, B_i)\}|, \quad (2)$$

where rel (a, b) - function, which the truth in only case when when profiles an and b are interconnected (the relation of friendship or manifestation of activity is established).

2. The algorithm Label Propagation [6] realized in GraphX API which solves the problem of a clustering is applied to the received graph and finds communities in the graph.

3. For each community RDD with a set of profiles of VKontakte, Facebook, Instagram and Twitter which are connected to one or several members of the initial community is generated.

Thus the data set (dataset) on the basis of which we can speculate about accessory of group of accounts of different social networks (without obvious indication of interrelations) to one person is formed.

Grouping of the common features given on the basis of the analysis is final stage in the solution of an objective. The following procedure is applied to each data set from a dataset:

1. Creation of the full multiple-count graph in which information on profiles of social networks and the potential characterizing probability of their accessory to one person is stored;

Tops of the graph contain information on profiles which is used at their comparison.

For comparison of two profiles the multilayered neural network is used. On an entrance layer of network the vector of dimension 12 containing the following data moves:

- Name  $\leftrightarrow$  Name'
- max (Name  $\rightarrow$  Username', Name'  $\rightarrow$  Username)

- max (Name → E-mail', Name' → E-mail)
- max (Name → Skype', Name' → Skype)
- Username ↔ Username'
- max (Username → E-mail', Username' → E-mail)
- Username ↔ Skype'
- max (Skype → Username', Skype' → Username)
- max (Skype → E-mail', Skype' → E-mail)
- E-mail ↔ E-mail'
- Phone ↔ Phone'
- Website ↔ Website'

We will determine operations  $a \rightarrow b$  and  $a \leftrightarrow b$ :

1.1 Completeness of entry of a into b:

$$a \rightarrow b = 1 - \frac{d+r+s}{\text{len}(a)} \in [0, 1], \quad (3)$$

where d - the number of operations of removal for transformation a to b;

r - the number of operations of replacement for transformation a to b;

s - the number of operations of a transposition for transformation a to b;

len(x) - function of calculation of length of an argument.

1.2 Comparison of an and b:

$$\forall i \in [1, \text{len}(a)], j \in [1, \text{len}(b)] \quad d[i, j] = 1 - \frac{\text{dist}(a[i], b[j])}{\text{len}(b[j])} \in [0, 1], \quad (4)$$

$$a \leftrightarrow b = \frac{\sum_i^{\text{len}(a)} d[i, \text{fit}(i)]}{\min(\text{len}(a), \text{len}(b))} \in [0, 1], \quad (5)$$

where dist (a, b) - the function calculating Damerau-Levenstein [7] distance for lines an and b;

fit(i) - the function returning an index of the word of a line b put in compliance to the word a[i].

Operation of comparison doesn't consider a word order. All words of initial lines are in pairs compared, and then, by means of Kuhn-Munkres [8] algorithm, to each word of a line the word of a line b is put in compliance so that the similarity sum on all couples of words was maximum. Also punctuation marks and other symbols aren't considered (except for letters and figures).

Before processing all symbols of entrance data are given to Latin by rules of a transliteration. The summary table of the main alphabets is for this purpose used (Russian, Ukrainian, Bulgarian, Indian, Arab).

#### 4. Results and Discussion

The training and control selections are collected on the basis of primary data. The amount of the training selection ~ 106 couples.

As negative examples both casual couples of profiles, and couples found by means of full text search in different parameters were used (name, username, email, skype).

In generated to the column for each couple of shares the following algorithm is carried out:

1. edges are sorted in decreasing order of scales;
2. edges which weight is less than threshold value are removed or one of incidental tops is already connected with any top of an opposite share.

As a result of these transformations the count in whom everyone a component of connectivity represents group of accounts of different social networks which belong to one person turns out.

Because the person can belong at the same time several community and also if the same group has been created in several communities, then it is possible to believe that accounts of this group really belong to one user.

The data obtained at the last stage register in the same base where primary data are stored. However they aren't used as entrance data for the realized algorithm in view of the unreliability.

#### 5. Conclusion

The processing and data analysis of social networks allows to personalize a product or service for a specific segment of target audience. The program complex received as a result of operation erases boundaries between social networks for different services, allowing them to integrate API and to operate with an entity of "people", but not "profile" that does their operation more effective.

The further research can be continued regarding upgrade of algorithms of aggregation and the analysis of data retrieved. It is necessary for implementation of the decision of the following tasks:

- the detection of popular social networks and services which users mention on the pages (for example, LinkedIn, Last.fm) and development of parcers for them;
- the analysis of additional information sources on pages of users in VKontakte, Instagram and Twitter (for example, a news feed on Twitter);
- the analysis of pages of users on target social networks and search of additional parameters based on which one user can compare profiles on accessory.

#### Acknowledgements

The work has been performed with partial financial support from the Ministry of Education and Sciences of the Russian



## References

- [1] Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. *IEEE Internet Computing* 2013; 5: 62–69.
- [2] Khotilin MI, Blagov AV. Visualization and Cluster Analysis of Social Networks. *CEUR Workshop Proceedings* 2016; 1638: 843–850.
- [3] Protsenko VI, Serafimovich PG, Popov SB, Kazanskiy NL. Software and hardware infrastructure for data stream processing. *CEUR Workshop Proceedings* 2016; 1638: 782–787. DOI: 10.18287/1613-0073-2016-1638-782-787.
- [4] Popov SB. The Big Data methodology in computer vision systems. *CEUR Workshop Proceedings* 2015; 1490: 420–425. DOI: 10.18287/1613-0073-2015-1490-420-425.
- [5] Apache Spark Documentation. Access mode: <http://spark.apache.org/docs/2.1.0>.
- [6] Smetanin N. Fuzzy search in the text and the dictionary. URL: <https://habrahabr.ru/post/114997>. (in Russian)
- [7] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A: Statistical Mechanics and its Applications*. 389(7): 1493–1500.
- [8] Hungarian algorithm for solving the assignment problem. URL: [http://e-maxx.ru/algo/assignment\\_hungary](http://e-maxx.ru/algo/assignment_hungary). (in Russian)

# Algorithms of the information stimulation system of Russian citizens' socio-optimal actions

M.I. Geraskin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The problem of the development of the state information stimulation system of Russian citizens' socio-optimal actions is considered according to the optimum of collective utility function as criterion. Conceptual model of the system is formed according to the conditions of individual rationality, Pareto efficiency, nonmanipulability and dynamic quasi-optimality. The algorithms of the information system are developed such as a process of step-by-step approximations. The system's criterion on each approximation does not decrease and the constraint on the stimulation fund is fulfilled.

*Keywords:* information system; distribution incentives mechanism; additive collective utility function; stimulation system; nonmanipulability; quasi-optimality

---

## 1. Introduction

In a transitional economy [1,2] trends of individual rationality are growing in the society. To overcome this trends the state develops the moral improving programs [3,4]. In this case, the purpose of the state is the social effect of citizens' acts, performing on the basis of maximization of collective but not individual utility function, hereinafter referred to as socio-optimal actions. Achieving this purpose requires the involvement in socially useful activity of large group of the population and personified registration of socio-optimal actions. It needs to organize the state information system, based on the information resources of currently working in Russia programs [5-7].

The concept of socio-optimal actions' stimulation provides for the establishment of information system of personified registration of the actions of citizens (hereinafter, agents). The system also includes the distribution of state stimulation fund in the form of incentives between agents according to certain mechanisms. The dynamics of the system is a two-period. In the first period (registration period) performed socio-optimal actions are recorded, and in the end of this period the stimulation fund is distributed. In the next period (period of stimulation) the earlier distributed incentives are used.

In addition to the utilitarian stimulating function information system also solves the problem of the formation of the agent's status in the hierarchy of the citizens, used for non-material motivation. On a longer time horizon, the state's social priorities could changed by varying the attributes of socio-optimal actions and their monetary valuation can be varied as a result of inflation. Therefore, to comparability of agents' statuses the system accumulates not only incentives as the current cash equivalent of social activity, but also agents' rating in comparable dimension.

The object of stimulation is socio-optimal actions of citizens, that is, actions that correspond to certain attributes. The actions should maximize collective utility function without increasing the individual utility function. Therefore, the attributes correspond to the terms of gratuitousness, public utility and unconnectedness with professional activities of citizens. Consequently, socio-optimal actions do not require special qualification, whereby the stimulation object's dimension is duration of action excluding the content of the action. The subject of stimulation is citizen, performing a socio-optimal action in certain period. The apparatus of stimulation is the state represented by certain ministries (departments).

## 2. The object of the study

The system under consideration is constructed outside epy architecture of market relations. Therefore, formed by market equilibrium relationship between the socio-optimal actions and incentives as their monetary valuation does not exist. For this reason, in this system it is possible some disproportions. Firstly, dynamic inconsistencies between the stimulation fund and the cash equivalent of the socio-optimal actions leads to a deficit or proficit of the fund; the deficit will not allow; the proficit expresses disinterest of citizens in stimulation. Secondly, the impossibility of compliance control between registered in the information system and the actual socio-optimal actions leads to inaccurate registration (overcharged) information. Therefore, for system's equilibrium incentive distribution mechanism must meet the following conditions: 1) individual rationality, in which the agents' utilities with the incentives are not lower than any alternative, that is, agents are interested in stimulation; 2) Pareto efficiency, which means that stimulation fund if distributed fully between agents, that is, there is no deficit and proficit; 3) nonmanipulability (compatible with incentives), in which each agent reported accurate information about its action according to criterion of individual rationality; 4) optimal distribution according to criterion of collective (additive) utility function. Thus, the object of study is the state information system providing equilibrium humanitarian goals of the state, the tools of their realization in the form of incentive distribution mechanism, as well as socio-optimal actions performed by citizens.

### 3. Methods

The investigations of stimulation systems and distribution mechanisms produce the following mechanisms corresponding the individual rationality. Competitive mechanism is developed with noncooperative [8] and cooperative [9] behavior of agents, its Pareto efficiency and optimality according to additive utility function criterion are proved. The step-by-step resource distribution mechanism (SRDM) is obtained [10], for which proved [11] that nonmanipulability and Pareto efficiency simultaneously only for SRDM; also SRDM, as shown in [12], is equivalent to mechanisms of direct and reverse priorities. It was shown [13] that unique SRDM exists, in which the incentive is distributed [14] as minimum of agent's information and the average undistributed rest of incentives. The approach to the distribution based on the penalty and incentive functions [15] showed the Pareto efficiency and optimality according to additive utility function criterion for compensatory mechanisms; according to a compensatory mechanism incentives are equal to agents' costs. Thus, only SRDM satisfies [16] all above conditions. Since SRDM implies consistent registration of agents' actions and further distribution of the incentives, it is impossible to use in the system, where actions perform independently and record simultaneously. Therefore, the development of adaptive distribution algorithm, satisfying the individual rationality, Pareto efficiency, nonmanipulability and additive optimality is important.

The model of information system of stimulation is considered. We introduce the following sets. The set of socio-optimal actions attributes

$$Z = \{z_i, i = 1, \dots, I\}$$

defines the attributes of action to be stimulating; index  $I$  is the number of the types of actions. The set of agents

$$K(t) = \{1, \dots, n(t)\}$$

includes citizens, performing in the period  $t$  actions corresponding  $Z$ ; index  $n(t)$  is the number of agents. The vector of socio-optimal actions

$$A(Z, t) = \{a_k(Z, t), k \in K\}$$

includes quantitative estimates of  $k$ -th agent's actions corresponding  $Z$  in  $t$ -th period in terms of time taken to perform these actions. The vector  $A(Z, t)$  is contained in allowable set

$$\bar{A} = \{a_k \in [0, a^{\max}], a^{\max} > 0, k \in K\},$$

where the symbol  $a^{\max}$  denotes the upper limit of agents' disposable time. For example, the vector  $Z$  may include attributes such as  $z_1 = \text{«carrying out socially useful activities»}$ ,  $z_2 = \text{«provision of free services to the citizen»}$ ; the components of the vector  $A(Z, t)$  express the time registered in the  $t$ -th period.

The stimulation fund in the  $t$ -th period is

$$F(t) \in (0, F^{\max}], F^{\max} > 0.$$

Then we omit the index  $t$ , assuming that all the parameters of the model correspond to a specific period.

We introduce the dimensionless registration function of socio-optimal actions

$$u_k = \psi(a_k), k \in K,$$

which the score value  $u$  corresponds to a time value  $a$ . Thus, the vector of action  $A$  corresponds to the vector of scores

$$U = \{u_k, k \in K\} \in \bar{U},$$

where  $\bar{U}$  – allowable set of scores. Let a function  $\psi(\bullet)$  is continuously differentiable, satisfies the conditions of saturation and set is defined  $\bar{U}$  as follows:

$$u_k = \psi(a_k): \psi'_a(\bullet) > 0, \psi''_a(\bullet) < 0, k \in K, \bar{U} = \{u_k \in [0, u^{\max}], u^{\max} > 0, k \in K\} u^{\max} = \psi(a^{\max}) > 0. \quad (1)$$

We introduce the social utility function of agent  $f(a, x)$  as the sum of the individual utility function and distributed incentive

$$f(a, x) f_k(a_k, x_k) = f_k^0(a_k) + x_k(u_k), k \in K, \quad (2)$$

where  $f^0(\bullet)$  - individual utility function;  $x(u)$  - distributed incentive.

For the function (2) property of individual rationality is defined as: the social utility function of agent performing the action  $a$  must be no less than the individual utility function when such action does not perform ( $a = 0$ ), that is,

$$f_k(a_k, x_k) \geq f_k^0(0), k \in K.$$

Therefore, incentives at the individual rationality must satisfy to

$$x_k(u_k) \geq f_k^{\min}, k \in K,$$

where

$$f_k^{\min} = f_k^0(0) - f_k^0(a_k) \geq 0$$

characterize the loss of individual utility when making the socio-optimal action. Agents can't have individual utility losses when making socio-optimal action, for example, performing it in their spare time from gainful activity. Therefore, guaranteed estimate of these losses, the same for all agents, is

$$f^{\min} = \arg \max_{k \in K} f_k^{\min}.$$

We introduce the stimulation function  $x = \varphi(u)$ , which the incentive in cash corresponds to the score  $u$ . Let the function  $\varphi(\bullet)$  is continuously differentiable, satisfies unsaturation condition on  $u$  and saturation condition on the sum of  $u$ , individual rationality and Pareto efficiency:

$$x_k = \varphi(u_k): \varphi'_{u_k}(\bullet) > 0, \varphi''_{\sum_{k \in K} u_k}(\bullet) < 0, x_k(u_k) \geq f^{\min} \geq 0, \sum_{k \in K} x_k(u_k) = F, k \in K. \quad (3)$$

Generally formulas (1) - (3) are the stimulation system model  $S$  that represents the list of agents' set, agents' scores set and agents' utility functions set:

$$S = \langle K, \psi_k(a_k), f_k(a_k, x_k(u_k)), k \in K \rangle. \quad (4)$$

We introduce the additive criterion of social efficiency of the stimulation system as the sum of the socio-optimal actions of agents in the  $t$ -th period

$$E(S) = \sum_{k \in K} a_k(S). \quad (5)$$

Definition: a nonmanipulability - is a property of the system  $S$  having the Nash equilibrium, that is, a state in which there is no agent, for which the social utility function for some action vector  $\tilde{A} \in \bar{A}$  is greater than the equilibrium Nash vector  $A^N \in \bar{A}$ :

$$\exists k \in K: f_k(x_k(\tilde{A})) > f_k(x_k(A^N)), \tilde{A} = (\tilde{a}_k, a_{-k}), \tilde{A}, A^N \in \bar{A}, \quad (6)$$

where symbol «-k» denotes enviroining, i.e. agents other than  $k$ -th.

We define the set of allowable stimulation systems  $\bar{S}$  based on the conditions of Pareto efficiency and individual rationality (3), nonmanipulability (6) in the form of:

$$\bar{S} = \left\{ S: x_k \geq f^{\min}, \sum_{k \in K} x_k \leq F, \exists k \in K: f_k(x_k(\tilde{U})) > f_k(x_k(U)) \right\}. \quad (7)$$

The problem of selecting the model  $S$  from the set  $\bar{S}$  is considered as the optimal control problem

$$S^* = \arg \max_{S \in \bar{S}} E(S). \quad (8)$$

The control parameters in the problem (8) is a functions  $\psi(\bullet), \varphi(\bullet)$ . So the criterion and constraints in problem (8) implicitly depend on the control parameters, and the problem (8) does not, in general, the analytical solution. In these cases, the approximate methods are used [17-21] for solution of optimal control problem, and the resulting solution was called quasi-optimal [22].

Definition: a model of stimulation system  $S$  is called a dynamic quasi-optimal if

$$E'_t(t) \geq 0 \forall S \in \bar{S}. \quad (9)$$

The dynamic quasi-optimality means that in  $t$ -th period additive criterion system (5) is no less the value at  $(t-1)$ -th period when the constraints (7). Therefore, the condition (9) defines the process of step-by-step approximations for system (4), on each of which the criterion (5) does not decrease. We pose the problem of developing an algorithm that implements the model (4) as a dynamic quasi-optimal. Subsequently, the problem of developing an algorithm that implements the model (4) as a dynamic quasi-optimal, is considered.

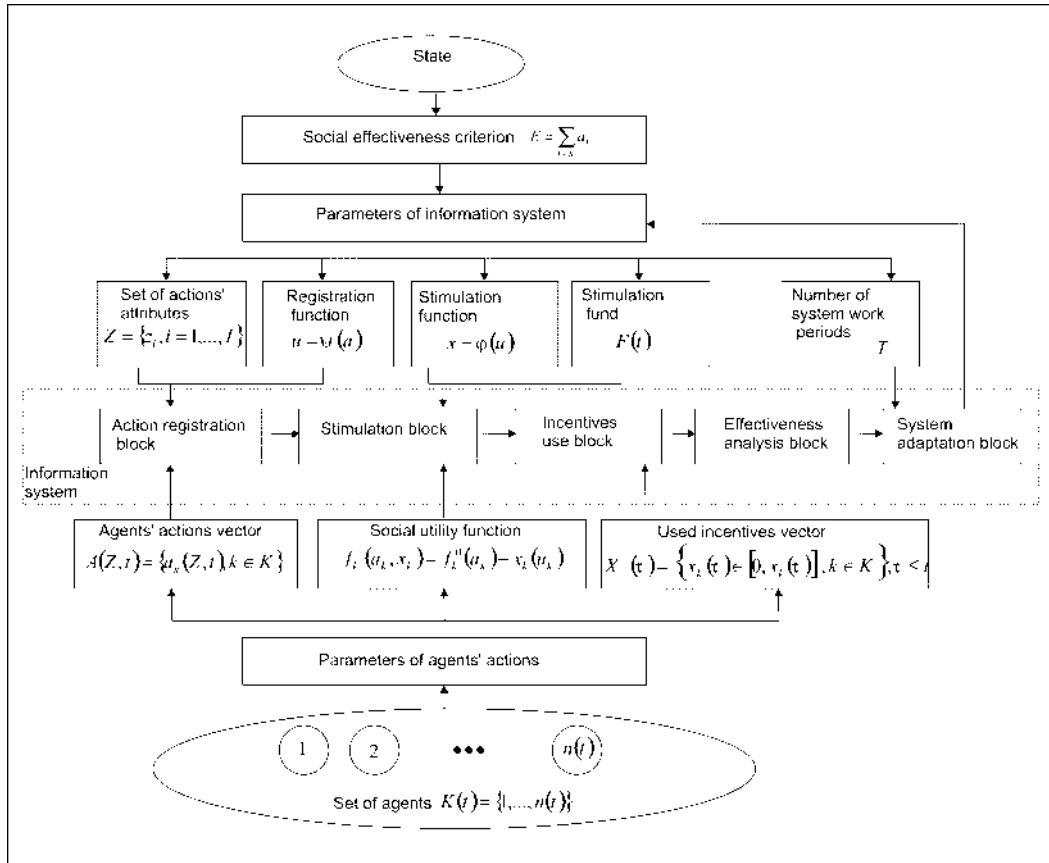


Fig. 1. Conceptual model of information system.

#### 4. Results and discussion

Conceptually, the projected information system (Fig. 1) implements the processes of interaction between state and citizens based on the goal of maximizing the total number of socio-optimal actions [23,24].

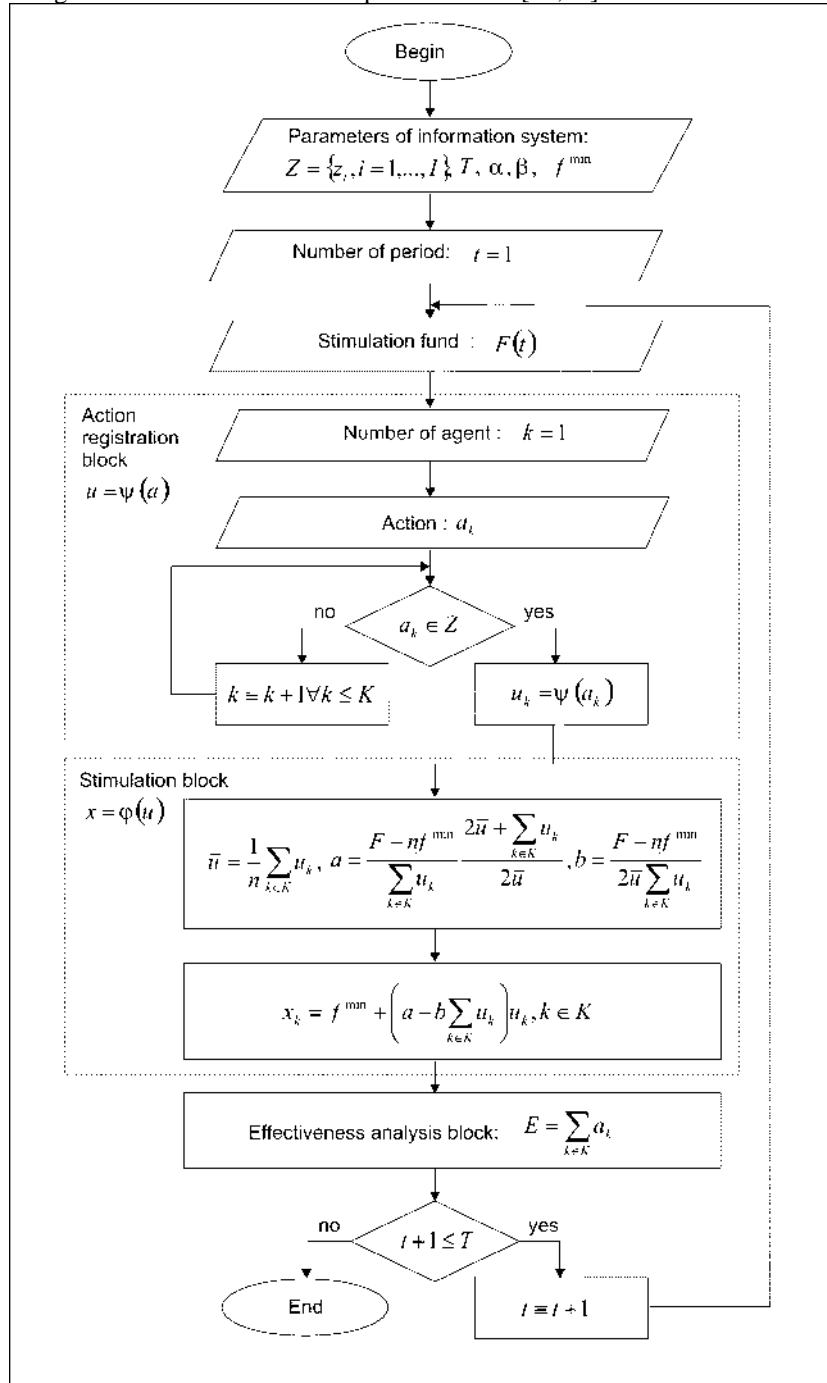


Fig. 2. Algorithm of one-period cycle of the information system.

The state defines such parameters of information system as a set of actions' attributes, the form of registration function, the form of stimulation function, the stimulation fund and the number of system work periods. The agents' actions are parameterized via the vector of action, a set of social utility functions and the vector of used incentives. The information system is the infrastructure of the state and citizens, consisting of five blocks. The action registration block identifies actions in the score values. The stimulation block is designed to distribute the fund, depending on the vector of score ratings. The incentives use block implements the functions of analyzing the dynamics of accrued and used incentives, as well as control the deficit (profit) of the stimulation fund. The effectiveness analysis block controls the dynamics of change in the social effectiveness criterion of system test on the selected time interval of up to the maximum number of work periods. The system parameters adaptation block implements a process of successive approximations for quasi-optimality.

A degree registration function  $\psi(\bullet)$  satisfying the conditions (1) is considered in the form

$$\psi(a_k) = \alpha a_k^\beta, \alpha \in (0, \alpha^{\max}], \beta \in (0, \beta^{\max}], \beta^{\max} \in (0, 1], k \in K, \quad (10)$$

where  $\alpha, \beta$  are constant coefficients.

We introduce the stimulation function  $\varphi(\bullet)$  relating to the class of direct compensatory priority functions [25]:

$$\varphi(u_k) = f^{\min} + \left( a - b \sum_{k \in K} u_k \right) u_k, k \in K, \quad (11)$$

where  $a, b$  are constant coefficients,  $a, b > 0$ , are selected to satisfy  $\varphi(\bullet)$  the conditions (3). Substantially compensatory function (11) defines guaranteed incentive  $f^{\min}$  of non-zero agent's action on the principle of "any socially useful action is rewarded", beyond which the incentive is distributed in proportion to the score  $u$  in accordance with the "cost"  $p$

$$p = a - b \sum_{k \in K} u_k,$$

decreasing with increasing the total number of agents.

It can be shown that if the coefficients  $a, b$  are defined by the formulas

$$a = \frac{F - n f^{\min}}{\sum_{k \in K} u_k} \frac{2\bar{u} + \sum_{k \in K} u_k}{2\bar{u}}, b = \frac{F - n f^{\min}}{2\bar{u} \sum_{k \in K} u_k}, \bar{u} = \frac{1}{n} \sum_{k \in K} u_k, \quad (12)$$

the system  $S$  provides firstly Pareto efficient distributed of stimulation fund, that is, the full distribution without deficit or profit and, secondly, the agents are not interested in the overstated information about performed actions, that is, the system is non-manipulable.

The algorithm for static (one-period) cycle of the information system is shown in Fig. 2. The one-period cycle does not include the incentives use phase, since the distribution of incentives is possible only after the registration of all actions of all agents in that period. The one-period cycle information system algorithm includes the action registration block, the stimulation block and the effectiveness analysis block.

The one-period cycle does not allow to achieve the stimulation system optimality according to criterion (5), in particular, to adapt the parameters of the blocks « $u = \psi(a)$ », « $x = \varphi(u)$ » so that the quasi-optimality condition (9) is performed in dynamics. In addition, the one-period cycle is not taken into account the dynamics of agents' number increasing and the dynamics of actions number, resulting in incentives inflation in case of constant stimulation fund, which may result in loss of efficiency. Also, in the one-period cycle the used incentives dynamics is not coordinated with the dynamics of stimulation fund, which could lead to its deficit or profit, and causes a reduction in efficiency.

The dynamic (multi-period) algorithm of the information system is considered. We introduce the following parameters of the dynamics of the system in the period  $\tau \in (0, t]$ : the accrued incentives vector is  $X(\tau)$ , the components of which are defined by (11), the used incentives vector is  $X^-(\tau)$ , the used stimulation fund is  $F^-(\tau)$ , the unused incentives residues vector at the end of the  $t$ -th period is  $R(t)$ , the total unused residue of incentives at the end of  $t$ -th period is  $R\Sigma(t)$ , the unused residue of stimulation fund at the end of  $t$ -th period is  $\Phi$ . These parameters are determined by the following formulas:

$$X(\tau) = \{x_k(u_k(\tau)), k \in K\}; X^-(\tau) = \{x_k^-(\tau) \in [0, x_k(\tau)], k \in K\}; F^-(\tau) = \sum_{k \in K} x_k^-(\tau), \tau \in [2, t]; \quad (13)$$

$$R(t) = \left\{ r_k(t) = \sum_{\tau=1}^t x_k(\tau) - \sum_{\tau=2}^t x_k^-(\tau), k \in K \right\}; R\Sigma(t) = \sum_{k \in K} r_k(t); \Phi(t) = \sum_{\tau=1}^t F(\tau) - \sum_{\tau=2}^t F^-(\tau), \quad (14)$$

where  $x^-(\tau)$  is incentive used by  $k$ -th agent in the period  $\tau \leq t$ .

In the case of full distribution of stimulation fund (Pareto efficiency), it can be shown that if the following conditions are met in the  $t$ -th period

$$\Phi(t) = R\Sigma(t) \geq 0, E(t) - E(t-1) > 0, t \in (2, T], \quad (15)$$

the information system is quasi-optimal. Violation of the conditions (13) indicates the non-optimality of the system (4) in the  $t$ -period, that is, the need to adapt such system parameters as registration function coefficients  $\alpha, \beta$  and fund  $F$ .

The algorithm of information system multi-period cycle is shown in Fig. 3. Blocks, detailed in the one-period cycle (Fig. 2), are shown in general. The incentives use block implements formulas (13), (14). The system parameters adaptation block is based on the analysis of conditions (15), the variations of system parameters are defined by the formulas:

$$\Delta\alpha \in [0, \alpha^{\max} - \alpha(t-1)], \Delta\beta \in [0, \beta^{\max} - \beta(t-1)], \Delta F \in [0, \min\{F^{\max} - F(t-1), |R\Sigma(t)|\}].$$

The algorithm provides a process of successive approximations, when the constraints are complied, resulting in the system to quasi-optimality state.

Simulation of the stimulation impact on the behavior of the population was carried out by changing the skewness and kurtosis of probability density function of the normal distribution

$$f(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{w(a-\bar{a})^2}{2\sigma^2}}.$$

where  $\bar{a}, \sigma$  – mathematical expectation and mean-square deviation of initial distribution;  $l$  – skewness coefficient ( $l < 1$  - left skewness,  $l > 1$  - right asymmetry) compared with a normal distribution ( $l = 1$ );  $w$  – kurtosis coefficient ( $w < 1$  - a more uniform distribution,  $w > 1$  - less uniform distribution) as compared with the normal distribution ( $w = 1$ ).

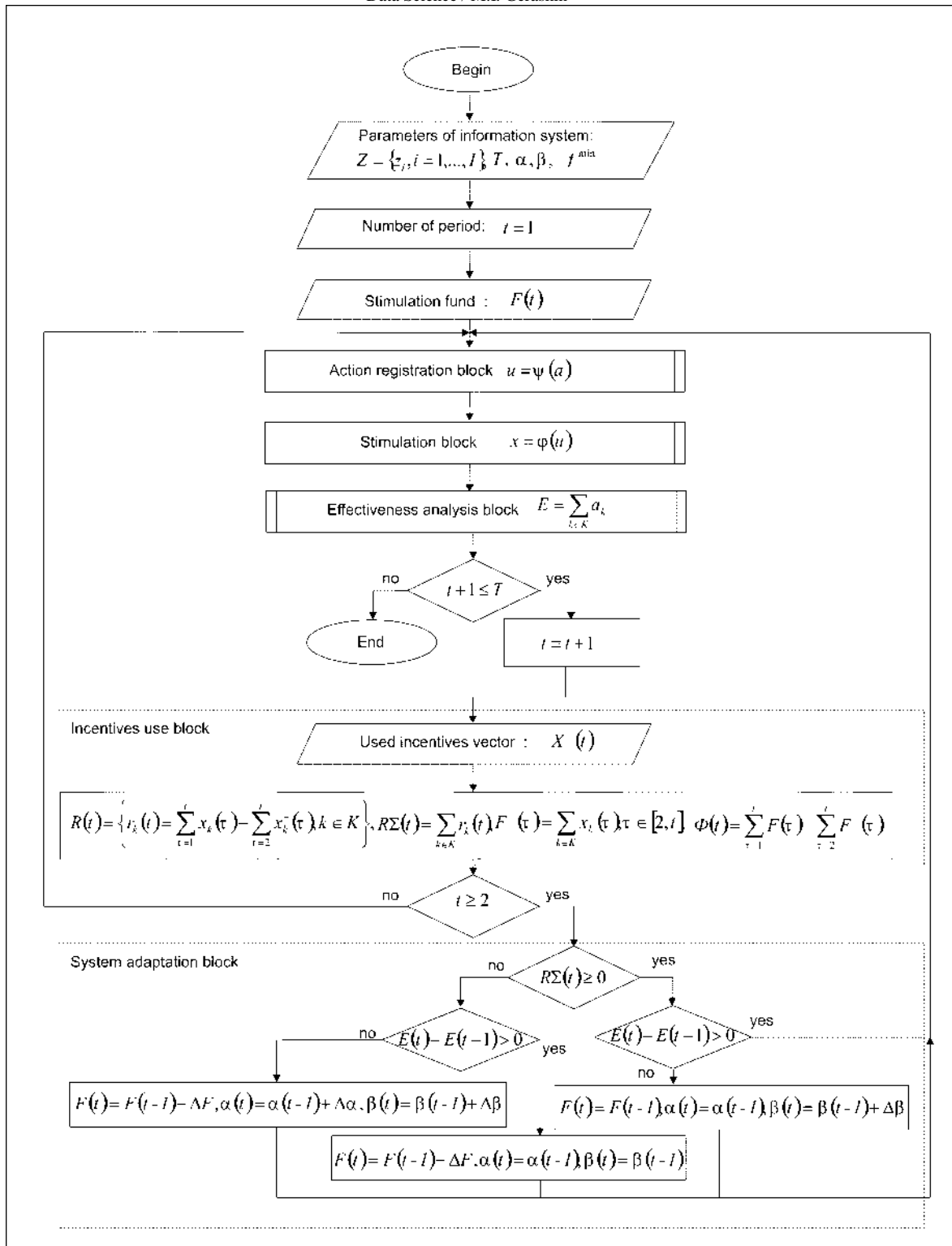


Fig. 3. Algorithm of multi-period cycle of the information system.

The simulation of the distribution of agents in the stimulation process was based on the following hypothesis: registration function coefficient  $\beta$  growth leads to right skewness, i.e., to increase of the expectation of the distribution of the agents' set, compared with the median of range  $[0, a^{\max}]$ ; coefficient  $\alpha$  growth leads to decrease in kurtosis, to increase the variance of the distribution of the agents' set in comparison with the initial value. The agents' number changing in the stimulation process was not more than 20% of the initial number of citizens, which was led to constraint on the coefficients of skewness and kurtosis  $l \in [1, 1.4], w \in [0.7, 1]$ . The simulation of information system was carried out for the following initial data:  $a^{\max} = 8$  hours,  $\bar{a} = 4$  hours,  $\sigma = 1.293$  hours,  $f^{\min} = 1$  thousand rubles,  $n = 9727$  thousand. The initial fund value was established  $F = 1000000$  thousand rubles, the initial registration function coefficients  $\alpha = 18,95, \beta = 0,8$  were chosen from the condition  $u^{\max} = 100$ ; the value of the used incentives coefficient, i.e. share of the used incentives to stimulation fund, was taken  $\Delta X = 0.8$ .

Three scripts of information system dynamics were considered: the first script (Fig. 4.) – the growth  $\beta$  when  $\alpha = const, F = const, \Delta X = const$  led to a right skewness; the second script (Fig. 5.) – the growth  $\alpha$  when

$\beta = const, F = const, \Delta X = const$  led to the reduction of kurtosis; the third script (Fig. 6). – the growth  $\beta$  when  $\alpha = const, F = const$  resulted in right skewness, and with the increase  $\Delta X$  stimulation fund  $F$  reduced on the relative value  $\Delta F$ .

The dynamics of the first script is shown (Fig. 4) in the periods  $t = 1, \dots, 7$ , for which the coefficient  $\beta$  was varied in the range  $[0.8, 0.92]$ , which led to increase in the coefficient  $l \in [1, 1.4]$ . As a result, the maximum average value of system efficiency reached  $E_{av.} = 5.68$ , the average stimulation fund residue  $R_{av.}$  increased to 200, the average score  $u_{av.}$  increased to 111, the score price decreased to  $p=1.28$ .

The dynamics of the second script (Fig. 5) for changing the coefficient  $\alpha$  in the range  $[18.95, 19.25]$  led to decrease in the coefficient  $w \in [0.7, 1]$ . As a result, the following figures were found:  $E_{av.}=4.76, R_{av.}=121, u_{av.}=57, p=1.48$ .

The dynamics of the third script (Fig. 6) repeated the first script of the coefficient  $\beta$  change in the range  $[0.8, 0.92]$ , which led to growth of coefficient  $l \in [1, 1.4]$ . However, at  $t=1, \dots, 6$  stimulation fund has been fixed ( $\Delta F = 0$ ) and the follow used incentives coefficient has been set:  $\Delta X = 0.8$  at  $t=1, 2, \Delta X = 1.1$  at  $t=3, \dots, 7$ . As a result, at  $t=6$  value  $R_{av.} = 0$  was reached, that led to the need at  $t=7$  to finish stimulation ( $\Delta F=1$ ); at  $t=7$  the following figures were obtained:  $E_{av.}=4.91, R_{av.}=0, u_{av.}=111, p=0$ .

The simulation showed the following results: 1) the average efficiency of the stimulation system is more sensitive to a change in the registration function coefficient  $\beta$  by right skewness of the distribution of the agents' set than to a change in the coefficient  $\alpha$  by reducing the kurtosis of the distribution, because of in the first case, the number of agents decreases, while in the second - increases; 2) stimulation system is non-manipulable, because of the score price decreases with increasing registration function coefficients; 3) stimulation fund deficit occurs when an excessive use of incentives, and deficit is compensated by fund decrease in the subsequent period.

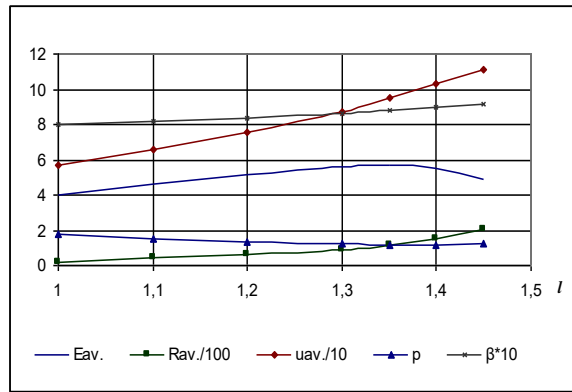


Fig. 4. The dynamics of the information system first script, the values of the coefficient  $l$  correspond to the periods  $t=1, \dots, 7$ .

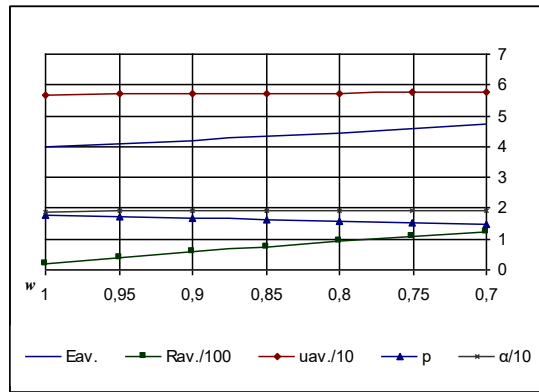


Fig. 5. The dynamics of the information system second script, the values of the coefficient  $w$  correspond to the periods  $t=1, \dots, 7$ .

## 5. Conclusion

The problem of information support of the state strategy of strengthening morality in society was considered. The information system of stimulation of citizens' actions was developed based on collective utility function maximizing. In the article the following results were obtained. The conceptual model of information system was formed on the base of individual rationality, Pareto efficiency, non-manipulablity and dynamic quasi-optimality. The model includes the action registration block, the stimulation block, the incentives use block, the effectiveness analysis block and the system parameters adaptation block.

The specific form of the compensatory stimulation function of the direct priorities class was proposed. In this function, the incentive consists of a guaranteed minimum and proportional to agent' score "premium". The score price decreases with the growth of the total number of agents' scores. This stimulation function implements the mechanism similar to the mechanism of oligopoly market equilibrium [26], but, unlike that, for certain function coefficients the agents' actions vector is Pareto efficient. Unlike the compensatory mechanism [27,28] proposed stimulation function ensures Nash equilibrium actions vector in such case than the stimulation fund does not depend on the actions vector. Thus, formed stimulation system satisfies the conditions of Pareto efficiency and compatibility with incentives.



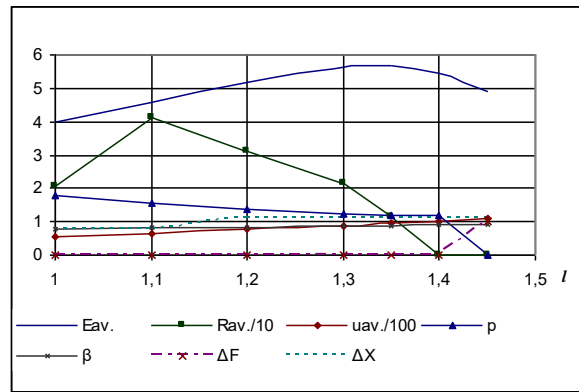


Fig.6. The dynamics of the information system third script, the values of the coefficient  $l$  correspond to the periods  $t=1, \dots, 7$ .

The dynamic algorithm for information system was developed as a multi-period cycle, which includes a one-period cycle of the actions registration and the stimulation fund distribution. The algorithm implements a process of step-by-step approximations with constraints, which result in a quasi-optimality system state. In this case, the system criterion does not decrease, and condition of stimulation fund sufficiency is fulfilled.

## References

- [1] Roland G. Transition and Economics. Politics, Markets, and Firms. Cambridge: MIT Press, 2000; 840 p.
- [2] Braguinsky S, Yavlinsky G. Incentives and Institutions. Transition to a Market Economy in Russia. Princeton. NJ.: Princeton University Press, 2000; 420 p.
- [3] RF Government Decree of December 30 2015 N 1493 "On State program" Patriotic Education of Citizens of the Russian Federation for 2016 - 2020".
- [4] RF Government Decree of December 27 2012 N 2567-р "On the state program of the Russian Federation" Development of Culture and Tourism "2013 - 2020".
- [5] RF Government Decree of April 15 2014 N 313 (as amended on 10.21.2016.) "On approval of the Russian Federation, the state program" Information Society (2011 - 2020)".
- [6] RF Government Decree of December 27 2012 N 1406 (as amended on 12.25.2015.) "On the federal target program" Development of the Russian judicial system for 2013 - 2020".
- [7] RF Government Decree of April 15 2014 N 320 "On approval of the state program of the Russian Federation" Public Financial Management and regulation of financial markets".
- [8] Burkov VN, Danev B, Enaleev AK, Nanev TB, Podvalny LD, Yusupov BS. Competitive mechanisms in problems of distribution of scarce resources. *Avtomatika i telemekhanika* 1988; 11: 142–53.
- [9] Burkov VN, Enaleev AK, Kalenchuk VF. Coalition with the competitive mechanism of resource distribution. *Avtomatika i telemekhanika* 1989; 12: 81–90.
- [10] Burkov VN, Enaleev AK, Lavrov YG. Synthesis of optimal planning and incentive mechanisms in the active system. *Avtomatika i telemekhanika* 1992; 10: 113–120.
- [11] Burkov VN, Iskakov MB, Korgin NA. Application of generalized median schemes for the construction of non-manipulable mechanism multicriterion active expertise. *Automation and Remote Control* 2010; 71(8): 1681–1694.
- [12] Korgin NA. Equivalence of non-manipulable and non-anonymous priority resource distribution mechanisms. *Upravleniye bol'shimi sistemami* 2009; 26(1): 319–347.
- [13] Burkov VN, Gorgidze II, Novikov DA, Yusupov BS. Models and cost and revenue distribution mechanisms in the market economy. *Moskva: Institut problem upravleniya* 1997; 356 p.
- [14] Korgin NA. Use of intersection property for analysis of feasibility of multicriteria expertise results. *Automation and Remote Control* 2010; 71(6): 1169–1183.
- [15] Enaleev AK. Optimal incentive-compatible mechanisms in active systems. *Automation and Remote Control* 2013; 74(3): 491–505.
- [16] Burkov VN, Korgin NA, Novikov DA. Problems of aggregation and decomposition mechanisms of management of organizational and technical systems. *Problemy upravleniya* 2016; 5: 14–23.
- [17] Krylov IA, Chernousko FL. A method of successive approximations for solution optimal control problems. *USSR Computational Mathematics and Mathematical Physics* 1963; 2(6): 1371–1382.
- [18] Gindes VB. A method of successive approximations for solution linear optimal control. *USSR Computational Mathematics and Mathematical Physics* 1970; 10(1): 297–307.
- [19] Fedorenko RP. An approximate solution of the optimal control problems. *Moskva: Nauka*, 1978; 680 p.
- [20] Chernous'ko FL, Kolmanovsky VB. Computational and approximate methods of optimal control. *Journal of Soviet Mathematics* 1979; 12(3): 310–353.
- [21] Lyubushin AA. Modifications of the method of successive approximations for solving optimal control problems. *USSR Computational Mathematics and Mathematical Physics* 1982; 22(1): 29–34.
- [22] Aleksandrov VM, Dykhta VA. Approximate solution to the resource consumption minimization problem. I. Construction of a quasioptimal control. *Journal of Applied and Industrial Mathematics* 2011; 5(4): 467–477.
- [23] Ivanov DU, Orlova CU, Ajupov AA, Bogatirev VD, Pavlova EV. Venture capital management technique based on real options. *International Business Management* 2016; 10(22): 5286–5290.
- [24] Gerasimov KB, Gerasimov BN. Modeling the development of organization management system. *Asian Social Science* 2015; 11(20): 82–89.
- [25] Novikov D. Theory of Control in Organizations. New York: Nova Science Publishers, 2013; 341 p.
- [26] Geraskin MI, Chkhartishvili AG. Structural Modeling of Oligopoly Market under the Nonlinear Functions of Demand and Agents' Costs. *Automation and Remote Control* 2017; 78(2): 332–348.
- [27] Geraskin MI. The optimal mechanism for the distribution of the effect in an integrated strongly coupled system of anonymous agents with transferable utility. *Problemy upravleniya* 2017; 2: 27–41.
- [28] Geraskin MI. Transferable utility distribution algorithm for multicriteria control in strongly coupled system with priorities. *CEUR Workshop Proceedings* 2016; 1638: 542–551.

# Design patterns of database models as storage systems for experimental information in solving research problems

D.E. Yablokov<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

The article deals with design patterns of relational databases that are used as storage systems of experimental data. The classification of these patterns based on their complexity and level of detail in the description of entities and their relations is given. For each pattern, specific features of its application in the process of data modeling are shown. Databases created on the basis of simple patterns are less adapted to changes. Their design corresponds only to a context of a solvable task. Databases created using more complex patterns have a more flexible design. It allows considering requirements which can arise in the future that minimizes need for redesign.

*Keywords:* data model; design pattern; declarative pattern; advanced declarative pattern; contextual pattern; typed contextual pattern; advanced contextual pattern

---

## 1. Introduction

An important factor in the design of storage systems for experimental information is the problem of the correct choice of the data modeling strategy. It is a choice of the basic concept which would allow to provide the main context of this information and would be enough flexible for possible extensions [1]. We will mean a set of holistic and systematic ideas can be used to express a certain way of understanding or an interpretation of any objects, events, processes or the phenomena, which have any information value as the concept. Following the definition, we can say that data modeling, as actually selecting the data model, is very important stage in development process of the database. In addition, it lays the foundation of a conceptual framework in terms of which we will work with the storage system. The data model is a kind of database pattern [2], i.e., fixed reproducible means of describing the way to represent and store information. For each pattern, it is necessary to consider the selected abstraction level related to a context of subject domain that in the future will be the data source.

## 2. Declarative pattern

With this approach [3], the data are described by the principle “as is” (Fig. 1). Many computing applications for experimental research often use a set of elements interrelated by a certain set of connections. For example, an instance of any abstract data structure “graph” may contain a set of entities with the semantics of the graph “vertex” behavior. The relation of these entities to the mentioned above data structure is described by additional entity “graph\_vertex”. In addition, these entities can be combined into pairs by using the “edge”, associative entities having semantics of the graph edge behavior. Vertices and edges can represent objects of any kind [4]. Usually they have any characteristic allowing identifying them among a set of similar objects in the description. Moreover, they may have some additional attributes relating, for example, to the description of the position of status of a vertex, weight or direction of an edge. Also they may contain data about the properties of the abstraction, which can be considered as vertex or an edge. Along with the properties directly relating to such concepts as a graph and vertex, the pattern supports attributes of the relations, such as attributes of the associative entities “graph\_vertex” and “edge”.

Foreign key attributes for relationships between vertices (“from\_vertex\_id” and “to\_vertex\_id”) indicate their direction in case where data in the form of planar or spatial oriented graph. In the case of undirected graph information on forward edge shall be duplicated only in the backward direction, so that values of “from\_vertex\_id” and “to\_vertex\_id” attributes in the row containing information on edge in the backward direction are interchanged. This will allow us to represent data on the connections between vertices of a simple undirected graph in terms of parallel edges of an oriented multigraph.

This pattern is a highly specialized solution, so it is easily to define correspondence between the object of data domain and abstractions used in modeling process. It has the following advantages. First, easy to understand, because a small number of abstractions allows easily model subject domain, with use of simple concepts, which intuitively is clear. Second, easy to support, because of the possibility to manipulate data without the need for knowledge about more difficult features of storage structure. Third, simple queries to fetch data and fast implementation of a data access layer. However, it has some disadvantages, which hinder development of the corresponding databases. The storage structure is initially fixed and has to be changed before adding any new entity or attribute. It will be necessary to introduce the new table as abstraction for the description of some object of data domain or attribute as an analog of its any property. For this purpose, a refactoring procedure like “Introduce New Table” or “Introduce New Column” [5] has to be performed. In case of weak adaptation strategy of the storage system to incoming requirements affecting the content or the quality of an already existing or new information such actions can cause negative consequences.

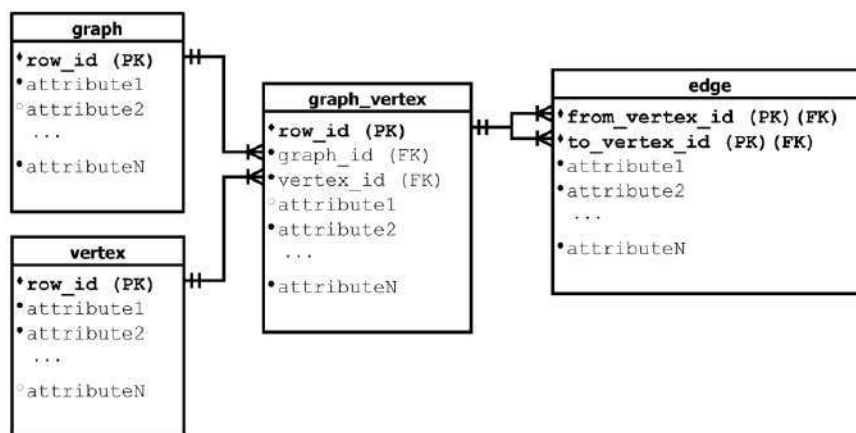


Fig. 1. Advanced declarative pattern.

Using this kind of pattern is possible if change of a data domain context is not expected in the future and development of the storage structure will be done to introduce of sub entities characterizing additional information about the objects already exists in the database.

### 3. Advanced declarative pattern

This pattern has structure similar to the previous pattern, but realizes the additional indirection level in describing attributes and their values. For example, the data model for storage of crystallochemical information using already mentioned graph abstractions could look as follows (fig. 2).

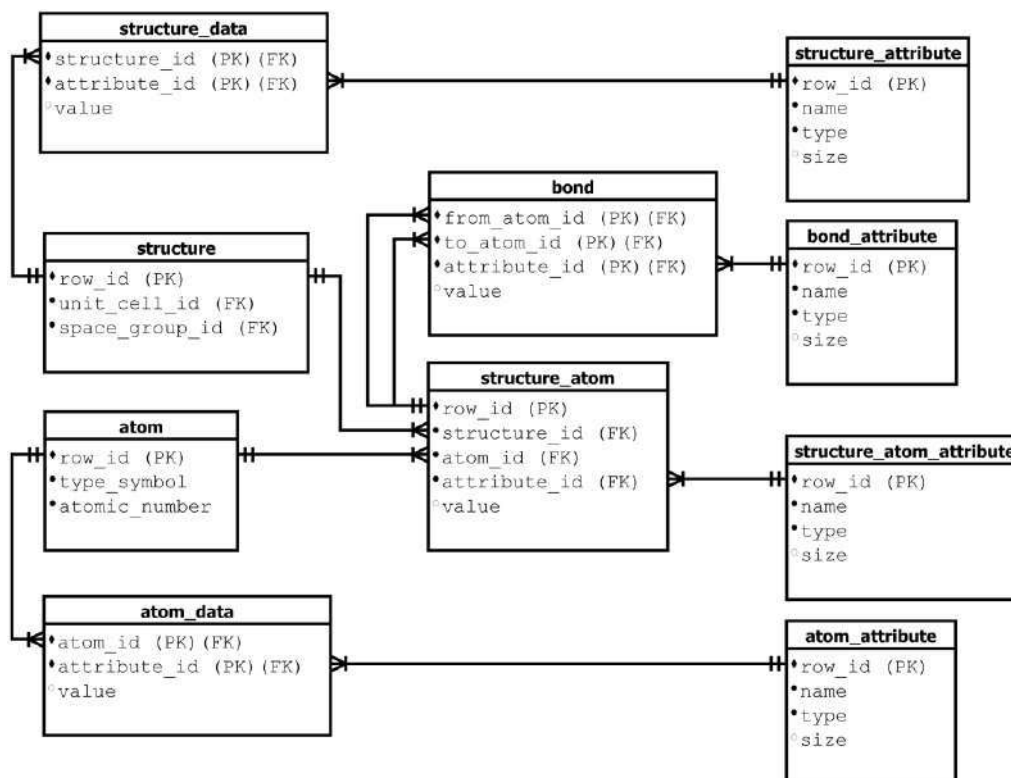


Fig. 2. Advanced declarative pattern.

Foreign key attributes "unit\_cell\_id" and "space\_group\_id" of an entity "structure" specify relationships with the entities "unit\_cell" and "space\_group" describes the concept of unit cell [4] and space group [4]. These data are necessary to unique identification of the chemical structure instance in crystal chemistry. "Type\_symbol" and "atomic\_number" attributes of an entity "atom" describe the main characteristics of chemical elements from the periodic table. The attribute "value" of the associative entities "structure\_data", "atom\_data", "structure\_atom" and "bond" is needed for store the property values of these entities. Properties without values are described by means the entities "structure\_attribute", "atom\_attribute", "structure\_atom\_attribute", "bond\_attribute" containing the same set of fields "name", "type" and "size".

This pattern has all the advantages of the declarative approach but has a more developed mechanism for the attribute description. It is allow describing various states of entity instances of different classes and their relationships without

restrictions. If new data about any characteristics or statuses of chemical structures atoms and their relationships are in the future obtained, this pattern allows saving this information without system redesign.

However, even with the flexibility of the storage structure for attributes and their values the introduction of the new entity is impossible without redesigning storage system entire or its specific part. Descriptions of some attributes can be repeated for different entity classes and this indicates the redundancy of attributes data. Because of the data redundancy there is possible to update the attribute description for only one entity class. The database will contain different descriptions for identical attributes, and it is potential inconsistency when processing or updating data. In data access applications type conversion operations to cast attribute values to the type specified in the attribute description must be implemented. The scenario illustrating the possibility of using this pattern can be the following. If the number of abstractions used in the modeling process to describe domain objects is constant, but the information related to their properties is changed, then it is motivation to use an advanced declarative pattern.

#### 4. Contextual pattern

This pattern [3] implies independence of information context when different objects can be represented differently depending on a situation. Moreover, the change or assignment of their state or behavior can be made even in runtime. Using associative entities for interrelation between objects and attributes (“object attributes”) and also between attributes and relationships (“relationship attributes”) allows assigning to any object or relationship any context-related number of attributes. Based on the conceptual diagram of the contextual pattern (fig. 3) and its simplistic data model (fig. 4) it is possible to conclude that the description of any data domain independently of the context can be represented in terms of objects, their attributes, object attributes, object relationships and relationship attributes.

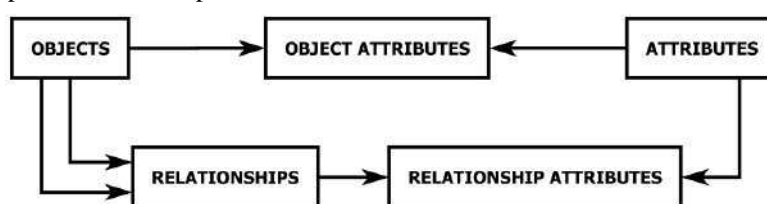


Fig. 3. Contextual pattern (conceptual diagram).

At the description of each object there are “name” and “description” fields for specifying the conceptual context of the instance. For each attribute there is its formal description containing the “name”, “type” and “size” fields. Thus, there is a formal declaration of attributes without specifying of the actual values. The actual values of the attributes associated with the object or objects can be defined using the field “value” of the associative entity “object\_attribute”. In the same way, the actual values of attributes for entities of “relationship” are defined in the field of “value” of the associative entity “relationship\_attribute”. Semantics of such approach is very similar to an advanced declarative pattern, but the contextual pattern does not depend on the context of the subject area.

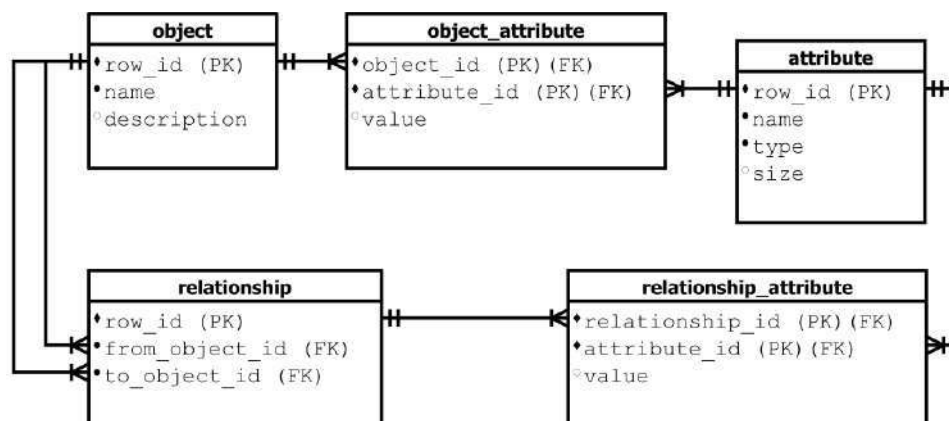


Fig. 4. Contextual pattern (ER-model).

The pattern of this kind is easy to use if you follow a set of implicit rules, but it is necessary to pay for the additional level of flexibility of structure of storage. There is a potential possibility of change the strategy of working with data and there can be difficulties with semantics of storage and data fetching. Not oriented on the subject domain logic, the storage structure will require additional data access level in the form of a set of the user functions or views simplifying access to the information stored in the database. Without the clear specification of the relation of an object to any domain segment, the performance can be reduced when data is fetched because of the large number of objects with an identical set of attributes. The need of information storage about all attributes in one place can lead to data redundancy mentioned in the disadvantages of the advanced declarative pattern.

### 5. Typed contextual pattern

The pattern solves the problem of missing in the object description the characteristic of its relation to a certain class depending on a context of data domain (fig. 5). By entering user-defined types for objects and their relationships, this kind of pattern allows classifying information about them by means predefined criteria. Creation of such criteria in storage system can be step-by-step, for example, when forming necessary level of understanding of features of data domain. These features could be unknown at the initial stage of working with the database. Moving of information on attribute types to the separate table allows avoiding the problem of data redundancy. The structure of the universal storage [6] created by the principles of the typed contextual pattern can be conceptually partitioned into several main components. First, objects that combine such concepts as object type ("object\_type") and an object instance ("object") [7]. Second, attributes ("attribute") and types of attribute values ("data\_type") which allow describing signatures of object properties [7] separately. Third, object interrelations associated with types of object relationships ("relationship\_type") to formalize the entity class "relationship". Fourth, "object\_attribute" and "relationship\_attribute" allow storing attribute values of objects and their relationships.

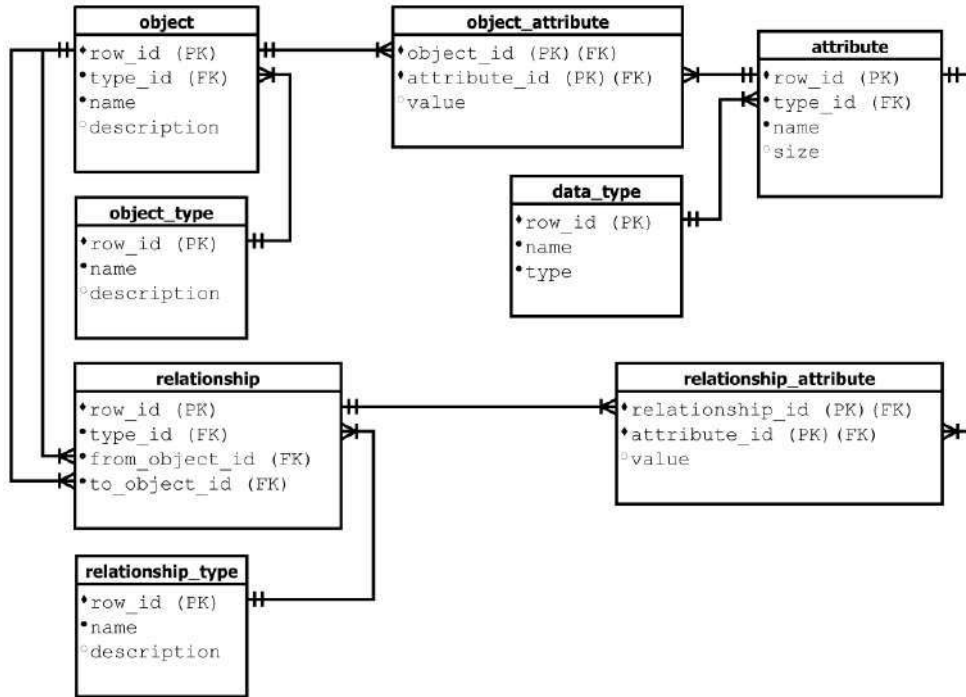


Fig. 5. Typed contextual pattern.

Unfortunately, one of key disadvantages of an advanced declarative pattern complicates use of the typed contextual pattern. Type casting that is necessary for converting attribute values according to their types should still be implemented in the text of queries or views, or in an application at the data access layer.

### 6. Advanced contextual pattern

In the database created by the principle of an advanced contextual pattern (fig. 6) it is important to define common data for most users primitives based on elementary concepts. This will allow identifying logically associated with these concepts unstructured data independently of the subject area and will provide portability of a data model from one project to another. The detailing level in this approach allows developers of new applications to consider at the early stages of design only the most important issues. All others details of the storage system can be implemented later, when the understanding of problem areas of data domain develops to necessary level.

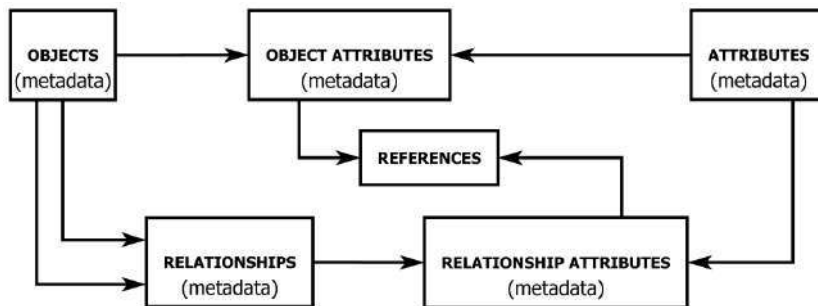


Fig. 6. Advanced contextual pattern (conceptual diagram).

The introduction of additional meta-data level allowed avoiding the disadvantages of the declarative and contextual pattern. The advanced contextual pattern supports all the necessary features to conform to requirements for storage systems based on the universal data model. First, objects description from any subject domain. Second, data representation with using a domain-specific language. Third, data redundancy limitation and support for all possible operations on data processing. Fourth, evolutionary design with adaptation to new requirements and minimal impact on existing data.

Data are described with use of relational approach and bases of object-oriented programming. The main idea is when using a relational kernel of storage system it is extended by the most successful object-oriented technologies. These extensions can be the user-defined type system and the means of describing hierarchical data, such as inheritance and composition which allow to represent relations of objects according to the principles “is a” (similar behavior) or “has a” (part of). Object-oriented approach allows you to represent data in the form of a set of interacting objects, each of them associated with the specific entity class. It promotes the correct and more effective structuring storable information and makes possible to perform an object-oriented decomposition to form the conceptual boundaries of the data model.

As with the typed contextual pattern, the storage structure created by the principles of an advanced contextual pattern can be conceptually partitioned into several main components. The following is a short description of these components is provided, and explanations to use of some categories of the ideas on the basis of which the proposed solutions and methodologies for working with data.

6.1. Objects

Any object is an instance of an entity class and considered as pure abstraction without binding to subject domain (fig. 7). Specification of properties of objects according to the object-oriented approach to the relational storage model is made at the level of object type ("object\_type"). According to storage semantics each object type inherits to any base type ("meta\_type") which in terms of the elementary primitives to define a context as a criteria for possible classification of all child data elements. The special attention needs to be focused on the field "parent\_row\_id" in the description of "object\_type". This field is needed to create tree-like hierarchical structures, which in the storage semantics of the advanced contextual pattern means inheritance.

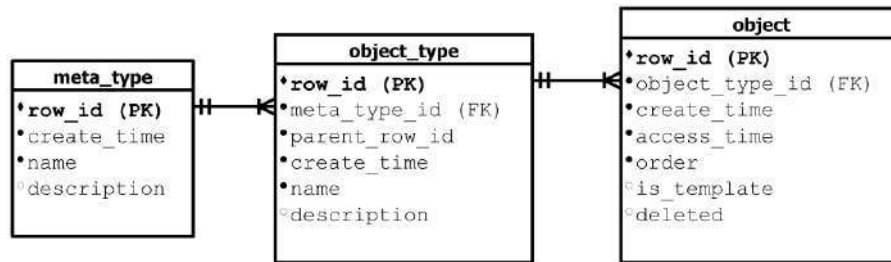


Fig. 7. Advanced contextual pattern (objects).

6.2. Relationships

Each instance of "relationship" is associated with a specific instance of "object\_type". Thus, the association of a relationship with a type of objects is set (fig. 8). For example, the bonds type between atoms (“object\_type”) may be “HB” (Hydrogen Bond) and base type ("meta\_type") can be the "Edge". In the analysis or decomposition of this atomic bond it can be considered in terms of primitive graph abstractions. The description of each relationship contains the link to the “relationship\_type” defining the necessary abstraction level for specifying of criteria that separate a concrete relationship instance from other relationships.

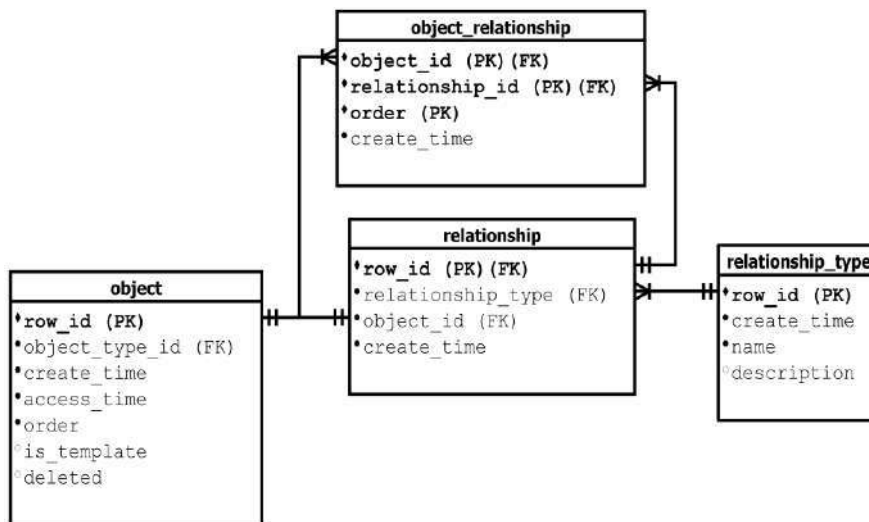


Fig. 8. Advanced contextual pattern (relationships).

For example, the “relationship\_type” can specify the form of the relationship between objects on levels such as acquaintance, aggregation or composition. The associative entity “object\_relationship” allows creating correlation between parent and descendants, for example, specifying a subset of the graph elements using only the edges or vertices.

### 6.3. Attributes

Each “attribute” contains the “structural\_code” field in the description, specifying the semantics of its storage. It can be as primitive attributes, associated with data types, and composite, consisting of primitive or same composite attributes (fig. 9). The associative entity “attribute\_data\_type” contains a reference to the data type (“data\_type\_id” field), and also the “attribute\_exid” field, which is used to construct a table alias for storing the values of primitive attributes.

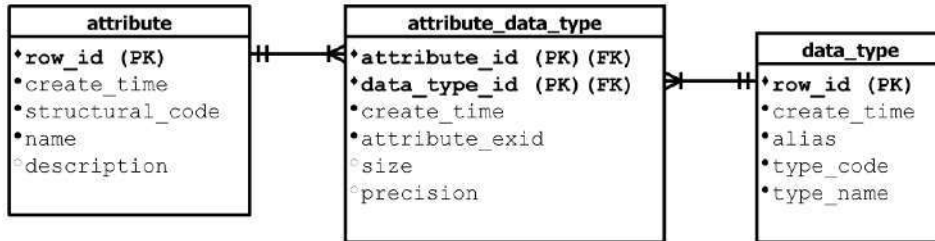


Fig. 9. Advanced contextual pattern (relationships).

### 6.4. Objects and relationships attributes

Each “object\_type” can be assigned any number of attributes (fig. 10). The associative entity “object\_type\_attribute” allows to specify by means of the “parent\_row\_id” field what attribute is composite and what of attributes are child of it. By default the access specifier to properties of ancestor always will be public. It means that in the description of object hierarchies the Liskov Substitution Principle (LSP), one of the basic principles for working in object-oriented style is used. Excluding attributes from the descendant scope is possible when the value of the field “not\_inherited” is set to “true”, the access level is defined to private.

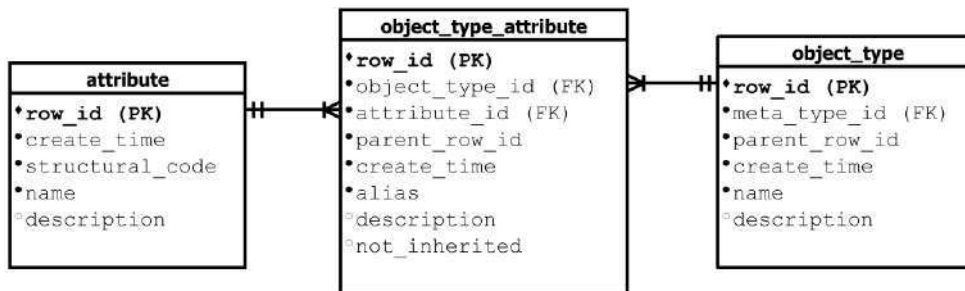


Fig. 10. Advanced contextual pattern (objects and relationships attributes).

### 6.5. Attribute values

The values of primitive attributes are stored in separate tables for example “attribute\_value\_18F19DCCA4F24B27B3D0BAB50AAE740B” (fig. 11). It allows during the query design not worry about converting of attribute values. Aliases of these tables contain the ID got from value of the field “attribute\_exid” declared in the description of “attribute\_data\_type”.

### 6.6. Attribute references

The introduced mechanism of references (“attribute\_reference”) allows defining the relative value for the attribute of another object or relationship (fig. 12). When organizing dictionaries containing any information or a set of constants used in describing experimental data, references to the values of dictionary records can be used as relative values for attributes of other objects or relationships. This will reduce duplication of information and make the data more normalized without any problems with referential integrity.

## 7. Conclusion

The selection of a data model pattern is in fact selection of a paradigm of working with data on the basis of the domain model and the used abstraction level. When using any of patterns we get a lot of technologies corresponding to the complexity level of a pattern allow controlling the complexity of subject domain and complexity of storage system as well as the complexity of software for data processing. Whatever was the selected approach it does not replace the experience and style of thinking necessary to choose the correct strategy for working on data processing and analysis projects.

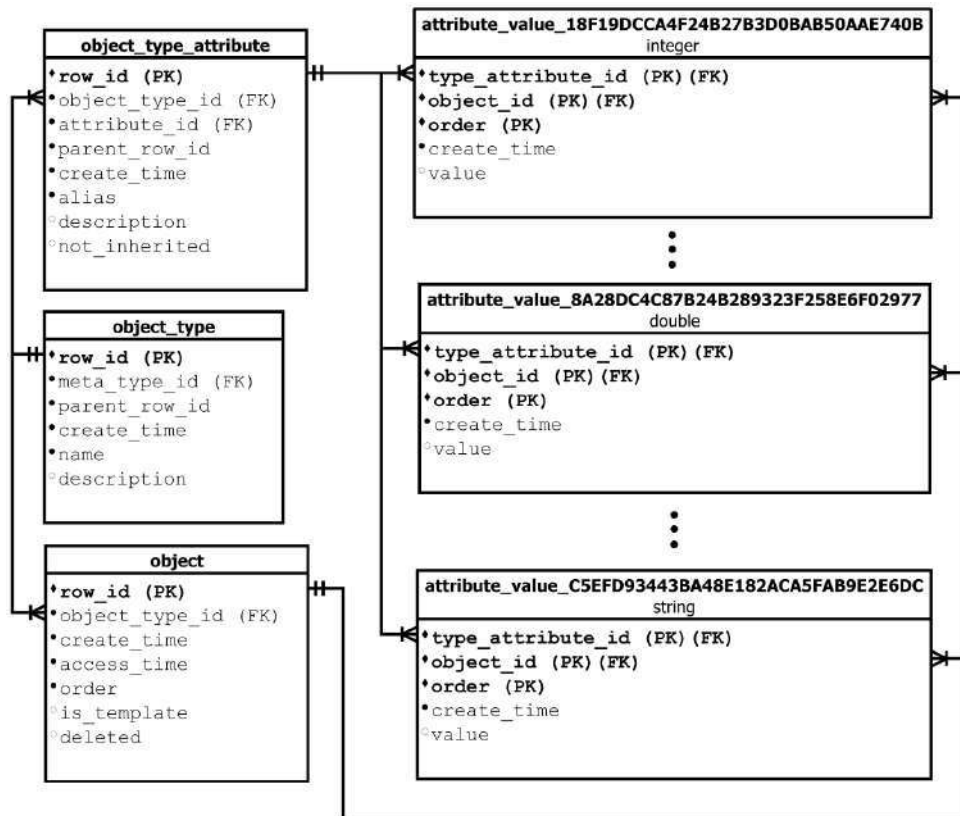


Fig. 11. Advanced contextual pattern (attribute values).

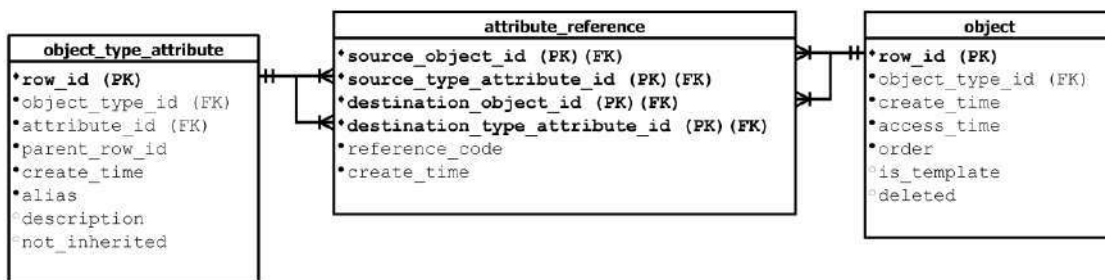


Fig. 12. Advanced contextual pattern (attribute references).

### Acknowledgements

This work was supported by the Russian government (Grant 14.B25.31.0005).

### References

[1] Simson GC, Witt GC. Data Modeling Essentials, Third Edition. Morgan Kaufmann Publishers, 2005; 560 p.  
 [2] Hey DC. Data Model Patterns: Conventions of Thought. Dorset House Publishing, 1996; 288 p.  
 [3] Silverstone L. The data Model Resource Book. Vol. 3: Universal Patterns for Data Modeling. Wiley Computer Publishing, 2009; 648 p.  
 [4] Blatov VA, Proserpio DM. Periodic-Graph Approaches in Crystal Structure Prediction. Edited by Oganov AR. Modern Methods of Cristal Structure Prediction. Wiley-VCH, 2011; 1–28.  
 [5] Ambler S, Sadalage PJ. Refactoring Databases: Evolutionary Database Design. Addison-Wesley, 2006; 384 p.  
 [6] Silverstone L. The Data Model Resource Book. Vol. 1: A Library of Universal Data Models for All Enterprises. Wiley Computer Publishing, 2001; 542 p.  
 [7] Fowler M. Patterns of Enterprise Application Architecture. Addison-Weatley, 2003; 736 p.



# Comparative Analysis of CRM-systems

E.Z. Glazunova<sup>1</sup>, V.V. Kovelskiy<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The comparative analysis of CRM-systems is provided in this article in order to define the most effective software, which would be able to solve problems appearing in companies selling coupons.

*Keywords:* CRM-system; sales; customer relationship management system

## 1. Introduction

Nowadays CRM system is the most effective concept for modern business development. Therefore every company tries to implement such software to increase productivity and find new approaches to customers.

The goal of this research is to find appropriate CRM system for company, which is selling coupons.

Main tasks:

- 1) To define the main problems of companies which are selling coupons
- 2) To make a comparative analysis of the most popular CRM-systems;
- 3) To choose the optimal system for solving the problems.

During the research the following problems were determined:

- 1) Potential clients search;
- 2) Motivation of service providers to provide a large discount;
- 3) Search and processing of analytical information of company\$
- 4) Determination of individual approach to each client;
- 5) Growth of coupons sales;
- 6) Integration with social networks and emails;
- 7) Determination of real value of business proposal;
- 8) Determination of criteria of evaluating work with clients in different sections;
- 9) Determination of a clear procedure for processing of customers applications;
- 10) Creation of flexible reporting system;
- 11) Monitoring of any changes in the database.

## 2. CRM-systems analysis

To achieve the goals mentioned above, coupons selling companies need CRM, which will most effectively help achieve such goals. Therefore, it is to analyze the popular CRM-systems necessary in the framework of this research (table. 1).

Each system was analyzed on 10 parameters on the basis of a scale from 1 to 5 points.

Table 1. Comparative characteristics of CRM-systems on a five-point scale.

Comparative characteristics	InvGate	Bpm'online	Insightly CRM	AmoCRM	Salesforce
Intuitive interface	1	4	2	3	5
Implementation speed	1	4	3	2	5
Project management	3	1	4	2	5
Configuring user restrictions	4	2	3	1	5
Client module	3	4	2	1	5
Conducting transactions	4	2	3	1	5
Invoicing	4	1	2	3	5
Financial resources management accounting	1	3	2	4	5
Speed of operations	3	4	1	2	5
Price	1	4	5	2	3
Average rating	2.5	2.9	2.7	2.1	4.8

Next, the characteristics of each system will be considered.

## 3. CRM-system characteristics

BPM'online allows companies to manage the full consumer life cycle through a single CRM platform, into which three products have been integrated: marketing, sales, service.

Marketing Bpm'online is a multichannel marketing software that enables specialists to carry out sales with subsequent interaction with customers.

Sales Bpm'online - a tool that automates the sales system. The advantage of this function is the ability to combine sales, financial transactions, accounts, communications, etc. into a single system.

Bpm'online service is a tool for servicing and attracting customers.

Bpm'online can work both in the cloud mode and local, where the data is hosted on the company's servers. In both cases, users can access bpm'online from both the web browser and the mobile application.

Additional system features:

- management of any fields in the customer's card;
- the establishment of access modes allows to edit customer cards only to those employees who created counterparties, the others can only scan;

- storage of documents;

- formation of the knowledge base with preservation of all the teaching material or the material necessary for the work;

- generation of standard reports with visualization.

Based on all of the above, we can conclude that the CRM-system BPM-online is functional. But in order to work comfortably in it, you need to download add-ons.

Sales management service AmoCRM is a web-based platform, available from anywhere in the world, where there is an Internet connection. Users of the system can manage sales, personnel and receive analytical information and reports.

Functions of the system:

- 1) Mail integration;
- 2) Field for new users is supported;
- 3) Unique tags allow users to create new agreements and contacts;
- 4) Existing clients may be uploaded from Outlook and Gmail databases;
- 5) Ability to control actual tasks with the help of calls and emails.

AmoCRM, provides full visibility of the sales pipeline: it shows the number of sales, income of sales representatives based on tags.

Platforms like Facebook, MailChimp, Zendesk, Dropbox and Xero can integrate directly with AmoCRM. The mobile app is available for download in the App Store or Google Play.

In addition to the advantages identified shortcomings of this system. First, there is no possibility to differentiate access to information. Secondly, transactions schemes, management and financial reports are not available.

Insightly CRM allows small companies to manage projects, contacts, sales and documents using a single platform that is accessible via the Internet and on mobile devices.

Insightly connects users to online applications, such as Google Calendar, Gmail, MailChimp, Evernote, Dropbox, QuickBooks, Xero, and others.

The Insightly dashboard provides real-time information on current tasks. Advanced Insightly reports allow you to create tables and graphs. The program reports such details as the responsible user, the stage of the task, the deadline, the probability of winning the client.

Insightly also offers an integrated project management function. Once the transaction is concluded, users can track and manage the subsequent project obligations directly from the CRM. The side panel Insightly also saves emails directly from Gmail, provides user access to the conversation history, connects all correspondence with projects and events.

InvGate - is a service management platform that offers help with query registration, customer service and technical support, self-management knowledge and much more. The system is compatible with Mac, Windows and iPad. Companies of almost any industry and any size can use InvGate. This software is scalable and configurable, depending on the needs of the company.

InvGate allows users to run advanced reports and view analytics, automate workflows and much more. Detailing allows you to have the most recent data that is available to all team members.

The main disadvantage of this system is the lack of integration with social networks.

The Salesforce system offers a wide range of CRM applications for all types of businesses, with a focus on sales and support.

The Salesforce system offers vertical solutions for wealth management and financial services segments. Its partners offer a wide range of additional industry solutions. Applications built on the force.com platform are modern architecture, which provides increased flexibility and scalability for organizations of any size.

The Salesforce application has the ability to manage sales, automate marketing, manage relationships with partners, and serve customers. They help organizations manage customer accounts, track sales, conduct and monitor marketing campaigns, and provide maintenance.

The Salesforce system interacts with social networks and performs collaborative work across the organization. For specialized organizations, force.com provides the ability to develop custom applications. Developers can access the application development environment and access the tools and resources needed to design, create, and custom applications for the organization.

The salesforce system for Outlook allows users to synchronize contacts, calendars, messages, and tasks to eliminate double entries through cloud work.

Salesforce allows you to manage an unlimited number of contacts, track transactions, manage tasks and events, and track performance, transactions, and generate reports. It is this system that will achieve the goal of the companies of coupons, which confirms the comparative analysis (Table 1). Therefore, in the industry under consideration, it is preferable for companies to implement Salesforce CRM as an organizational and managerial innovation.

By implementing this particular CRM system, a flexible reporting system will be implemented, that is, all transactions can be sorted by priority (auto-determination of the most important transactions), name, offer, amount and client, and status (for

example, completed and unfinished). In Salesforce there is also a convenient analytical tool "sales funnel", showing the effectiveness of the work of one or another employee or department as a whole.

#### 4. Conclusion

In the work, a quantitative and qualitative analysis of the influence of the CRM system modules from the basic set on the levels of the company's strategic indicators was conducted. As a result, integral indicators of effects and levels of their certainty were obtained.

Thus, the purpose of this study (the selection of CRM-system for the companies of coupons is achieved, and the proposed project of CRM-system implementation meets the planned goals of the strategic development of the industry in question.

#### Acknowledgments

Vice-rector on educational and foreign affair, Professor, Doctor of Economics. Bogatyrev V.D. Associate Professor of Management Department, PhD. Kirillov A.V. Professor of Management Department, PhD in Economics. Osmankin N.N. Rector of Samara University, Professor, Academician, Shakhmatov E.V. President of Samara University, Professor, Academician, Soifer V.A.

#### References

- [1] Kravetch AG, Bershadskiy AM. Systems of management of companies resources. Volgograd State Technical University 2013; 25–45.
- [2] Snyder Mike, Steger Jim. Microsoft Dynamics CRM 3.0. Moscow: ECOM Publishers, 2014; 688 p.
- [3] Nerdinger FB. Orientation to client. Modern practices of working with clients. Moscow: Humanitarian Center, 2014; 180 p.
- [4] Joe H. War to client. Loyalty for once and ever. Moscow, 2010; 112 p.
- [5] Cherkashin P. Strategy of management of interaction with clients (CRM). Moscow: Binom. Laboratory of knowledge, Internet-university of information technologies 2007; 376 p.
- [6] AIN. Choose your own CRM-system: comparative review of Saas-solutions. URL: <http://ain.ua/vyberite-svoyu-crm-sistemu-sravnitelnyj-obzor>.
- [7] Software Advice. InvGate Service Desk Software. URL: <http://www.softwareadvice.com/crm/invgate-profile>.

# The role of subprocess-connector in business process modeling

K. Shoilekova<sup>1</sup>, K. Grigorova<sup>1</sup>, E. Malysheva<sup>2</sup>

<sup>1</sup>Angel Kanchev University of Ruse, 8 Studentska str., Ruse 7017, Bulgaria

<sup>2</sup>Volga Region State University of Services, 4 Gagarina str., 445677, Togliatti, Russia

---

## Abstract

The paper emphasizes the importance of the correct and consistent business process planning. The most important objective in business process generation is the proper execution of the activities in a business organization and studying the links between them. To visualize all business processes in a system a subprocess – connector has been created that helps to detect the deadlock markings in a system. It is impossible to change a component without interfering the operation with the others. Several aspects of Petri Nets as a tool for process simulation and surmounting the deadlocks in a system are presented.

*Keywords:* Business Process; Petri nets; Simulation; Connector; Subprocess

---

## 1. Introduction

Each enterprise is based on models representing its inside processes. A company success is determined by the rational organization of its business processes that are subject to in-depth analyses and continual optimization, which is within the priorities of the business processes management but not a single-time initiative. From an enterprise point of view the management of business processes is becoming increasingly important: business process, or process for short, controls which piece of work will be performed by whom and which resources are exploited for this work, i.e. a business process describes how an enterprise will achieve its business goals.

A business process can be defined as a sequence of activities distinctly specified within an organization involving people, equipment, applications, information and other resources, aiming to create products, respectively, values.

Figure 1 shows a three-dimensional view of a business process including:

- case dimension;
- process dimension;
- resource dimension.

The case dimension signifies that all cases are handled individually. From business process point of view, cases do not interact directly. They influence each other indirectly by sharing resources and data. The process dimension specifies the activities in the workflow process, i.e. the tasks and routing along the tasks. The resource dimension signifies the resources, their roles as well as the organizational units. A business process can be visualized by a number of nodes in a three-dimensional view as shown in Fig. 1. Each node represents either a work item (case + task) or an activity (case + task + resource). It can be seen from Fig. 1 that the business process management is an adhesive of cases, tasks, and organization [1].

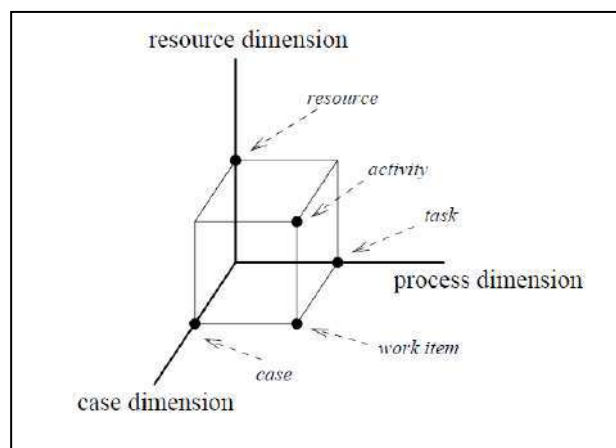


Fig. 1. A three-dimensional view of a business process [1].

## 2. Petri nets

Petri nets were introduced by Carl Adam Petri in 1962]. They became the first standard adopted for business process modeling. Since then Petri nets have been used to model and analyze different processes with applications ranging from embedded systems to flexible manufacturing systems, user interaction, and business processes. Petri nets offer a graphical representation of stepwise processes including choice, iteration and concurrent execution. They are distinguished with an exact

mathematical definition of their execution semantics, and a well-developed mathematical theory for process analysis. Petri nets are recognized as one of the known techniques for describing business processes in a formal and abstract way [5], [6], [8], [9].

Over the last three decades the classical Petri net has been extended with color, time and hierarchy specifications [2], [4]. These extensions facilitate complex process modeling, where data and time are important factors. Petri nets are used as a powerful tool in business process modeling for several reasons. They offer:

- formal semantics;
- graphical language;
- support of basic primitives needed to model business processes;
- analysis based on four general approaches:
  - Reachability Analysis: involves the enumeration of all reachable markings, but it suffers from the state-space explosion issue;
  - Matrix Equation Approach: in many cases it is applicable only to special sub-classes of Petri nets or special situations;
  - Invariant Analysis: determines sets of places or transitions with special features, as token conservation or cyclical behavior;
  - Simulation: discrete-event simulation is an option to check system's properties.

While modeling refers to the development of mathematical representation of modeled object processes, simulation concerns data processing by means of algorithms and procedures for solving all mathematical calculations related to the model, i.e. computer simulation is a system's representation by its model activation.

Business processes' simulation based on Petri nets is a way to show the system characteristics. The main idea is to use an appropriate execution algorithm in order to detect its unwanted properties.

Studies in [3] and [7] helped to compile a list of over 200 different techniques and tools for business processes simulation using Petri nets. Regrettably, a significant number of websites are out of operation being without support and updates for more than 15 years, or they are just source-codes lacking any descriptions, which fences off the understanding of their functions.

In result of the initial filtering, the list was reduced to 47 techniques appropriate for further examination and comparative analysis. Even these are too many for the purpose of detailed investigation, and speaking about an effective software product, at least it is expected to have gained minimum popularity within society and scientific communities. As there is no information about a worked out comparison, a number of Google Scholar results and Google search results were used.

The above stated approach reduced the filtering criteria to:

1. The tool was created or updated over the last 16 years (after year 2000).
2. There is an operational website of the tool that can be used in reality, and it is provided with relevant documentation, help menus or other scientific papers.
3. The tool has minimum 200 citations in scientific papers (ascertained with Google Scholar search).
4. The tool has minimum 5000 results ascertained with Google search.

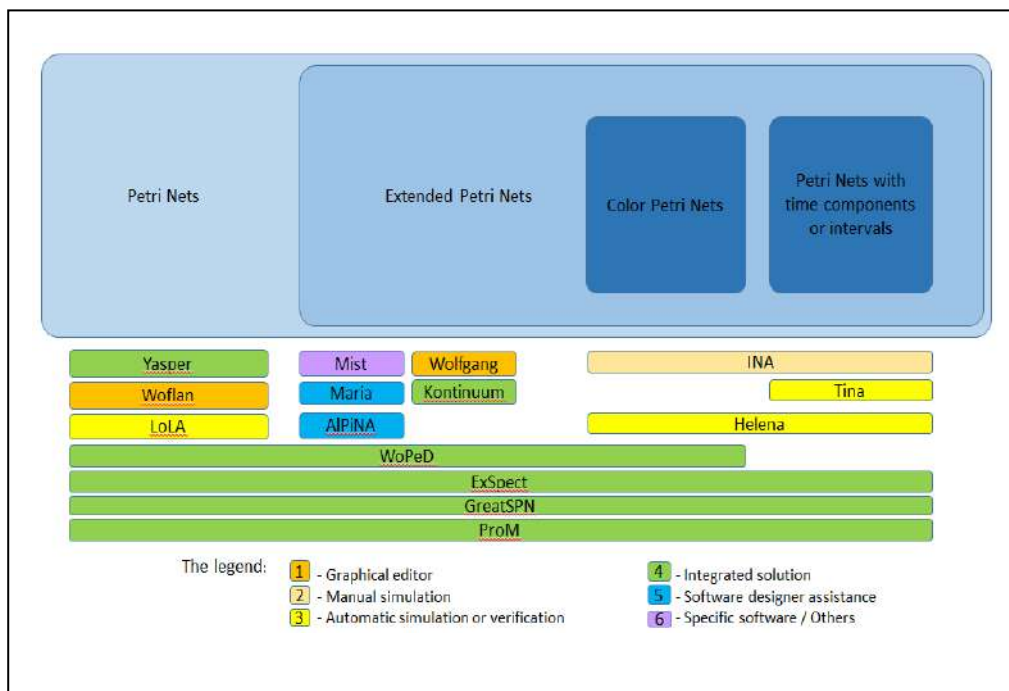


Fig. 2. Tools for business process analyses and transformations.

Concluding filtering, the final list includes 15 tools ranked by the number of Google search results:

1. Kontinuum – this tool is business process management software for use by business decision makers and administrators as a tool to maximize control, save time and cut unnecessary costs.

2. Maria - Modular Reachability Analyzer is a reachability analyzer for concurrent systems that uses Algebraic System Nets (a high-level variant of Petri nets) as its modelling formalism.
3. Wolfgang – it is a lightweight tool that allows users to easily create and edit Petri nets and check them against general and workflow specific net properties.
4. Mist – it is tool to check safety properties against Petri Net like models.
5. INA - Integrated Net Analyzer is a tool package supporting the analysis of Place/Transition Nets (Petri Nets) and Coloured Petri nets.
6. Yasper - Yet Another Smart Process EditoR is a tool for modeling, analyzing and simulating automated business processes. Yasper’s models are based on Petri nets.
7. ProM - Process Mining Framework is an open source framework directed to RapidMiner 5, a system supporting the design and documentation of the whole data mining process.
8. LoLA - a Low Level Petri Net Analyzer has been implemented for the validation of reduction techniques for place/transition net reachability graphs. LoLA features symmetric as well as stubborn set based methods. Net symmetries are automatically computed. Stubborn sets are customized to the particular analysis task.
9. Tina - TIme Petri Net Analyzer is a toolbox for the editing and analysis of Petri Nets, with possibly inhibitor and read arcs, Time Petri Nets, with possibly priorities and stopwatches, and an extension of Time Petri Nets with data handling called Time Transition Systems.
10. Helena - a High LLevel Net Analyzer is a model checker developed at the CNAM university in Paris. It is a free software available under the terms of the GNU general public license.
11. WoPeD - Workflow Petri Net Designer is an open source software product developed at the Cooperative State University Karlsruhe, Germany (distributed under GNU license).
12. GreatSPN - GRaphical Editor and Analyzer for Timed and Stochastic Petri Nets is a tool package for modeling, validation and performance evaluation of distributed systems.
13. ExSpect – it is a software tool designed for discrete process modeling.
14. Woflan - Workflow Analyzer is an analysis tool which can be used to verify the correctness of a workflow procedure.
15. ALPiNA – Algebraic Petri Nets Analyzer is a model checker for Algebraic Petri Nets created by the SMV Group at the University of Geneva.

After multiple filtering and exploring different tools for business processes simulation based on Petri nets, a comparative analysis was worked out (Fig. 2), and they were placed in categories: graphical editor, tools for manual simulation, tools for automatic simulation or verification, integrated solution combining previous solutions, as well as tools for software designer assistance and tools for specific software development.

The analysis confirms that Petri nets can be regarded as an appropriate tool for business processes modeling and simulation.

A main task in business process modeling is the verification of process models regarding syntactical and structural errors. A syntactical error is given if modeling elements are used in an invalid manner. While the former might be checked with low efforts, the latter usually requires a very complicated analysis to prove properties like deadlock in the models. A deadlock in a process model is given if a certain instance of the model (one or more but not necessarily all) can not continue working, while it has not yet reached its end.

Verification is concerned with determining, in advance, whether a process model exhibits certain desirable behaviours. By performing this verification at design time, it is possible to identify all or part of potential problems. In this case the model can be modified before it is used for execution. As part of the systems rely on process models for execution of work, careful analysis of process models at design time can greatly improve the reliability of such systems.

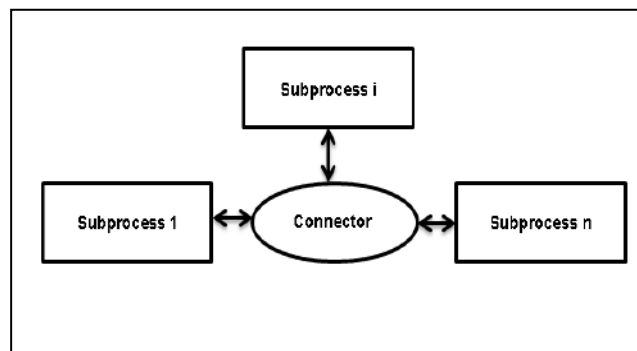


Fig. 3. Subprocess – connector.

The only drawback is that every business process is part of another process, which makes the visualization of all processes incorporating the main business process a difficult task. This problem can be solved by creating a subprocess and connector aiming to describe how business processes interact in order to achieve a complete system functionality (Fig. 3).

To manage the size and complexity of business process models, the use of subprocesses is widely advocated. But there is no solid evidence for benefits of modularization of business process models as well as clear criteria for identifying subprocesses. The modularization may foster the understanding of a complex business process model by its “information hiding” quality and

force the effectiveness of business process modeling. It is even possible to specify some criteria that can be used to automatically derive process fragments that seem suitable to capture as subprocesses.

A subprocess – connector is composed of one basic module called connector and a number of secondary modules called subprocesses. The main business process is described within the connector while subprocesses serve to describe those business processes, which interact with the main process. The subprocess – connector is designed to show the communication between separate processes that is accomplished by determining the input and output characteristics of every other subprocess (Fig. 4).

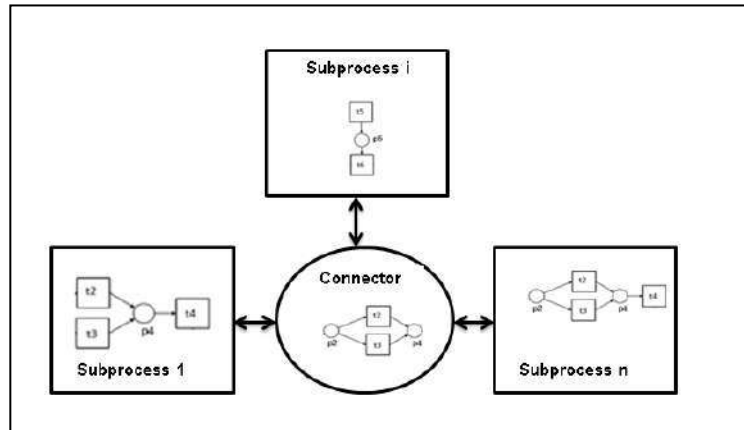


Fig. 4. Communication between separate processes in subprocess – connector.

A main task in subprocess-connector in business process modeling is to produce an abstraction of the process that serves as a basis for detailed definition, study, and possible reengineering to eliminate non-value-added activities. The subprocess-connector must allow for a clear and transparent understanding of the activities being undertaken, the dependencies among the activities, and roles (people, machines, information, etc.) necessary for the process.

Using subprocess-connector, one defines not only the sequence of activities, but also the transmission of data between activities and the conditions that define how the process continues. Since activities might not be executed arbitrarily, they are bound together via connectors.

The execution of subprocess is triggered by start conditions that are determined in the subprocess - connector. The start condition may specify that all incoming control conditions must evaluate to TRUE, or it may specify that at least one of them must evaluate to TRUE. Whatever the start condition is, all incoming conditions must be evaluated before the activity can start. If an activity has no incoming control conditions, it becomes ready when the process or block containing it starts. In addition, a boolean expression called transition condition is associated with each subprocess - connector.

Fig. 5 represents a business process describing the separate stages of fuel oil loading at a filling station. The business process is represented with the help of subprocess – connector. This business process can be described with the following options:

Scenario 1: The driver parks the car, fills oil into the car tank and leaves the filling station.

Scenario 2: In case that the driver is in shortage of money to fill in fuel.

The subprocess - connector is convenient for verifying the target properties of a business process and/or a system as a whole. The complete process provides an opportunity to report a number of deadlock markings, which can prevent serious errors like an incorrectly developed process as part of a series production that can cause dramatic problems to the firm, as well as additional work, legal problems, angry clients, managerial worries and depressed employees. Therefore, it is of crucial importance to check whether a given business process is correct before being started to function.

Deadlock is a state of a system in which no action can take place. Deadlock usually appears in systems that contain subsystems that run in parallel and share some form of common resources. Because Petri nets are a formal model of concurrent systems, they are appropriate for deadlock detection and prevention. Using subprocesses and subprocess – connector allows to simulate the performance of each subprocess and to detect potential deadlock cases in them. This facilitates the deadlock detection in the entire process.

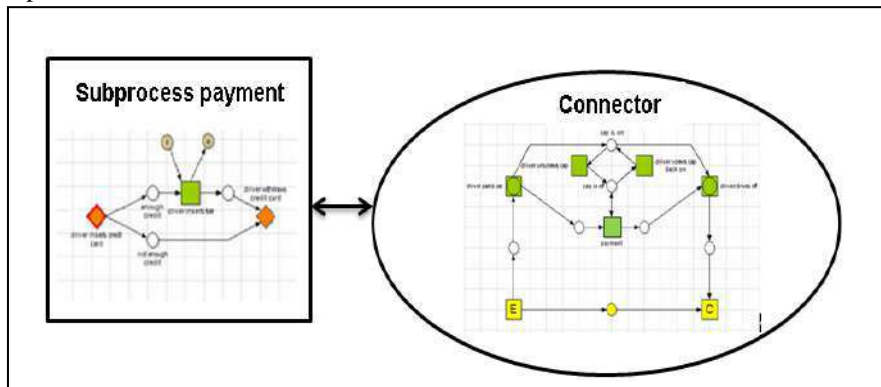


Fig. 5. Communication between separate processes in the subprocess – connector during a concrete business process execution.

Once the processes describing the whole system have been modeled and each process simulated, they can be united by means of the created subprocess – connector in order to achieve:

- visualization of processes within the whole system;
- simulation of business processes within the whole system;
- establishing the interaction between separate business processes.

### 3. Conclusion

The paper discusses the use of subprocess-connector that facilitates business process modeling. An approach to detect deadlocks in process models is described. The subprocess-connector requires a deeper understanding of the business process modeling and Petri nets.

Petri nets are a suitable tool for modeling and simulation of business processes within a system. To visualize all business processes in a system a subprocess – connector has been created that helps to detect:

- the deadlock markings in a system;
- the lack of communication between separate subprocesses and the main business process that may cause doubling of a given subprocess;

It is preferable to model initially the entire process of the enterprise and to separate it into several subprocesses and to model them in details. The defined subprocesses also could be divided into their subprocesses if necessary. The entire process leads to optimizing the business processes of the enterprise.

### Acknowledgements

This work is supported by the Bulgarian National Scientific Research Fund under the contract DFNI - I02/13.

### References

- [1] Van der Aalst WMP. The Application of Petri Nets to Workflow Management. Department of Mathematics and Computing Science, Eindhoven University of Technology.
- [2] Van der Aalst WMP. Putting Petri nets to work in industry. *Computers in Industry* 1994; 25(1): 45–54.
- [3] TGI Petri Nets Tools and Software. URL: <https://www.informatik.uni-hamburg.de> (30.02.2017).
- [4] Jensen K. Coloured Petri Nets. Basic concepts, analysis methods and practical use. EATCS monographs on Theoretical Computer Science. Berlin: Springer-Verlag, 1996.
- [5] Murata T. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE* 1989; 77(4): 541–580.
- [6] Reisig W. Petri Nets. An Introduction. Berlin: Springer-Verlag, 1999.
- [7] Thong WJ, Ameen MA. A Survey of Petri Net Tools. *Advanced Computer and Communication Engineering Technology*. Springer International Publishing, 2015; 537–551.
- [8] Vijverberg W. Translation of Process Modeling Languages. Eindhoven, 2006.
- [9] Wang J. Petri Nets for Dynamic Event-Driven System Modeling. URL: <http://bluehawk.monmouth.edu/~jwang/Ch024.pdf> (03.02.2017).



# Application of Data Mining and Process Mining approaches for improving e-Learning Processes

K. Grigorova<sup>1</sup>, E. Malysheva<sup>2</sup>, S. Bobrovskiy<sup>2</sup>

<sup>1</sup>Angel Kanchev University of Ruse, 8 Studentska str., Ruse 7017, Bulgaria

<sup>2</sup>Volga Region State University of Services, 4 Gagarina str., 445677, Togliatti, Russia

---

## Abstract

The article describes the basic principles and methods of Data mining and Process mining, their similarities and differences. The authors examine the research in Educational Data Mining field, associated with the use of Data mining techniques in education, give examples of problems to be solved with the use of Data mining and Process mining techniques in the area of traditional and e-learning, describe the possibilities and limitations of different methods. Some examples of special software for Data mining and Process mining are presented. A review of major scientific conferences and journals devoted to the research in Educational Data Mining is made.

*Keywords:* Data Mining; Process Mining; Education Data Mining; e-Learning

---

## 1. Introduction

Modern information systems have accumulated a huge amount of data about processes taking place in the various domain areas. Many of today's information systems, including e-Learning system, collect and store data about the events occurring during the systems' performance in so-called event logs. Data mining and Process mining technologies allow the use of the event log data for analysis and improvement of the processes. Availability of advanced software dealing with Data mining and Process mining, allows to test these techniques on data obtained from real processes. A stimulus for the growing interest in Data mining and Process mining is the constant increase in the amount of data recorded in the information systems, including data about events that provide detailed information about the history of the processes, and the need to improve and support business processes in competitive and rapidly changing environment. Data mining and Process mining are complementary approaches that can reinforce each other. Process models detected and aligned with the event log data confirm the value of data analysis and provide a basis for further development as of Process mining, as well as of Data mining.

## 2. Data Mining and Process Mining: An Overview

At the core of both methods (Process mining and Data mining) are the data. They have a lot in common, as they use the same mathematical algorithms and techniques. The main difference is that Data mining operates with the data in general, whilst Process mining works with the data about events, which contain information about the processes [1].

### 2.1. Definitions and Methods of Data mining

Data mining - a multidisciplinary area, which has arisen and developed on the basis of such science fields as applied statistics, artificial intelligence, pattern recognition, machine learning, algorithmization, database theory and others. Data mining might consist of the following steps: identification of patterns and associations (free search), the use of the association rules to predict unknown values (predictive analytics), identification and analysis of the exceptions in the identified rules (anomaly detection). Here are some definitions of the concept. Gartner Group, the agency that analyzes the information technology markets, defines Data mining as follows: "The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques" [2]. SAS Institute, a developer of analytical software, mentions in his definition of big data and its practical usefulness: "Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more" [3]. In the Data mining Curriculum [4] the following definition is met: "Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems".

Data mining methods and algorithms include: decision trees, symbolic rules, cluster analysis, nearest neighbor method, Bayesian networks, artificial neural networks, support vector machines, linear regression, correlation and regression analysis, association rules support, evolutionary programming and genetic algorithms, a variety of methods for data visualization and many others. Most of the analytical methods used in Data mining technology are well-known mathematical algorithms and methods. New in their application is the possibility to use them in solving various concrete problems, due to existing appropriate hardware and software.

## 2.2. The Basic Principles and Methods of Process mining

Process mining is a relatively young research discipline. The idea of Process mining is to detect, control and improve the actual occurring processes by extracting knowledge from event logs readily available in modern information systems [1], [5]. Process mining sits between Big data and Data mining on the one hand, and Business Process Modeling and Analysis on the other. Large volumes of data that business generates, and deployment of business logic across all levels of the business, providing an opportunity for theoretical and practical research on these interrelated and topical areas. Applying the principles of Data science on various aspects of business processes represents a new approach to their modeling and management.

More and more data about business processes is recorded by means of information systems in the form of so-called records of events (event logs), which can advantageously be used as an input information for business process models retrieval. Although the event data are available in the organizations, they often lack of understanding of their real-life processes. A knowledge hidden in event logs can be converted into useful management information.

Process mining includes automated process detection (extraction the process models from event logs), conformance checking (monitoring deviations by comparing model and event logs), defining the organizational structure, automated construction of simulation models, model extension and recovery, the prediction of process behavior in order to develop recommendations on the basis of the process history.

Although this technology has only been recently developed, it can be applied to any type of operational processes in different organizations and systems. Process mining techniques provide new means for detecting, monitoring and improvement of processes in various fields of application, offer opportunities for a stricter conformance checking and the validation and reliability of information about the basic processes of the organization. It is an important tool for modern organizations that need to manage non-trivial operational processes, since on the one hand, there is an incredible growth of event data, on the other hand, the processes should be aligned with the need for effective customer service.

One of the main directions of modern Data mining application is Educational Data Mining (EDM). The main goal of EDM is to use the huge amount of data about the educational processes, coming from different sources in different formats and with different levels of detail. The data represents information about the educational process, provides better understanding of learning and improving its outcomes.

## 3. Data and problems in EDM

Nowadays in the field of education there are a wide variety of educational environments and information systems. CBE (Computer-based education) refers to the use of computers in education to provide directed training to generate control instructions for the student. The first CBE systems are a stand-alone educational applications that work on your computer without the use of artificial intelligence for student modeling, adaptation, personalization, and so on. Global use of the Internet has led to development of many new Web based educational system, such as e-learning systems, distance learning systems, on-line training systems, and so on, and the increasing use of artificial intelligence has led to the emergence of new intelligent and adaptive educational systems. The main types of currently used systems include: LMS (Learning management systems) [7], ITS (Intelligent tutoring systems) [8], AIHS (Adaptive intelligent hypermedia systems) [9], Test and quiz systems [10] and others. Each of them provides a variety of data sources that need to be processed in different ways depending on the nature of the available data and the specific problems and tasks that are solved by using Data mining techniques.

During Educational Data Mining researchers use data of educational systems such as distance learning systems, intelligent computer-based training, electronic manuals, school information systems, online classes and discussion forums, computer-aided testing system [11]. The data have typical characteristics, such as multiple levels of hierarchy (a level for subject, a level for grading, a level for question), the context (a specific student in a particular class answers to a specific question in a particular time on a particular date), short time data (recording data with different resolutions to facilitate various analyses, for example, to record data every 20 seconds) and long periods of time data (a big amount of data recorded over many sessions over an extended period of time, for example, covering semester and yearly courses) [12]. EDM analyzes the data by any type of information system, supporting training or education (universities, schools, colleges and other academic or professional education institutions, providing traditional and modern forms and methods of training, and informal learning). These data are not limited to the interaction of individual students with the educational system (for example, data entry in the tests, navigating through the training and testing system, interactive exercises), but may also include data about the cooperation of students (e.g. text chat), administrative data (e.g. school, district, teacher), demographics (e.g. gender, age, school classes), student emotionality (e.g. motivation, emotional state) and so on.

Since the main purpose of Data mining in the field of education is to greatly improve the quality of training, it is more difficult to get quantitative measurements than in other areas, and the results should be evaluated through indicators like improving efficiency. Thus, a data-driven decisions are formed aiming to improve the current educational processes and teaching materials. EDM is often used when working with educational programs, in solving problems of modeling student's behavior and forecasting of the course results. Examples of problems solved with the help of EDM, are:

- Monitoring the progress of learning to detect in real time the undesirable behavior of students, such as the termination of training, low motivation, incorrect use of educational forums, abuse, fraud, etc., creating warnings to the parties concerned [13], provision feedback to the teachers in order to support decision-making on the improvement of student learning, the adoption of pre-emptive actions to remedy the situation [10];

- Predicting student achievement, assessment of knowledge and learning outcomes [10], formation of recommendations to students based on their interests and activities in the learning process [14];
- Individual approach, adapting training to each student, including course content, navigation on the course, the presentation of the material [15], [16], identification the groups of students according to their individual characteristics, personal characteristics, features of the training, etc. [17] [18];
- Building a curriculum and educational content [19], [20], planning and scheduling of future courses, course planning, planning of resource allocation, organization of access to learning materials, planning consultations, curriculum development, etc. [21];
- Development and validation of scientific theories on learning technology, the formation of new scientific hypotheses [22], simulation the domain teaching instructions in terms of concepts, skills, training modules and their relationships [23]; User / Student modeling (Cognitive models of students presenting their skills and knowledge) [24], estimation of parameters of probability models based on data about learning to determine the likelihood of events of interest [25].

A variety of problems and their educational performance leads to the need to adapt methods of Data mining and Process mining to these data and problems. The applicability of Data mining techniques in the field of education are considered in [12], [26].

#### 4. Data Mining and Process Mining methods in EDM and e-Learning systems

In Educational Data Mining, the most commonly used methods are Classification, Clustering, Text mining (text data mining and text analytics) and Relationship mining, Knowledge tracing, Bayesian modeling, Social network analysis, as well as the Detection of anomalies, Discovery with models, Distillation of data for human judgment, Nonnegative Matrix factorization and techniques and algorithms of Process mining, such as Alpha-algorithms, Heuristic algorithms, Probabilistic algorithms, Genetic algorithms, etc.

Prediction – a definition of how the target attribute depends on a combination of other attributes. The types of prediction methods are: classification (target variable is a category), regression (target and background variables are numbers), the density score (predicted value is the probability density function). Using these methods to predict student performance and to determine the pattern of student behavior is considered in [27] and [28].

Clustering is identification of groups of similar instances. Typically, to determine the similarity the distance measure is used. After the set of clusters is determined, new items can be classified according to the nearest cluster. The clustering in EDM can be used to group similar course materials or to form groups of students based on their knowledge and patterns of interactions [29], [30]. Examples of the applicability of various types of clustering algorithms in EDM are discussed in [31].

Text Mining is a method of producing high-quality information from text. Typical tasks include text mining categorization of text, text clustering, concept / entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. In the EDM, text mining was used to analyze the content of discussion boards, forums, chats, Web pages, documents, and so on. [32].

Relationship Mining allows us to determine the relationships between the variables and presenting them in the form of rules for subsequent use. There are different types of relationship mining, such as association rule mining (relations between variables), sequential pattern mining (temporal association between variables), correlation mining (linear correlation between variables) and causal data mining (the causal relationships between variables). Relationship mining can be used to determine the relationships in student behaviors (behavior patterns) and to diagnose difficulties in teaching or the mistakes that often occur together. [33]

Knowledge Tracing (KT) is a popular method to assess student skills, which is used in effective cognitive tutor systems [34]. KT uses a cognitive model that maps problem-solving item required skills and records correct and incorrect responses of students as evidence of their knowledge of a particular skill. It monitors students' knowledge for some time, and parameterizes them by four variables. KT corresponds to the method of Bayesian network.

Social Network Analysis (SNA) is to understand and to measure the relationship between the entities in the network information. SNA considers social relationships in terms of network theory consisting of nodes (representing individual actors within the network) and the connections or ties (which represent relationships between individuals, such as friendship, kinship, organizational position, etc.). In the EDM Social Network Analysis can be used to obtain information to interpret and analyze the structure and relationships in the interaction tasks, including interaction with the communications [35].

Outlier Detection - is to identify the data that are significantly different of rest of the data. Abnormal values correspond to the observations (or measurements), which are usually more or less than other values. The EDM anomaly detection can be used for the detection of students with learning difficulties, deviations in the actions or behavior of a student or a teacher, and for the detection of irregular learning processes [36].

Discovery with Models is to use previously tested phenomena model (using a prediction, clustering, or manual knowledge engineering) as a component of another kind of analysis such as prediction or relationship mining [37]. This method is often used in EDM and supports the identification of the relationship between the student's behavior and its characteristics, the use of psychometric modeling systems in machine-learning models, the analysis of research in various fields of study [38].

Distillation of Data for Human Judgment is to present the data in an understandable form using generalization, visualization and interactive interfaces to extract useful information and to support decision making. This method comprises obtaining statistical data about the learning process to determine the common characteristics, obtaining summary data and reports on the

behavior of the trainee. Data visualization and graphical techniques help to see, explore and understand large amounts of educational data immediately. In the EDM is also known as distillation for human judgment [39] and it has been used to assist teachers with the visualization and analysis of the students activity and the use of the information [40].

Nonnegative Matrix Factorization (NMF) is a technique that involves a clear interpretation in terms of Q-matrix, also referred to as transfer model [41]. There are many NMF algorithm, and they can give different solutions. NMF uses an array of positive numbers is the product of two smaller matrices. For example, when a learning process is considered, the matrix may represent the results of students' testing and can be decomposed into two matrices: Q, which represents learning elements and S, representing each student's skills.

The extraction of knowledge about the process in the learning systems from event logs for the full representation of the entire process, its analysis and improvement is the purpose of Process mining.

In the EDM Process mining can be used to present the students' behavior according to the records in the event log. Data about each event contain the time stamp and the data about learning process. This may be information about students' knowledge assessment [42], information on participation in forums and chats, about lectures and other educational materials viewing, information about passing tests [43], data describing the collaborative learning processes [44], information about events related to the metacognitive prompts [45]. Depending on the behavior of students, they can be combined into different groups.

It is important to define the concept of the event (it could be a mouse click) and the concept of the sequence of events. For visualization of individual events Dotted Chart diagrams are often used. Further a construction of process models and conformance checking take place. To construct and test learning process models the general and special Process mining algorithms are used (alpha-algorithms, probabilistic, heuristic and genetic algorithms) as well as the Data mining methods and algorithms. The process model is usually presented in the form of a BPMN model or as a Petri net. Building of the learning process model is complicated by the existence of loops and parallel tasks, the presence of "noise", the mutual influence of some tasks to others.

Unfortunately, in the Russian scientific journals, in spite of the considerable amount of work in the field of data mining, there are still little scientific papers related to the study of the application of Data mining and Process mining technology in the learning process. Among them there are the use of artificial neural networks in the modeling of educational process in high school [46], the study of the structure of high school students values by means of cluster analysis [47], the use of methods of Educational Data Mining and Learning Analytics in the educational qualifications [48], the study of the factors of adaptation of students to training conditions with the help of the analysis of variance method [49], an overview of the tasks and methods of Data mining in the field of education and the use of classification algorithms for data analysis of training systems [50].

## 5. Software products with the capabilities of Data mining and Process mining

Special software is necessary for the implementation of Data mining and Process mining. More and more software vendors add to their software products such features. Examples of software products with the capabilities of Data mining and Process mining are presented in Table 1.

Table 1. Examples of software products with the capabilities of Data mining and Process mining.

Tool Name	Vendor	Website
Celonis Process Mining	Celonis GmbH	<a href="http://www.celonis.de">www.celonis.de</a>
Disco	Fluxicon	<a href="http://www.fluxicon.com">www.fluxicon.com</a>
Minit	Gradient ECM	<a href="http://www.minitlabs.com">www.minitlabs.com</a>
NLTK	Open Source	<a href="http://www.nltk.org">www.nltk.org</a>
Orange	Open Source	<a href="http://orange.biolab.si">orange.biolab.si</a>
Perceptive Process Mining	Lexmark	<a href="http://www.lexmark.com">www.lexmark.com</a>
ProM	Open Source	<a href="http://www.promtools.org">www.promtools.org</a>
ProM Lite	Open Source	<a href="http://www.promtools.org">www.promtools.org</a>
QPR ProcessAnalyzer	QPR	<a href="http://www.qpr.com">www.qpr.com</a>
RapidProM	Open Source	<a href="http://www.rapidprom.org">www.rapidprom.org</a>
RapidMiner	Open Source	<a href="http://www.rapidminer.com">www.rapidminer.com</a>
Rialto Process	Exeura	<a href="http://www.exeura.eu">www.exeura.eu</a>
SNP Business Process Analysis	SNP AG	<a href="http://www.snp-bpa.com">www.snp-bpa.com</a>
SPSS	IBM	<a href="http://www-01.ibm.com/software">www-01.ibm.com/software</a>
WEKA	Open Source	<a href="http://www.cs.waikato.ac.nz/ml/weka/">www.cs.waikato.ac.nz/ml/weka/</a>

One of the commonly used software is freeware ProM. ProM has over 1,500 plug-ins, allowing the use of different methods and algorithms for Data mining and Process mining, different types of data and models, to convert the data and models, etc., and

the version ProM Lite contains the most commonly used modules. Most commercial software products, including Data mining and Process mining, are easy to use. Approximately 40 software products, often used in Data mining in the field of education are given in [6].

## 6. Scientific conferences and journals in the field of Educational Data Mining

EDM became an independent research area in recent years. It includes research on the training of intellectual systems - Intelligent tutoring systems (ITS), Artificial intelligence in education (AIED), User modeling (UM), Technology-enhanced learning (TEL), as well as Adaptive and intelligent educational hypermedia (AIEH).

The first conference EDM2008 is held in Montreal, Canada; EDM2009 in Cordoba, Spain; EDM2010 in Pittsburgh, USA; EDM2011 in Eindhoven, the Netherlands; EDM2012 in Chania, Greece, EDM2013 in Memphis, USA, EDM2014 in London, UK, EDM2015 in Madrid, Spain, and EDM2016 in Raleigh, USA, EDM2017 in Wuhan, China.

Table 2 summarizes some of the conferences that correspond to the field of EDM.

Table 2. Scientific conferences that correspond to the category EDM.

Title	Short title	Type	Starting year
International Conference on Artificial Intelligence in Education	AIED	every two years	1983
International Conference on Educational Data Mining	EDM	annual	2008
International Conference on Intelligent Tutoring Systems	ITS	every two years	1988
International Conference on Learning Analytics and Knowledge	LAK	annual	2011
International Conference on User Modeling, Adaptation, and Personalization	UMAP	annual	2009

Table 3 provides examples of journals corresponding to the field of EDM.

Table 3. Examples of journals corresponding to the field of EDM.

Title	Short title	Publisher
ACM Special Interest Group on Knowledge Discovery and Data Mining, Explorations	SIGKDD Explorations	ACM
Computer and Education	CAE	Elsevier
IEEE Transactions on Knowledge and Data Engineering	TKDE	IEEE
IEEE Transactions on Learning Technologies	TLT	IEEE
Internet and Higher Education	INTHIG	Elsevier
International Journal of Artificial Intelligence in Education	IJAIED	AIED Society
Journal of Educational and Behavioral Statistics	JEBS	SAGE Publications
Journal of Educational Data Mining	JEDM	EDM Society
Journal of the Learning Sciences	J Learn Sci	Taylor&Francis
User Modeling and User-Adapted Interaction	UMUAI	Springer

Most accurately the theme of the domain is presented in Journal of Educational Data Mining (<http://www.educationaldatamining.org/JEDM/>), published since 2009. Journal of Educational Data Mining is available as an online journal with free access.

## 7. Conclusion

The paper discusses the basic principles of research in EDM domain, some examples of tasks that can be solved by the use of data mining and Process mining in the area of traditional and e-learning are given, the possibilities and limitations of different methods are described, an overview of the major scientific conferences and journals devoted to the application of Data mining and Process mining techniques in education is presented.

EDM allows investigation on the content of learning materials in e-learning systems and the processes performed in it to be carried out.

The use of Information and Communication Technologies in education generates a large amount of data that contains comprehensive information for students, the processes through which they pass in the course of education. The data derived and used by stakeholders (teachers, instructors, etc.) to understand the learning habits of students, the factors affecting their performance and skills they acquire can be examined. To answer these questions, the research interest in the use of Data mining in education increases. EDM is a discipline aimed at developing specific methods to study educational databases generated by any type of information system supporting training or education (schools, colleges, universities, or vocational training institutions offering traditional and/ or modern methods teaching and informal learning). EDM brings together researchers and practitioners from computer science, education, psychology, psychometrics, and statistics.

The basic idea of Process mining is detecting, monitoring and improvement of real processes by extracting knowledge from event logs automatically recorded by information systems. This approach can be applied to the problems of education. The main goals in this direction are:

- The extraction of process-related knowledge from large education event logs, such as: process models following key performance indicators or a set of curriculum pattern templates.
- The analysis of educational processes and their conformance with established curriculum constraints, educators' hypothesis and prerequisites.
- The enhancement of educational process models with performance indicators: execution time, bottlenecks, decision point, etc.
- The personalization of educational processes via the recommendation of the best course units or learning paths to students (depending on their profiles, their preferences or their target skills) and the on-line detection of prerequisites' violations.

It can be concluded that the use of complementary methods of Data mining and Process mining in e-Learning systems can improve the quality of teaching, increase its availability and effectiveness.

## Acknowledgements

This work is supported by the Bulgarian National Scientific Research Fund under the contract DFNI - I02/13.

## References

- [1] Van der Aalst WMP. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin: Springer-Verlag, 2011; 370 p.
- [2] Gartner Inc. IT Glossary. URL: <http://www.gartner.com/it-glossary/data-mining> (21.01.2017).
- [3] SAS Institute Inc. URL: [http://www.sas.com/en\\_us/insights/analytics/data-mining.html](http://www.sas.com/en_us/insights/analytics/data-mining.html) (21.01.2017).
- [4] SIGKDD. URL: <http://www.kdd.org/curriculum/index.html> (21.01.2017).
- [5] IEEE Task Force on Process Mining. *Process Mining Manifesto*. URL: [http://www.processmining.org/blogs/pub2012/process\\_mining\\_manifesto](http://www.processmining.org/blogs/pub2012/process_mining_manifesto) (21.01.2017).
- [6] Slater S, Joksimović S, Kovanovic V, Baker RSJd, Gasevic D. Tools for Educational Data Mining: A Review. *Journal of Educational and Behavioral Statistics* 2017; 42(1): 85–106.
- [7] Romero CE, Ventura S, Salcines E. Data mining in course management systems: Moodle case study and tutorial. *Comput Edu* 2008; 51: 368–384.
- [8] Mostow J, Beck J. Some useful tactics to modify, map and mine data from intelligent tutors. *J Nat Lang Eng*. 2006; 12: 195–208.
- [9] Merceron A, Yacef K. Mining student data captured from a web-based tutoring tool: initial exploration and results. *J Interact Learn Res* 2004; 15: 319–346.
- [10] Romero C, Zafra A, Luna JM, Ventura S. Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Syst J Knowl Eng*. 2013; 30(2): 162–172.
- [11] Romero C, Ventura S, Pechenizkiy M, Baker RSJd. *Handbook of Educational Data Mining*. Chapman & Hall/CRC Press, 2011; 526 p.
- [12] Romero C, Ventura S. Data mining in education. *The Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2013; 3: 12–27.
- [13] Kotsiantis S, Patriarchas K, Xenos MN. A combinational incremental ensemble of classifiers as a technique for predicting student's performance in distance education. *Knowl-Based Syst*. 2010; 23: 529–535.
- [14] Tang T, Daniel BK, Romero C. Recommender systems for and in social and online learning environments. *Expert Syst J Knowl Eng*. 2015; 32(2): 261–263.
- [15] Romero C, Ventura S. Preface to the special issue on data mining for personalised educational systems. *User Model User-Adapted Interact*. 2011; 21: 1–3.
- [16] Bannert M, Reimann P, Sonnenberg C. Process mining techniques for analyzing patterns and strategies in students' self-regulated learning. *Metacognition and Learning* 2014; 9(2): 161–185.
- [17] Bouchet F, Harley JM, Trevors GJ, Azevedo R. Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. *Journal of Educational Data Mining* 2013; 5(1): 104–146.
- [18] Ayers E, Nugent R, Dean N. A comparison of student skill knowledge estimates. *International Conference On Educational Data Mining*. Cordoba, Spain, 2009; 1–10.
- [19] Cairns AH, Gueni B, Fhima M, Cairns A, David S, Khelifa N. Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining. *The Fourth International Conference on Advances in Information Mining and Management*, 2014; 53–58.
- [20] Garcia E, Romero C, Ventura S, Castro C. Collaborative data mining tool for education. *International Conference on Educational Data Mining*. Cordoba, Spain, 2009; 299–306.
- [21] Hsia T, Shie A, Chen L. Course planning of extension education to meet market demand by using datamining techniques—an example of Chinkuo Technology University in Taiwan. *Expert Syst Appl J*. 2008; 34: 596–602.
- [22] Siemens G, Baker RSJd. Learning analytics and educational data mining: towards communication and collaboration. *Proceedings of the 2<sup>nd</sup> International Conference on Learning Analytics and Knowledge*. Vancouver, British Columbia, Canada, 2012; 1–3.
- [23] Pavlik P, Cen H, Koedinger K. Learning factors transfer analysis: using learning curve analysis to automatically generate domain models. *Int Conf Edu Data Min*. 2009; 121–130.
- [24] Frias-Martinez E, Chen S, Liu X. Survey of datamining approaches to user modeling for adaptive hypermedia. *IEEE Trans Syst Man Cybern C*. 2006; 36(6): 734–749.
- [25] Wauters K, Desmet P, Noortgate W. Acquiring item difficulty estimates: a collaborative effort of data and judgment. *International Conference on Educational Data Mining*. Eindhoven, The Netherlands, 2011; 121–128.
- [26] Baker R, Siemens G. *Educational data mining and learning analytics*. Cambridge Handbook of the Learning Sciences: 2nd Edition, 2014: 253–274.

- [27] Romero C, Espejo P, Zafra A, Romero J, Ventura S. Web usage mining for predicting marks of students that use Moodle courses. *Comput Appl Eng Edu J*. 2013; 21: 135–146.
- [28] Baker RSJd, Gowda SM, Corbett AT. Automatically detecting a student's preparation for future learning: help use is key. *Fourth International Conference on Educational Data Mining*. Eindhoven, The Netherlands, 2011; 179–188.
- [29] Bogarín A, Romero C, Cerezo R, Sánchez-Santillán M. Clustering for improving educational process mining. *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. ACM - New York, NY, USA, 2014; 11–15.
- [30] Vellido A, Castro F, Nebot A. Clustering Educational Data. *Handbook of Educational Data Mining*. Boca Raton, FL: Chapman and Hall/CRC Press, 2011; 75–92.
- [31] Dutt A, Aghabozrgi S, Ismail MAB, Mahroeian H. Clustering Algorithms Applied in Educational Data Mining. *International Journal of Information and Electronics Engineering* 2015; 5(2): 112–116.
- [32] Tane J, Schmitz C, Stumme G. Semantic resource management for the web: an e-learning application. *International Conference of the WWW*. New York, 2004; 1–10.
- [33] Merceron A, Yacef K. Measuring correlation of strong symmetric association rules in educational data. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, 2011; 245–256.
- [34] Corbett A, Anderson J. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model User-Adapted Interact* 1995; 4: 253–278.
- [35] Rabbany R, Takaffoli M, Zaïane O. Analyzing participation of students in online courses using social network analysis techniques. *International Conference on Educational Data Mining*. Eindhoven, The Netherlands, 2011; 21–30.
- [36] Ueno M. Online outlier detection system for learning time data in e-learning and its evaluation. *International Conference on Computers and Advanced Technology in Education*. Beijing, China, 2004; 248–253.
- [37] Baker RSJd, Yacef K. The state of educational data mining in 2009: a review and future visions. *J Edu Data Min*. 2009; 3–17.
- [38] Bienkowski M, Feng M, Means B. Enhancing teaching and learning through educational data mining and learning analytics: an issue brief. Washington, D.C.: Office of Educational Technology. U.S. Department of Education, 2012; 1–57.
- [39] Baker RSJd. Data mining for education. *International Encyclopedia of Education*. 3rd ed. Oxford, UK: Elsevier, 2010; 7: 112–118.
- [40] Mazza R, Milani C. GISMO: a graphical interactive student monitoring tool for course management systems. *International Conference on Technology Enhanced Learning*. Milan, Italy, 2004; 1–8.
- [41] Desmarais MC. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explor*. 2011; 13: 30–36.
- [42] Trčka N, Pechenizkiy M, van der Aalst W. Process mining from educational data. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, 2011; 123–142.
- [43] Mukala P, Buijs J, Leemans M, van der Aalst W. Learning Analytics on Coursera Event Data: A Process Mining Approach. *5th International Symposium on Data-driven Process Discovery and Analysis*. Vienna, Austria, 2015; 18–32.
- [44] Schoor C, Bannert M. Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior* 2012; 28: 1321–1331.
- [45] Sonnenberg C, Bannert M. Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics* 2015; 2(1): 72–100.
- [46] Petrova MV, Anufrieva DA. Investigation of the possibilities of methods of intellectual data analysis in modeling the educational process in the university. *Vestnik Chuvashskogo Universiteta* 2013; 3: 280–285. (in Russian)
- [47] Avadehni YuI, Kulikova OM, Radionova VA. The study of the structure of values of university students with the use of data mining technologies. *Sovremennye problemy nauki i obrazovaniya* 2013; 6: 841 p. (in Russian)
- [48] Veryaev AA, Tatarnikova GV. Educational Data Mining i Learning Analytics - directions of development of educational qualification. *Prepodavatel' XXI vek* 2016; 2: 150–160. (in Russian)
- [49] Shumetov VG, Lyaskovskaya OV. Study of the factors of adaptation of the students of the 2000s to the training in the university by the methods of data Mining. *Srednerusskij vestnik obshchestvennyh nauk* 2015; 6: 49–56. (in Russian)
- [50] Gorlushkina NN, Kocyuba IY, Hlopotov MV. The tasks and methods of intellectual analysis of educational data to support decision-making. *Obrazovatel'nye tekhnologii i obshchestvo* 2015; 1: 472–482. (in Russian)

# Teacher attitudes in the design of learning activities through technology

R. Martinez-Lopez<sup>1,2</sup>, C. Yot<sup>2</sup>, M. Sacchini<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Universidad de Sevilla, C/Pirotecnia s/n, Sevilla 41013, Spain

---

## Abstract

This study explores the use of technology in the design of learning activities by Russian teachers and their relationship with the technology self-efficacy in higher education. The Inventory of Learning Activities with Technologies in the University was translated and adapted to Russian context and validated by retest method. Answers were classified through content analysis. Findings suggest that access to technology, on-line courses, and data elaboration software among teachers should be enhanced. Teachers' self-confidence and use of technologies are related: one increases the level of another and vice-versa.

*Keywords:* learning activity; teacher, attitude; self confidence; technology use; supercomputing education; high performance computing; open question; Russian university

---

## 1. Introduction

Since supercomputing is considered a strategic area [1], the relevance of supercomputing education is increasing [2, 3]. The effort taken by Russia is justified [4, 5]. Nevertheless, despite its success, it is considered that improvement could be developed [6]. An analysis of the state of things of 2010 [4], showed that supercomputing education is narrowed to studying only several simple technological subjects at the University. Due to that, an appropriate education of users in supercomputer centers is critically low [3]. A recent analysis of the current state of the High Performance Computing (HPC) and Computational Science research [7], briefly highlighted the necessary scalability at all levels and highly trained computational scientists with the ability and skills to approach complex scientific problems.

The skills necessity claim of specialists in supercomputing [7], are not new. The workforce and the technology disparity has been researched for a long time [9]. Proposals to bridge the skills gap have covered, for example, the HPC competency as a requirement in the research and engineering curricula [1]. Following this approximation, the unification of skills at the university level and advanced research methodology has been proposed [10, 7], curriculum contents and knowledge assessment integration [11]. Other studies have been focused on the diffusion of supercomputing education to improve the use of supercomputer systems and the change of the higher educational system [3]. Proposals and research have tended to focus on the skills and knowledge required in a wide range of computer science issues of a specialist nature in the area of supercomputing technologies [7], rather than attitudes. For example, the recent analyses about training students in clusters competition in Thailandia, found that attitude is one of the missing elements [8], due to the Thai personality characteristics, only interested students counted.

Studies on the learners and the teachers could help to determine if they are ready for a new technology [12]. Student-centered approaches to learning have encouraged teachers to integrate technologies into their teaching. Empirical research reported that teacher attitudes and personal use of technology, accounted for 55% of the variance [13]. Considering teachers as facilitators, the incorporation of technology into their teaching is critical [14]. In fact, it is necessary to understand how the implementation of a technology could improve the perceived competence and use of teachers in their teaching [15]. Recent research to determine the possibilities of using technology in high education highlighted the increase of favorable circumstances for a professional competency development [16]. From a technology-enhanced learning perspective, understand the reasons of teachers using or not technology and what they should know in order to use it, requires further research [17, 18]. In this context, teachers requirement of more preparation is a relevant issue in Russia [19]. Moreover, what technologies do Russian teachers use related to learning activities in HPC is missing.

The "Inventory of Learning Activities with Technologies in the University" (IAATU) [20] is applied to assess what type of technology Russian university teachers are using to design learning activities and which is they level of self-confidence to use it. The problem of training specialists in the field of supercomputing has been widely discussed but, no from this perspective. This study is a contribution to new knowledge in the field of HTC, from the approach of teacher attitudes, specifically confidence, with the use of technology in the design of learning activities.

## 2. The object of the study

### 2.1. Supercomputing in education

The HPC concept circumscribes the computer server systems market, software, networking and services used to manage computationally intensive, data simulation and analytic problems [1]. The inclusion of parallel computing technologies in supercomputing education [3], specifically into the engineering curricula [1, 5] is recommended. In fact, software is considered relevant into the HPC leadership [1], and the access to industry tasks, supercomputer technologies and systems [5].

Nevertheless, not all is about tools [3]. Indeed, skills demand a mixed understanding of a scientific discipline and computer



technique [1]. Moreover, supercomputer centers require diverse educational activities [3], for example, complex scientific problems like methods of scientific inquiry [7], parallel programming technologies or architecture computer [3]. For this reason, a collaborative learning approach [7] could be useful to understand and explain the high complexity of computing systems [3] and to integrate HPC into the curricula.

## 2.2. *The use of technology in the design of learning activities*

Despite the recognition of HPC based simulation as belonged to the scientific inquiry and recommendations of integrating computational science methods in universities [1], it is difficult to add new content to the science and engineering curricula [1]. An option to solve this problem is the use of learning activities, which increase the motivation and the training results. Learning activities design is associated with the learning purpose, for example, test questions are used to evaluate knowledge and skills, test assignments for practical skills in parallel and distributed computing tasks, and demonstrations for teaching problem-oriented specialists [5]. Furthermore, HPC and computational functional skills enrich through the use of technology in the learning process. For example, the use of technology let students experience by themselves the challenge of “writing meaningful simulations” without any access to HPC devices [21]. This learning activity through technology let them to understand HPC simulations that can be replicated on a big scale. This example illustrates how the use of technology in the design of learning activities can promote supercomputer technologies throughout the curriculum [5].

Instead of focusing in “what should we teach” or “what should be included in the new curricula” [5], the purpose of this research is related to explore how teachers are using technology and what kind of technology they are using to design learning activities.

## 2.3. *High Performance Computing competence: teacher attitudes*

Teachers are one of the target groups of supercomputing education infrastructure [5] in fact, due to they are often busy [5] their effort in teaching HPC is very valuable [21]. Knowledge and skills in parallelism concepts [3] are critical but, from the social psychology theory approach, also attitude. Attitude to technology influence its use [22]. Teachers’ own technology practices and the type of technology activities they assign to students, let understand the development of teaching HPC from the teachers’ perspective. A previous knowledge about teacher self-confidence in the technology used for learning activities design can be relevant to address an education program related to HPC [5]. Even more useful, if these teachers are going to train potential qualified candidates to HPC positions like “university graduates in mathematics, engineering, or physical sciences” [1].

Through IAATU, this study explores the use of learning activities with technologies and self-confidence by teachers at Russia universities.

## 3. Methods

### 3.1. *Population*

In the original study, 103 answers were collected from the online survey, since February to April, 2016. 52.4 % females and, 47.6 % males. 43.7%, in the age group of 31-40, 17.5 under 31 and 9.7% over 61. 44.7% of the teachers from Russia Universities. Re-test (n=48) was realized at Samara National Research University. 47,9% women and 52,1% men. In relation to age 16,7% between 20 and 30 years old, 27,1% (31-40), 27,1% (41-50), 12,5% (51-60) and 16,7% (61-70).

### 3.2. *Instrument*

In order to explore to what extent university professors are using technology as a pedagogical support resource, IAATU was used [20]. The adapted Russian version of IAATU [28], with 38 items distributed among 1 to 5 on a double Likert-type scale, collect demographic information such as: gender, age, university, field of knowledge and professional category. One scale refers to use frequency asking “to what extent do you perform the activities described in the item?” while the other refers to the degree to which the teacher feels confident using the technology with the question “If you perform them, to what extent do you feel comfortable?” Two open questions in relation to technology not included and the use of technology at the University, are contemplated.

### 3.3. *Analysis*

Descriptive statistical methods were employed to analyze the level of the participants in self-confidence and in the use of technology. IBM SPSS Statistics and univariate were used to describe the characteristics and activities learning technologies frequency of the participants. In addition to pretest the Russian target language version of IAATU, re-test with target language subjects was conducted [23]. Temporal stability of the responses were analyzed on the same group of respondents with an interval of one month by means of a method based on the use of IAATU [24]. Moreover, the correlation coefficient between the two intervals of IAATU were examined [25], through Pearson Correlation Coefficient. Likewise, considering coefficient alpha and retest as index of reliability, were calculated [25]. The estimated internal consistency of each scale in the retest is provided in order to increase confidence in measure [26]. A Spearman's correlation was run to assess the relationship between technology use in learning activities and self-confidence on that technology, of a small sample of 48 teachers aged 31-50 years old at Samara National Research University.

In order to analyze the content of open questions a frequency criterion was adopted. That is, higher was the number of the repetition of the same or similar terms in the answers, higher was the importance of such words. In this case, the words of the answers were also to be evaluated as word-concepts to count in order to establish which specific problems of access to technology are present among the teachers of Russian Universities. For example we got a word-concept as <lack of software> from analyzing and summing up an answer like: “I can not use technology in class, because there is no software to be used for economic tasks. Or at least I do not know them.”

#### 4. Results and Discussion

Scales of level of use (Cronbach's  $\alpha = .91$ ) and self-confidence ( $\alpha = .93$ ). Results re-test in relation to use ( $\alpha = .93$ ) and self-confidence ( $\alpha = .94$ ) were reported with a value above Cronbach's  $\alpha = .95$  scales. It showed very good reliability and internal consistency, which meets the criteria of reliability [27]. As previously reported, association between the use of learning activity and self-confidence in the Russian adaptation of IAATU were established [28].

The test-retest reliability coefficient of use scale (Figure 1) showed that there was a moderate positive correlation, which means there is a tendency for high use variable (n=103) scores of the original test go with high use (n=48) variable scores of re-test (and vice versa),  $r = .56$ ,  $p = \leq .001$ , with a  $R^2 = .323$ . In the confident scale (Figure 2), there was a moderate positive correlation, which means there is a tendency for high self-efficacy variable in the original test (n=103) scores go with high self-efficacy variable scores in the re-test (n=48), and vice versa,  $r = .56$ ,  $p = < .001$ , with a  $R^2 = .322$ .

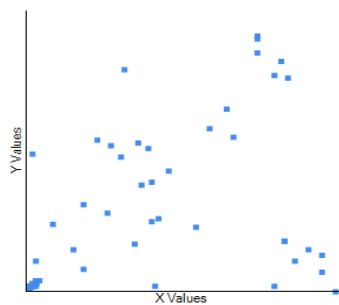


Fig.1. Pearson correlation use.

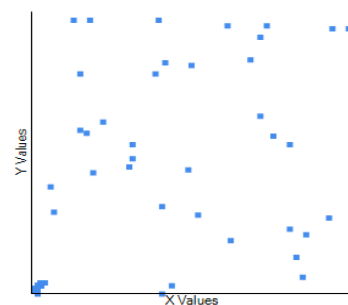


Fig.2. Pearson correlation self-confidence.

In relation to the re-test analysis, the average use and confidence score for each item were calculated (Table 1). Three groups of learning activities are identified in relation to the mean: low level (mean 1-2.5), medium (2.5- 3.5) and high (3.5-5). The average confidence is higher than the use. According to the same statistic (Table 2), up to 8 items have a high average use and confidence score (3.5-5). According with previous research, this results suggest that teachers' technology previous practice and confidence could determine their use of technology in their teaching [29].

Table 1. Results of the Spearman Rho correlation coefficient by items.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Coefficient	,521	,699	,287	,483	,698	,625	,720	,541	,845
Sig. (bil)	,000	,000	,066	,001	,000	,000	,000	,002	,000
	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18
Coefficient	,675	,822	,766	,731	,719	,929	,667	,890	,791
Sig. (bil)	,000	,000	,000	,000	,000	,000	,000	,000	0
	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27
Coefficient	,636	,853	,908	,839	,603	,752	,723	,925	,913
Sig. (bil)	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Item 28	Item 29	Item 30	Item 31	Item 32	Item 33	Item 34	Item 35	Item 36
Coefficient	,947	,942	,812	,803	,919	,949	,910	,664	,971
Sig. (bil)	,000	,000	,000	,000	,000	,000	,000	,000	,000
	Item 37	Item 38							
Coefficient	,644	,935							
Sig. (bil)	,000	,000							

In addition for five items, trust is high (3.5-5) and the use is medium (2.5-3.5): Item 4 “During my presentations and to facilitate my students’ understanding of given concepts and ideas, I use video segments found on Internet” ( $M=3.11$   $SD=1.18$ ,  $M=4$   $SD=1.04$ ), Item 6 “Using the virtual platform, I provide my students with videos, demonstrations, simulations, experiences and/or cases to expand the information they received” ( $M=2.52$   $SD=1.41$ ,  $M=3.56$   $SD=1.48$ ), Item 9 “I select text documents and I make them available to my students on the virtual platform in an effort to improve the reading understanding of my subject

content” ( $M=3$   $SD=1.53$ ,  $M=3.78$   $SD=1.45$ ), Item 13 “I design practical cases, using digital resources (videos, presentations, specific software, etc.), so that students can apply the theory learned to practical cases” ( $M=3.25$   $SD=1.37$ ,  $M=3.95$   $SD=1.18$ ), Item 22 “I design problems in which students have to solve complex problems, using digital resources, similar to those a professional would use” ( $M=3.38$   $SD=1.56$ ,  $M=3.95$   $SD=1.37$ ). Further research would be necessary in order to explore the reasons of this dissonance.

Spearman correlation (Table 1) was calculated. There was a strong positive correlation between use of technology and self-confidence, which was statistically significant,  $r_s = .7217$ ,  $p = .0$  in all cases, except for item 3 ( $M=3.65$   $SD=1.15$ ,  $M=4.38$   $SD=0.85$ ), “During my presentations, I show students some type of simulations, demonstrations or examples based on digital resources, either my own, or available on the web, to clarify concepts and ideas”. Unlike the initial results, the strength of the association of the items is not coincident in most cases. It would be necessary further research in order to determine the reasons.

According to the Mann-Whitney U tests only the level of use of item 18 “I design activities in which students must provide comments or given their point of view by means of personal or group blogs” differs according to sex (sig. 0.32). The confidence level of items 17 “I facilitate interaction with students outside the classroom by means of cellphone applications such as WhatsApp, Line, Twitter, Facebook, etc. to motivate the exchange of information, the resolution of doubts...” (sig. 0.15), 18 “I design activities in which students must provide comments or given their point of view by means of personal or group blogs” (sig. 0.15), 29 “I use virtual platform tools so that students can turn in homework/papers for my subject” (sig. 0.40), 30 “When assessing students, I use electronic portfolios, created on the actual platform or with specific online tools, for continual assessment” (sig. 0.002) differ according to gender. According to Kruskal, the level of use of items 20 “I ask students to write reports, essays, articles, etc. using reference management tools such as Zotero, Refworks, Mendeley, Endnote...” (sig. 0.009) and 38 (sig. 0.039) differed according to age. Only the confidence of 38 “During my teaching activities, I attend the terms of use for the digital materials that have a Creative Commons license” (sig. 0.46) differed according to age.

Table 2. Average use and confidence score.

Item	Use		Confidence	
	M	SD	M	SD
Item 1	3.90	1.13	4.58	.65
Item 3	3.65	1.15	4.38	.85
Item 10	4.31	1.05	4.61	.68
Item 16	4.17	1.05	4.48	.94
Item 19	3.51	1.33	4.21	1.04
Item 23	3.70	1.35	4.29	1.01
Item 35	4.02	1.29	4.28	1.14
Item 37	4.18	1.05	4.27	1.04

In order to complete the study content analysis of open questions from the original study ( $n=103$ ), (Table 3) and (Table 4), following frequency criteria was applied. The first open question (Table 3) was: “If your learning process involves technologies that are not listed in the questionnaire, please describe them”. It is highlighted that 6 answers specifically mentioned the System Management Learning (LMS) Moodle, as a technology not mentioned in IAATU. Due to Moodle is the main LMS used in Russian universities [30], and it is recommended in the learning administrative process and the training [7], its mention is also reasonable because of the geographically distributed country of Russia. Further research using IAATU should include an item in reference to Moodle, in order to accomplish a better adaptation of the questionnaire in the Russian higher education context.

Table 3. Technologies not listed in IAATU.

Type of technology	Number of answers
Courses	11 / 55
Software	9 / 55
Others	9 / 55
No answers	26 / 55

The second open question (Table 4) refers “If you want to leave a comment on the questionnaire with questions on the educational process with the use of technology at the University”. 13 answers mentioned the University as responsible of the use of technology by teachers. 9 of them referenced the lack of access and availability of technology and 3, training needs.

Table 4. Problems to implement technology.

Type of answer	Number of answers
Questionnaire	17 <sup>[A1]</sup> / 47
University	13 / 47
Self-confidence	6 / 47
No answers	16 / 47

[A1] count 2 comments of one person as 2

Analysis showed that the more frequent problems reported by teachers at Russian University are technology acknowledge. This allowed us to understand that at the present day Russian Universities suffers of an insufficient usage of technological means and structures. Another problem connected to the access of technology could be explained also with the actual lack of software knowledge of the newest technology adapted to be used for specific task in the didactic by the Engineering and Economy Faculties. Related to this, it is recommended the application of collaborative pedagogical approaches after the implementation of IT solutions [7], and it can be done by the design of learning activities.

The analysis of the self-confidence answer category cluster related to the teacher behaviour in relation to the use of technology for learning activities. 5 answers are related to the use of technology in relation to self-confidence, for example: “If I

do not use any technology, it is not always due to the fact that I don't feel comfortable with it I can not know it, or do not have access or do not have enough time to do it.” or “I can not use ready-made test (...)” and 1 to familiarity with ICT. 3 answers were a mix between comments in reference to the questionnaire (2), for example: “Not all the tools described in the questionnaire are familiar but, answers are formulated on the basis of analog products with similar functionality in the educational process”, “I do a lot of issues of teacher training and blended learning technology development of online courses so, I would be more interested in questions of technology implementation strategies (...)”; and the University (2), “I'm open and ready to cooperate with the organization in e-learning. They need to train teachers in online methodology”, “Lectures, laboratory and practices can not be transferred into the virtual space. Moodle is not available for all students”.

It is plausible that this research may have limitations that could have influenced the results obtained. First, the sample size of our analysis are not enough to make generalizations about the more frequent problems reported by teachers at Russian University. As such, the findings should be taken with caution. Second, the high value of no answers to the open questions can be interpreted as perceived ambiguities in the meaning due to the fact that back-translation was not applied. Although it is not considered mandatory [31], it provides an assertion that the instrument is the same in two languages [23]. Third, retest assessments could introduce bias, due to the risk that respondents desire to appear consistent [32]. Furthermore, due to the IAATU was translated into a new language, from Spanish to Russian, to avoid the assumption of hypotheses about the dimensionality of a given set of items, exploratory factor analysis (EFA) could be applied [33]. Finally, considering that the support and maintenance of supercomputer centers require specific technology in order to automatic decision making on emergency situations, monitoring, or high performance tasks for the infrastructure, and others [3], an adaptation of the items to this context could be required.

Despite of the limitations, the results of this study coincide with previous research [34, 20] level of use of learning activity in teachers depend on their technology self-efficacy [35]. Furthermore, IAATU could serve as a tool to identify teachers that already have the attitude to serve as a bridge between the specialized knowledge of scientists and practitioners on the one hand, and scholars [21]. Even more, further research could consider the identification of teacher profiles in relation to their level of use of ICT in the design of learning activities [20]. Concerning to the use of digital technology, recent empirical studies pointed that patterns of technology use emerge from the frequency of use and by the nature of the activity [36]. Similarly, IAATU could be applied to define what kind of learning activity teachers are using: lecturer's presentation, communication, information management, application, evaluative or productive [20]. Due to the recommendation of a collaborative learning approach in supercomputing teaching [7] further research is required in order to determine what kind of learning activities could promote it.

Moreover, evidence has showed how teacher confidence in a task can be regulated by self-efficacy [37]. However, as confidence does not necessarily specify what the certainty is about [37], further data collection is necessary in order to determine exactly how confidence affects the use of technology in the design of learning activities at Russian Universities context.

## 5. Conclusion

This study contributes to the understanding of technology use and confidence (self-efficacy) in the design of learning activities from the teacher perspective into the higher education system in Russia. Relationship between teachers' own technology practices and the type of technology activities they assign to students has been examined. From a competency-based approach that acquires more holistic structure [38], this study highlight the importance of attitude, specifically, self-confidence as complementary to skills and knowledge in HPC. Previous research pointed advanced training for university teachers in various applied areas where supercomputing systems can be used for problem solving [4]. This study suggests that supercomputing education could enrich from an approach that take into account personal beliefs and actions based on attitudes, for example, teachers' confidence in their technology use. The approach of this research is a first step to understand the development of teaching HPC from the teachers' perspective.

## Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation and the 5-100 program. As part of first author's doctorate thesis, the main author desires to transmit thanks to Professor Elenev Valeriy Dmitrievich, Director of the Institute of Aviation Technology at Samara National Research University, for invaluable support of this study.

## References

- [1] IDC, High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy, 2015.
- [2] Zvacek S, Restivo MT, Uhomoihibi J, Helfert M. Computer Supported Education: 7th International Conference, CSEDU 2015. Lisbon, Portugal, May 23-25, 2015. *Commun. Comput. Inf. Sci.* 2016; 583: 152–168.
- [3] Voevodin V. Efficiency of Exascale Supercomputer Centers and Supercomputing Education. *Commun. Comput. Inf. Sci.* 2016; 14–23.
- [4] Voevodin V, Gergel V. Supercomputing education: the third pillar of HPC 2010; 11(4): 117–122.
- [5] Voevodin V, Gergel V, Popova N. Challenges of a systematic approach to parallel computing and supercomputing education. *European Conference on Parallel Processing*, 2015; 90–101.
- [6] Ministry of Education and Science of the Russian Federation. Supercomputing Education Project. Russian Presidential Commission on modernization and technological development of Russian economy.
- [7] Alexandrov N. Education and training for exascale. *J. Comput. Sci.* 2016; 14: 69–73.
- [8] Chantrapornchai C, Uthayopas P. A Road to Student Cluster Competition for Thailand. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) A, 2016.
- [9] Joiner DA, Gray P, Murphy T, Peck C. Teaching parallel computing to science faculty. *Proc. Elev. ACM SIGPLAN Symp. Princ. Pract. parallel Program*, 2006; 239.
- [10] Berzins M, Kirby R, Johnson C. Integrating teaching and research in HPC: experiences and opportunities. *Comput. Sci.* 2005; 36–43.
- [11] Gergel V, Meyerov I, Sysoyev A. Unified Assessment of Skills in Parallel and Distributed Computing. *Fac. Comput. Math. Cybern.* 2007; 5–6.

- [12] Yanuschik OV, Pakhomova EG, Batbold K. E-learning as a Way to Improve the Quality of Educational for International Students. *Procedia - Soc. Behav. Sci.* 2015; 215: 147–155.
- [13] Wozney L, Venkatesh V, Abrami P. Implementing Computer Technologies: Teachers Perceptions and Practices. *J. Technol. Teach. Educ.* 2006; 14(1): 173–207.
- [14] Çatma Z, Corlu MS. How special are teachers of specialized schools? Assessing self-confidence levels in the technology domain. *Eurasia J. Math. Sci. Technol. Educ.* 2016; 12(3): 583–592.
- [15] Lemon N, Garvis S. Pre-service teacher self-efficacy in digital technology. *Teach. Teach.* 2016; 22(3): 387–408.
- [16] Zagrebina EI, Sharafetdinova ZG, Lushchik IV, Konyushenko SM, Ermoshina NV, Kosyakova EY, Ashrafullina GS. The Electronic Learning System as a Means of Forming Professional Competencies among University Students. *J. Sustain. Dev.* 2015; 8(3): 178–184.
- [17] Ertmer PA, Ottenbreit-leftwich AT. Ertmer & Ottenbreit-Leftwich teacher knowledge confidence and beliefs 2010; 42(3): 255–284.
- [18] Schweighofer P, Grünwald S, Ebner M. Technology Enhanced Learning and the Digital Economy: *Int. J. Innov. Digit. Econ.* 2015; 6(1): 50–62.
- [19] Grigorevna MN. Pedagogical Maintenance of Future Teachers. *Practice-oriented Training.* 2015; 8.
- [20] Marcelo C, Domínguez C, Mayor Ruiz C. University Teaching with Digital Technologies. *Comun. Rev. científica Iberoam. Comun. y Educ.* 2015; 45: 117–124.
- [21] Hilpert J, Berlich R, Lürßen P, Zwölfer A, Barwind J. Teaching Simulations and High Performance Computing at Secondary Schools in the German State of Baden-Württemberg. *Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015; 731–738.*
- [22] Hogg MA, Vaughan GM. *Essentials of Social Psychology* 2010; 16(3).
- [23] Maneesriwongul W, Dixon JK. Instrument translation process: A methods review. *J. Adv. Nurs.* 2004; 48(2): 175–186.
- [24] Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila. Pa. 1976)* 2000; 25(24): 3186–3191.
- [25] Noar SM. The Role of Structural Equation Modeling in Scale Development. *Struct. Equ. Model. A Multidiscip. J.* 2003; 10(4): 622–647.
- [26] Lovelace M, Brickman P. Best practices for measuring students' attitudes toward learning science. *CBE Life Sci. Educ.* 2013; 12(4): 606–617.
- [27] Gliem JA, Gliem RR. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research to Practice Conference in Adult, Continuing, and Community Education* 2003; 1992: 82–88.
- [28] Martinez-Lopez R, Reznichenko M, Yot C, Marcelo C. Inventory of Activities of Learning Technologies at University: Cross-Cultural Adaptation in the National Context of Russia. *Eng. Educ.* 2016; 20: 57–63.
- [29] Chuang HH, Weng CY, Huang FC. A structure equation model among factors of teachers' technology integration practice and their TPCK. *Comput. Educ.* 2015; 86: 182–191.
- [30] Grigorievich D, Gennadievna N. *Russian Universities: Towards Ambitious Goals* 2016; 11(8): 2207–2222.
- [31] Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *J. Clin. Epidemiol.* 2015; 68(4): 435–441.
- [32] Polit DF. Assessing measurement in health: Beyond reliability and validity. *Int. J. Nurs. Stud.* 2015; 52(11): 1746–1753.
- [33] Polit DF, Beck CT. The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Res. Nurs. Health* 2006; 29: 489–497.
- [34] Yot C, Marcelo C. *De la tiza al teclado: Enseñar y aprender con tecnologías digitales*, Grupo de i, 2016.
- [35] Marcelo C, Yot C. Pedagogies of working with technology in Spain. *Adv. Res. Teach.* 2015; 22B: 331–357.
- [36] Area-Moreira M, Hernández-Rivero V, Sosa-Alonso JJ. Models of educational integration of ICTs in the classroom. *Comunicar* 2016; 24(47): 79–87.
- [37] Bandura A. Self-efficacy: The Exercise of Control. *Encycl. Hum. Behav.* 1997; 4: 71–81.
- [38] Erganova NE, Shutova TV. Cluster model of designing competencies of a future vocational school teacher. *Middle - East J. Sci. Res.* 2014; 19(1): 89–93.

# Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources

A.A. Zarubin<sup>1</sup>, A.R. Koval<sup>1</sup>, V.S. Moshkin<sup>2</sup>, A.A. Filippov<sup>2</sup>

<sup>1</sup>The Bonch-Bruевич Saint - Petersburg State University of Telecommunication, 61 Moika, Saint - Petersburg, 191186, Russia

<sup>2</sup>Ulyanovsk State Technical University, 32 Severny Venetz str., Ulyanovsk, 432027, Russia

---

## Abstract

The activities of any large organization requires the work of specialists with a large volume of unstructured information in order to obtain and extract the necessary knowledge to interact with partners, decision-making, etc. An array of unstructured textual information is not adapted to structuring and semantic search. Thus, development of intelligent algorithms and text analysis methods for dynamic generation of the knowledge base contents is needed. Extract of syntagmatic structure of a text and further representation of extracted knowledge in the form of a single unified ontology allows to get access to the knowledge base for solving complex problems.

*Keywords:* ontology; knowledge base; syntagmatic analysis; text resource

---

## 1. Introduction

In the process of any large modern organization activity, it is necessary to make urgent management decisions timely that requires specialists to have deep knowledge of the problem area (PrA). Moreover, they should be able to use different decision support systems and tools for work with knowledge.

The desire to automate and speed-up the process of obtaining necessary knowledge about the PrA drives the need in the unified multipurpose toolkit for knowledge management that does not require a user to have some additional skills in the field of knowledge engineering and ontological analysis.

Thus, one can identify a number of scientific problems besetting modern organizations. In order to be solved, such problems require the systematic approach and include the following ones:

- the need of developing the semantic basis for representation of electronic information storage content;
- the lack of integrative conceptual models using different approaches to the storage of knowledge about the PrA;
- the need of unified the automated processing of the stored knowledge;
- the need of simultaneous use of multi-aspect contexts of the PrA under consideration;
- the need of solving the problem of tracking the clarity of human reasonings.

Thereby, nowadays, the actual problem is providing specialists of a wide range of organizations with a universal tool allowing to address the knowledge management challenges [1]. Furthermore, the tool should not require some extra training of users.

At the moment, the ontological approach is most often used for organization of knowledge bases of expert systems. A lot of Russian and foreign researchers such as T.A. Gavrilova [2], V.N. Vagin [3], V.V. Gribova [4], Yu.A. Zagorulko [5], A.S. Kleschev [6], I.P. Norenkov, D.E. Palchunov, S.V. Smirnov [7], D. Bianchini, T.R.Gruber, A.Medche, G. Stumme and others address the problem of integration and search of information in order to provide management decision support on the basis of an ontology.

In a broad sense, ontologies are models representing knowledge within the individual contexts of the PrA in the form of semantic information-logical networks of interrelated objects where the PrA concepts with properties and relations between objects are the main elements.

Ontologies serve as integrators proving the common semantic basis in the processes of decision-making and data mining, and the unified platform for combination of different information systems [8,9].

## 2. Formal model of knowledge base

The knowledge base (KB) represents the storage of knowledge of different PrAs and contexts in the form of an applied ontology. The PrA ontology context is a specific state of the KB content that can be chosen from a set of the ontology states. The state was obtained as a result of either versioning or constructing the KB content from different points of views.

Formally, an ontology can be represented by the following equation:

$$O = \langle T, C^{T_i}, I^{T_i}, P^{T_i}, S^{T_i}, F^{T_i}, R^{T_i} \rangle, i = \overline{1, t},$$

where  $t$  is a number of ontology contexts,  $T = \{T_1, T_2, \dots, T_n\}$  is a set of ontology contexts,  $C^{T_i}$  is a set of ontology classes within the  $i$ -th context,  $I^{T_i}$  is a set of ontology objects within the  $i$ -th context,  $P^{T_i}$  is a set of ontology classes properties within the  $i$ -th context,  $S^{T_i}$  is a set of ontology objects states within the  $i$ -th context,  $F^{T_i}$  is a set of the PrA processes fixed in the ontology within the  $i$ -th context,  $R^{T_i}$  is a set of ontology relations within the  $i$ -th context defined as:

$$R^{T_i} = \left\{ R_C^{T_i}, R_I^{T_i}, R_P^{T_i}, R_S^{T_i}, R_{F_{IN}}^{T_i}, R_{F_{OUT}}^{T_i} \right\},$$

where  $R_C^{T_i}$  is a set of relations defining hierarchy of ontology classes within the  $i$ -th context,  $R_I^{T_i}$  is a set of relations defining the 'class-object' ontology tie within the  $i$ -th context,  $R_P^{T_i}$  is a set of relations defining the 'class-class property' ontology tie within the  $i$ -th context,  $R_S^{T_i}$  is a set of relations defining the 'object-object state' ontology tie within the  $i$ -th context,  $R_{F_{IN}}^{T_i}$  is a set of relations defining the tie between  $F_j^{T_i}$  process entry and other instances of the ontology within the  $i$ -th context,  $R_{F_{OUT}}^{T_i}$  is a set of relations defining the tie between  $F_j^{T_i}$  process exit and other instances of the ontology within the  $i$ -th context.

### 3. Extracting the core of ontology of the problem area based on the syntagmatic analysis of external wiki-resources

Wiki-resources are formed by a large number of users. Thus, applying of the automated methods for extracting the core of the ontology based on the knowledge contained in the Wikipedia, can reduce the degree of subjectivity and increase the number of experts involved in the process of the ontology building [11].

The algorithm of extracting the core of the ontology from the external wiki-resources is based on the methods described in [3].

The PrA features in the wiki-resource are represented as a hierarchy of associated hyperlinked HTML-pages with a certain semantics. The core of the ontology is automatically extracted from external wiki-resources in the process of data mining. The core of the ontology can be expanded in the process of the syntagmatic analysis of a set of thematic text documents.

The first method of extracting the core of the PrA ontology is based on the Lee algorithm [13]. Concepts are reduced to the initial form (lemmatization). Defining types of relations between concepts is in the process of the syntagmatic analysis of terms located on the right and the left of reference defines the concept. The rules for determining the type of relations are presented in the form of syntagmatic patterns (patterns contain a sequence of words).

The second method of extracting the core of the domain ontology based on the contents of wiki-resources allows the intelligent system to adapt dynamically to the changes in the domain [14]. Methods of automatic text processing (ATP) in a natural language (NL) can be used in order to extract knowledge from the text of the wiki resource pages.

The ATP process is usually carried out in several steps [15]:

1. Grafematic analysis is the process of initial analysis of the text in a NL. The grafematic analysis presents the input data in a convenient format for further analysis (separation of input text into words, delimiters, etc).
2. Morphological analysis (lemmatization) is a process of transforming the words of the input text to the initial form defining the part of speech, gender, case, etc.
3. Parsing is the process of selecting members of simple sentences and constructing a parse tree.
4. Semantic analysis consists of
  - construction of a semantic tree of sentences,
  - semantic interpretation of words and constructions,
  - definition of semantic relations between elements of the text.

Semantic representation of the text in a NL is the most complete of those that can be achieved only by linguistic methods. The core of the domain ontology can be extended by merging with the semantic tree extracted from wiki-resources. It is necessary to develop a method for translating a parse tree into a semantic tree in order to obtain a semantic tree.

It is necessary to determine the syntactic structure of the sentence for constructing the semantic tree of sentences in a NL. There are several parsing tools of texts in Russian, for example [16, 17, 18]:

- Lingo-Master;
- Treeton;
- Sreda RGTU;
- DictaScope Syntax;
- ETAP-3;
- ABBYY Compreno;
- Tomita-parser;
- AOT etc.

In the context of this work, AOT (tool for constructing a parse tree) was chosen [18]. Let us consider the application of the algorithm for translating a syntactic tree into a semantic tree using the example of a test fragment in Russian:

*Онтология в информатике — это попытка всеобъемлющей и подробной формализации некоторой области знаний с помощью концептуальной схемы.*

The parse tree for the test fragment is shown in the figure 1.

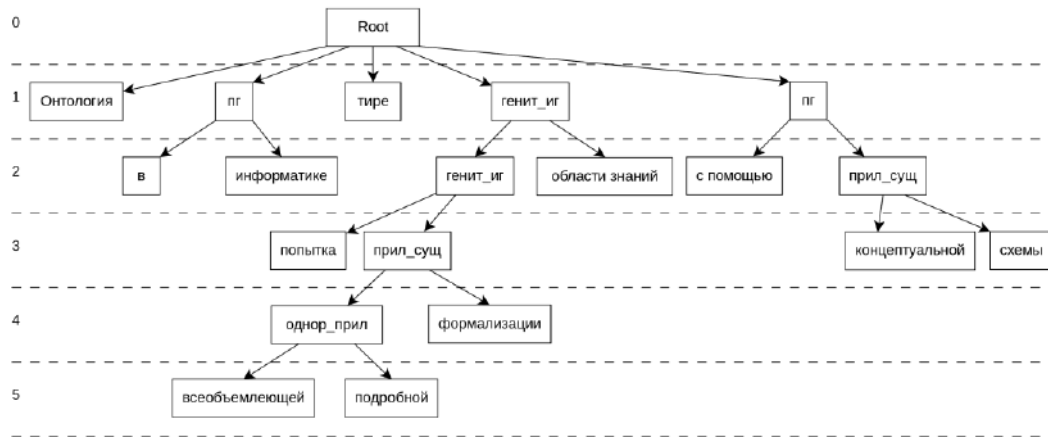


Fig. 1. Example of a parse tree.

Formally, the function of translating a parse tree into a semantic tree can be represented as follows:

$$F^{Sem} : \{N_{li}^{Synt}, P_j\} \rightarrow \{N^{Sem}, R^{Sem}\},$$

where  $N_{li}^{Synt}$  is the  $i$ -th node of the  $l$ -th level of a parse tree. For example, for the parse tree in Figure 1, the first node of the first level is the node “Онтология”, the second one is “пг”, the third one is “тире”, etc. The node of the parse tree can be a member of the sentence, for example, the node “Онтология”. Also, the parse tree node can be a syntactic label that defines the constituent members of the sentence, for example, “пг” (the prepositional group);  $P_j$  is the  $j$ -th syntagmatic pattern for defining the nodes of the parse tree. The nodes will be translated into nodes and relations of the semantic tree. The syntagmatic pattern is a collection of several words united according to the principle of semantic-grammatical-phonetic compatibility. Formally, syntagmatic pattern can be represented as follows:

$$(N_1^{Synt}, N_2^{Synt}, \dots, N_k^{Synt}) \rightarrow \{N^{Sem}, R^{Sem}\}, k = \overline{1, K},$$

where  $N_k^{Synt}$  is the  $k$ -th syntagmatic unit of the pattern corresponding to the node of the parse tree. It is necessary to use all the syntagmatic units included in it in order to use the syntagmatic pattern. Examples of syntagmatic patterns and the results of their use are presented in Table 1;

$K$  – number of syntagmatic units in the pattern;

$\{N^{Sem}, R^{Sem}\}$  are the sets of nodes  $N^{Sem}$  and relations  $R^{Sem}$  of the semantic tree obtained as a result of translation of the parse tree into a semantic tree. Formally,  $R^{Sem}$  can be defined as follows:

$$R^{Sem} = \{R_{isA}^{Sem}, R_{partOf}^{Sem}, R_{associateWith}^{Sem}, R_{dependsOn}^{Sem}, R_{hasAttribute}^{Sem}\}$$

where  $R_{isA}^{Sem}$  is a set of transitive relations of hyponymy;

$R_{partOf}^{Sem}$  is a set of transitive relations “part/whole”;

$R_{associateWith}^{Sem}$  is a set of symmetrical relations of association

$R_{dependsOn}^{Sem}$  is a set of asymmetric relations of associative dependence;

$R_{hasAttribute}^{Sem}$  is a set of asymmetric relations describing the attributes of nodes.

Table 1. Examples of syntagmatic patterns and the results of their application.

Initial data	Syntagmatic pattern	Result
попытка-генит_иг-формализации	{node1}-{генит_иг}-{node2} → {node1}-associateWith-{node2}	попытка-associateWith-формализация
в-пг-информатике	{node1}-{пг}-{node2} → {prevNode}-getRelation(node)-{node2}	lastNode-relation-информатика
тире	{тире} → {prevNode}-isA-{nextNode}	lastNode-isA-nextNode
концептуальной-прил_сущ-схемы	{node1}-{прил_сущ}-{node2} → {node2}-hasAttribute-{node1}	схема-hasAttribute-концептуальный
(всеобъемлющей, подробной)	{node1}-{однор_прил}-{nodes} →	формализация-hasAttribute-
однор_прил-формализации-	{node1}-hasAttribute-{nodes[1]}, {node1}-hasAttribute-{nodes[2]}, {node1}-hasAttribute-{nodes[...]}, {node1}-hasAttribute-{nodes[count]}	всеобъемлющий, формализация-hasAttribute-подробный

The algorithm for translating a parse tree into a semantic tree consists of the following steps:

1. Go to the first level of the parse tree.



2. Select the next node of the current tree level.
3. If the node is marked, go to step 2.
4. If the node is not a syntax label, go to step 9.
5. If the node is a syntax label and does not have child elements, go to step 9.
6. If the node is a syntax label and all its child nodes are not syntax labels, go to step 9.
7. If there is a temporary parent node, replace it, otherwise, create a temporary node.
8. If there is a previous node and there is no relation with it, add a temporary relationship with it and go to step 2.
9. Apply the syntagmatic pattern for translation.
10. Mark the processed nodes and go to step 2.
11. Go to the next level of the parse tree and go to step 2.

The resulting semantic tree for the test fragment is shown in Figure 2.

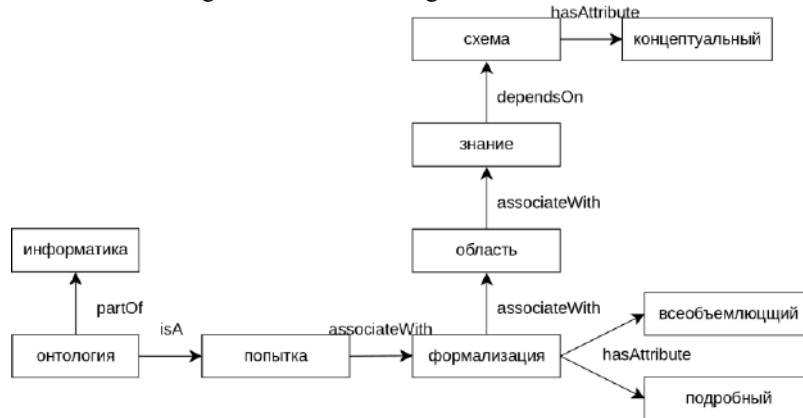


Fig. 2. Example of a semantic tree for a test fragment.

The result semantic tree can be merged with other semantic trees within the text. In addition, the semantic tree can be merged with the domain ontology compiled by an expert. Extending the knowledge base by merging semantic trees retrieved from semi-structured resources allows:

- provide a common terminology space for sharing and understanding by all users;
- determine the exact and consistent meaning of each term.

Ontology is a common terminological basis for complex iterative processes. Figure 3 shows the fragment of the core of the ontology “LAN Administration” extracted from the thematic wiki-resource.

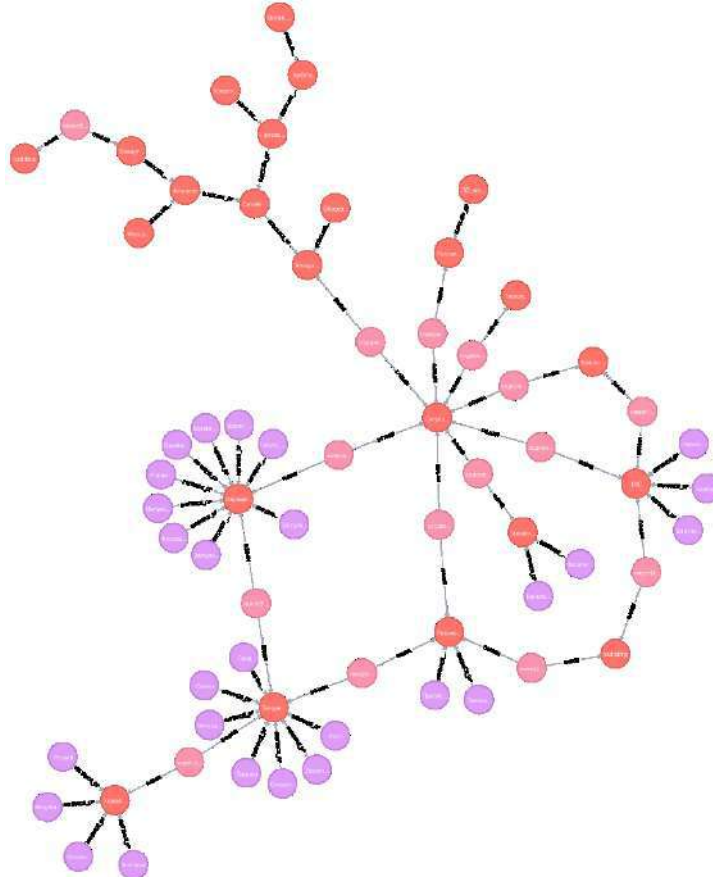


Fig. 3. The fragment of the core of the ontology “LAN Administration”.

#### 4. Construction of the PrA ontology based on the syntagmatic analysis of text documents

In the course of solving the problem of automated ontology expansion, two algorithms for terms extraction from domain texts using existing ontology core were developed:

- the thesaurus-based algorithm;
- the internal linkage algorithm [19].

The main feature of the developed algorithms is the term extraction from text documents by matching syntagmatic patterns with the lemmas of the objects from the core of the ontology. Syntagmatic patterns are extracted with the use of morphological analysis of text documents.

**The thesaurus-based algorithm.** A thesaurus is a reference work that lists words grouped together according to the similarity of meanings (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. Any ontology is a complicated version of the thesaurus.

The thesaurus approach assumes search of lemmas from the input words and their combinations among the terms defined in the ontology. For this purpose, each ontology class has a “HasALemma” property, which has a string value obtained by object name lemmatization.

The supporting ontology object used in the further analysis has the degree of proximity in relation to the input word / word combinations that is calculated by the following equation:

$$k_t = \max_{i=1}^m \frac{n_i}{p_i}, \quad (1)$$

where  $m$  is the number of all ontology objects,  $n_i$  is the number of words from the input sequence contained in the lemma of the current ontology object,  $p_i$  is the number of words in the current ontology object.

The process of assessing the proximity of the input words to the subject area terms is shown on Figure 4.

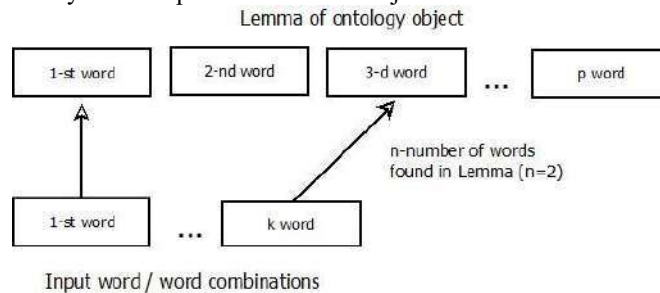


Fig. 4. Finding the supporting ontology object.

Each object in the ontology has an “IsATerm” property of Boolean type. The degree of proximity of input words to the terms of domain according to the Thesaurus algorithm is calculated by the following equation:

$$k_{Ont} = \frac{k_t}{c + 1}, \quad (2)$$

where  $k_t$  is the result of the first step of the analysis,  $c$  is the number of relations between the supporting ontology object and the nearest object with the true “IsATerm” value.

**Internal linkage algorithm.** The developed metrics allows extracting terminology by not only defining the termhood of single words but also comparing the terms from the text with ontology objects and lemmas combinations of those objects, using Radd relations. The Internal linkage algorithm is the implementation of the following one.

$$t_1 + R_1 + t_2 + R_2 + \dots + R_m + t_n, \quad (3)$$

where  $R_i \in R_{add}$ ,  $t_j \in T$ ,  $R_{add}$  is a set of relations that allow expanding the set of objects of the described domain through a combination of related objects lemmas, for example, properties “IsRelatedWith” and “IsAPartOf”.

Thus, extracted terms that are part of other terms consisting of more words are not considered as terms in order to avoid redundancy.

#### 5. Experiments

The text volume of about 62000 words from “LAN Administration” PrA was analyzed to assess the accuracy of the term extraction. OWL-ontology consisted of 261 classes and 46 relations.

Precision (P), Recall (R) and  $F_1$  measures were used to assess the effectiveness of the algorithms for each category of tokens. Experiments on term extraction using the most frequently applied statistical methods: Frequency, TF\*IDF, C-Value were also carried out. Results are presented in Table 2.

Thus, statistical methods showed significantly better results when retrieving one term tokens. The internal linkage algorithm first extracts terms related to existing knowledge base terms.

The internal linkage algorithm extracts less wrong terms in case of two and three term tokens. Statistical methods are more focused on the frequency of occurrences of phrases, regardless of the reference to the PrA features and can extract general scientific terms and terms from other problem areas. Moreover, statistical methods are more focused on the frequency of tokens without reference to the PrA and can extract general scientific terms and terms of other problem areas.

Table 2. Term extraction using statistical and syntagmatic methods.

Amount of words	Terms	Candidates	Right	P	R	F <sub>1</sub>
<b>Internal linkage algorithm</b>						
1	294	168	134	0.80	0.46	0.58
2	631	431	372	0.86	0.59	0.70
3	361	370	327	0.88	0.91	0.89
<b>Frequency</b>						
1	294	134	123	0.92	0.42	0.58
2	631	469	347	0.74	0.55	0.63
3	361	334	267	0.80	0.74	0.77
<b>TF*IDF</b>						
1	294	147	138	0.94	0.47	0.63
2	631	456	328	0.72	0.52	0.60
3	361	277	166	0.60	0.46	0.52
<b>C-Value</b>						
1	294	120	112	0.93	0.38	0.54
2	631	789	316	0.40	0.50	0.44
3	361	295	162	0.55	0.45	0.50

## 6. Conclusion

The use of mathematical and statistical approaches to the building of domain ontologies by extracting knowledge from text documents does not take into account morphological, semantic, and syntagmatic features used in the text of linguistic forms. The methods of syntagmatic analysis allows:

- to reduce all synonyms for the same concept;
- to include polysemous words for different concepts;
- to use the connections between the concepts and the appropriate terms to generate a new ontology entities.

Thus, the experimental results suggest a high efficiency of the methods described in the article. The methods were developed by combining linguistic algorithms of terminology extraction from large text corpora in the process of syntagmatic analysis and extracting the core of the ontology from external wiki-resources.

## Acknowledgements

This paper has been approved within the framework of the Federal Targeted Programme for Research and Development in Priority Areas of Development of the Russian Scientific and Technological Complex for 2014-2020, Government Contract No. 14.607.21.0164 on the subject "Development of architecture, methods and models in order to build software and hardware complex for semantic analysis of semi-structured information resources on the Russian element base" (Application Code "2016-14-579-0009-0687").

## References

- [1] Bova VV, Kureichik VV, Nuzhnov EV. Problems of representation of knowledge in integrated systems of support of administrative decisions. *News of SFedU* 2010; 108(7): 107–113.
- [2] Gavrilova TA. Ontological approach to knowledge management in the development of corporate information systems. *Artificial Intelligence News* 2003; 2(56): 24–29.
- [3] Vagin VN, Mikhailov IS. Development of the method of integration of information systems based on metamodeling and ontology of the subject domain. *Software Products And Systems* 2008; 1: 22–26.
- [4] Gribova VV, Kleshev AS. Managing the design and implementation of the user interface based on the ontology. *Management* 2006; 2: 58–62.
- [5] Zagorulko YuA. Construction of scientific knowledge portals based on ontology. *Computational Technologies* 2007; 12: 169–177.
- [6] Kleshchev AS. The role of ontology in programming. Part 1. Analytics. *Information Technologies* 2008; 10: 42–46.
- [7] Smirnov SV. Ontological modeling in situational management. *Ontology of Design* 2012; 2(4): 16–24.
- [8] Golenkov VV, Guliakina NA. Semantic technology of component design of knowledge-driven systems. *Fifth International Scientific and Technical Conference "OSTIS"*. Minsk, 2015: 57–78.
- [9] Namestnikov AM, Filippov AA. Implementation of the clustering system for conceptual indexes of project documents. *Automation of management processes* 2011; 3(25): 46–50.
- [10] Namestnikov AM, Filippov AA, Avvakumova VS. An ontology based model of technical documentation fuzzy structuring. *CEUR Workshop Proceedings, SCAKD* 2016; 1687: 63–74.
- [11] Shestakov VK. Development and maintenance of information systems based on ontology and Wiki-technologies. *13-th all-Russian Scientific Conference "RCDL-2011"*. Voronezh, 2011: 299–306.
- [12] Hepp M, Bachlechner D, Siorpaes K. Harvesting Wiki Consensus – Using Wikipedia Entries as Ontology Elements. *Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics. Annual European Semantic Web Conference (ESWC), 2006*: 124–138.
- [13] Subkhangulov RA. Ontologically-oriented method of searching for project documents. *Automation of management processes* 2012; 4(30): 83–89.
- [14] Konstantinova NS, Mitrofanova OA. Ontologies as a knowledge storage system. Portal "Information and Communication Technologies in Education". URL: <http://www.ict.edu.ru/ft/005706/68352e2-st08.pdf> (21.03.2017).
- [15] Sokirko AV. Semantic dictionaries in automatic processing: issertation for the degree of candidate of technical sciences. State Committee of the Russian Federation for Higher Education Russian State Humanitarian University. Moscow, 2001.

- [16] Boyarskiy KK, Kanevskiy YeA. Semantico-syntactic parser Semsin, Scientific and Technical Herald of Information Technologies. Mechanics and Optics 2015; 5: 869–876.
- [17] Artemov MA, Vladimirov AN, Seleznev KYe. Survey of natural text analysis systems in Russian. Scientific journal Bulletin of Voronezh State University. URL: <http://www.vestnik.vsu.ru/pdf/analiz/2013/02/2013-02-31.pdf> (22.02.2017).
- [18] Automatic text processing. Automatic word processing. URL: <http://aot.ru>(22.02.2017).
- [19] Yarushkina N, Moshkin V, Klein V, Andreev I, Beksaeva E. Hybridization of Fuzzy Inference and Self-learning Fuzzy Ontology Based Semantic Data Analysis. Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16), 2016: 277–285.

**Table of Contents**  
Mathematical Modeling

1. Modeling and coordinated control for the production and economic system E.V. Orlova.....	1-6
DOI: 10.18287/1613-0073-2017-1904-1-6	
2. One approach to control of a neural network with variable signal conductivity A. Olshansky, A. Ignatenkov.....	7-13
DOI: 10.18287/1613-0073-2017-1904-7-13	
3. A discrete phase problem in reconstruction of signals in space-rocket hardware A.A. Kuleshova, E.A. Shchelokov.....	14-22
DOI: 10.18287/1613-0073-2017-1904-14-22	
4. Mathematical models for forecast of geometrical parameters of assembly units V.A. Pechenin, M.A. Bolotov, N.V. Ruzanov.....	23-28
DOI: 10.18287/1613-0073-2017-1904-23-28	
5. Probability-theoretical model for product assembly parameters assessment N.V. Ruzanov, M.A. Bolotov, V.A. Pechenin.....	29-34
DOI: 10.18287/1613-0073-2017-1904-29-34	
6. About the attractor-repeller points during the descent of an asymmetric spacecraft in the atmosphere V.V. Lyubimov, V.S. Lashin.....	35-39
DOI: 10.18287/1613-0073-2017-1904-35-39	
7. Control of a one rigid-link manipulator in the case of non-smooth periodic trajectory N. Aksenova, V. Sobolev.....	40-42
DOI: 10.18287/1613-0073-2017-1904-40-42	
8. On stabilizability of the manifold of steady states in a model of the spread of a mutating viruses Ju. Ermoshkina.....	43-48
DOI: 10.18287/1613-0073-2017-1904-43-48	
9. Conditions for the loss of stability of equilibrium manifold in satellite model E. Shchepakina, V. Sobolev.....	49-51
DOI: 10.18287/1613-0073-2017-1904-49-51	
10. Viral evolution model with several time scales A.A. Archibasov.....	52-56
DOI: 10.18287/1613-0073-2017-1904-52-56	
11. Generalized model of pulse process for dynamic analysis of Sylov's fuzzy cognitive maps R.A. Isaev, A.G. Podvesovskii.....	57-63
DOI: 10.18287/1613-0073-2017-1904-57-63	
12. The method of augmented regularized normal equations for systems with sparse matrices S.Y. Gogoleva.....	64-66
DOI: 10.18287/1613-0073-2017-1904-64-66	
13. On possibilities for studying the problem of human society's evolution using simple mathematical models L.G. Teklina.....	67-71
DOI: 10.18287/1613-0073-2017-1904-67-71	
14. Mathematical model of power characteristics of the diagnostic fluorimeter V.N. Grishanov, V.S. Kulikov, K.V. Cherepanov.....	72-77
DOI: 10.18287/1613-0073-2017-1904-72-77	
15. Modeling and investigating the stability of a solution to the inverse problem of signal separation V.A. Zasov, Ye.N. Nikonorov.....	78-84
DOI: 10.18287/1613-0073-2017-1904-78-84	
16. Modeling and analysis of motion of a spacecraft with a tether aerodynamic stabilizer D. Elenev, Y. Zabolotnov.....	85-88
DOI: 10.18287/1613-0073-2017-1904-85-88	

17. Development of algorithms for diagnosing forms of lichen planus and predicting of the disease's course O.V. Serikova, V.N. Kalaev, N.A. Soboleva.....	89-92
DOI: 10.18287/1613-0073-2017-1904-89-92	
18. Calculation of the electrostatic field distribution formed by the generator of the off-electrode plasma M.A. Markushin, V.A. Kolpakov, S.V. Krichevskiy.....	93-99
DOI: 10.18287/1613-0073-2017-1904-93-99	
19. Approaches to the optimization of the placement of service-oriented cloud applications in the software-defined infrastructure of the virtual data center I. Bolodurina, D. Parfenov, K. Haenssgen.....	100-107
DOI: 10.18287/1613-0073-2017-1904-100-107	
20. The elaboration of numerical simulation error light pulse propagation in a waveguide of circular cross-section A.A. Degtuarev, A.V. Kukleva.....	108-112
DOI: 10.18287/1613-0073-2017-1904-108-112	
21. The calculation of the spatial spectrum of multidimensional fractals using the Fast Fourier transform O.A. Mossoulina.....	113-116
DOI: 10.18287/1613-0073-2017-1904-113-116	
22. Study of a singularly perturbed tuberculosis model E. Tropkina, E. Shchepakina.....	117-123
DOI: 10.18287/1613-0073-2017-1904-117-123	
23. Forecasting models generation of the electronic means quality R.O. Mishanov, S.V. Tyulevin, M.N. Piganov, E.S. Erantseva.....	124-129
DOI: 10.18287/1613-0073-2017-1904-124-129	
24. About scarce resources allocation in conditions of incomplete information N.L. Dodonova, O.A. Kuznetsova.....	130-134
DOI: 10.18287/1613-0073-2017-1904-130-134	
25. A model of milling process based on Morlet wavelets decomposition of vibroacoustic signals A.I. Khaymovich, S.A. Prokhorov, A.A. Stolbova, A.I. Kondratyev.....	135-140
DOI: 10.18287/1613-0073-2017-1904-135-140	
26. Intermediate asymptotic behavior of the stress and damage fields in the vicinity of the mixed-mode crack tip under creep regime L. Stepanova, E. Mironova.....	141-150
DOI: 10.18287/1613-0073-2017-1904-141-150	
27. Calculation of critical conditions for the filtration combustion model O. Vidilina, E. Shchepakina.....	151-157
DOI: 10.18287/1613-0073-2017-1904-151-157	
28. Mathematical modeling radio tomographic ionospheric parameters reconstruction via nanosatellites constellation for conditions of incomplete source data O.V. Phylonin, P.N. Nikolaev.....	158-167
DOI: 10.18287/1613-0073-2017-1904-158-167	
29. Modeling control over large space structure on geostationary orbit V.V. Salmin, A.S. Chetverikov, K.V. Peresypkin, I.S. Tkachenko.....	168-173
DOI: 10.18287/1613-0073-2017-1904-168-173	
30. On the method of step evaluation in construction descriptive models T.E. Rodionova, G.R. Kadyrova.....	174-177
DOI: 10.18287/1613-0073-2017-1904-174-177	
31. Server hardware resources optimization for virtual desktop infrastructure implementation K. Makoviy, D. Proskurin, Yu. Khitskova, Ya. Metelkin.....	178-183
DOI: 10.18287/1613-0073-2017-1904-178-183	
32. On delayed loss of stability in one mechanical problem E.V. Shchetinina.....	184-186
DOI: 10.18287/1613-0073-2017-1904-184-186	

33. Characteristics comparison of DTN networks routing protocols using hybrid model of nodes' mobility A.A. Tsarev, A.Yu. Privalov.....	187-190
DOI: 10.18287/1613-0073-2017-1904-187-190	
34. Two-stage approach to real-time assignment of Web Studio customer support tasks S. Begenova, T. Avdeenko.....	191-199
DOI: 10.18287/1613-0073-2017-1904-191-199	
35. Digital photoelasticity for calculating coefficients of the Williams series expansion in plate with two collinear cracks under mixed mode loading L.V. Stepanova, V.S. Dolgikh, V.A. Turkova.....	200-208
DOI: 10.18287/1613-0073-2017-1904-200-208	
36. Methods of bipolar microcircuits learning experiment S.V. Tyulevin, M.N. Piganov, E.S. Erantseva.....	209-213
DOI: 10.18287/1613-0073-2017-1904-209-213	
37. Frames and subspaces for phaseless reconstruction S.Ya. Novikov, M.E. Fedina.....	214-222
DOI: 10.18287/1613-0073-2017-1904-214-222	
38. Cellular automata-based model of group motion of agents with memory and related continuous model A.V. Kuznetsov.....	223-231
DOI: 10.18287/1613-0073-2017-1904-223-231	
39. Development of software system for analysis and optimization of taxi services efficiency by statistical modeling methods P. Azanov, A. Danilov, N. Andriyanov.....	232-238
DOI: 10.18287/1613-0073-2017-1904-232-238	
40. Detuning and dipole-dipole interaction effects on the entanglement of two qubits interacting with quantum fields of resonators E.K. Bashkirov.....	239-248
DOI: 10.18287/1613-0073-2017-1904-239-248	
41. Reduction of flexible joint manipulator mathematical model O.V. Vidilina, N.V. Voropaeva.....	249-253
DOI: 10.18287/1613-0073-2017-1904-249-253	
42. Model for constructing an option's portfolio with a certain payoff function M.E. Fatyanova, M.E. Semenov.....	254-262
DOI: 10.18287/1613-0073-2017-1904-254-262	
43. Stochastic Non-Markovian Schroedinger equation for a three-level quantum system V. Semin, A. Pavelev.....	263-265
DOI: 10.18287/1613-0073-2017-1904-263-265	
44. Numerical simulations of the quantum systems dynamics in the path integral approach A. Biryukov, M. Shleenkov.....	266-273
DOI: 10.18287/1613-0073-2017-1904-266-273	
45. Mathematical modeling of incentive mechanisms in projects for the development of new production O.V. Pavlov.....	274-279
DOI: 10.18287/1613-0073-2017-1904-274-279	
46. Nonlinear eigenvalue problems in fracture mechanics: eigenspectra and eigenfunctions A.A. Peksheva, L.V. Stepanova.....	280-288
DOI: 10.18287/1613-0073-2017-1904-280-288	
47. Study of the chain transfer agent's effect on the butadiene-styrene copolymer's properties based on the Monte-Carlo method T. Mikhailova, E. Miftakhov, S. Mustafina.....	289-292
DOI: 10.18287/1613-0073-2017-1904-289-292	

48. Reconstruction of realistic three-dimensional models of biological objects from MR-images for the radiation therapy purposes A.V. Lebedeva, V.V. Mamontova, S.A. Nemnyugin, A.V. Komolkin.....	293-295
DOI: 10.18287/1613-0073-2017-1904-293-295	
49. Neural network prediction model of the pilots' errors A.N. Danilenko.....	296-299
DOI: 10.18287/1613-0073-2017-1904-296-299	
50. Development of methods and algorithms for the classification of neurodegenerative diseases of the brain using MRI images O. Vasilchuk, A. Fedorov.....	300-305
DOI: 10.18287/1613-0073-2017-1904-300-305	
51. Use of graph-based and algebraic models in lifecycle of real-time flight control software A. Tyugashev.....	306-311
DOI: 10.18287/1613-0073-2017-1904-306-311	



# Preface

Vladimir Sobolev<sup>1</sup>, Dmitry Savelyev<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

The International Workshop “Mathematical Modeling” was organized and hosted by the Samara University and IPSI RAS – affiliate of the FSRC “Crystallography and Photonics” in the framework of the III International Conference and Youth School “Information Technology and Nanotechnology (ITNT-2017)” in the faculty of Information Technology at the Samara University, Samara, Russia on 25-27 April 2017 (<http://ru.itnt-conf.org/itnt17ru/>).

The work of this workshop is reported in this volume. The focus was on information technology and mathematical modeling across the sciences, a recurring theme of past three workshops reinvigorated recently by new collaborations of engineers and mathematicians. Participants repeatedly remarked on useful interactions with scientists from different disciplines. The Symposium was attended by experts working across a wide range of mathematical modeling fields from different countries, including Great Britain and USA, and provided a useful forum for them to share and exchange their work and ideas.

The goal of the ITNT-2017 Conference was to discuss problems of fundamental and applied research in information technology and nanotechnology, including but not limited to:

- Computer Optics;
- Diffractive Nanophotonics;
- Image Processing;
- High-performance Computing;
- Computer Vision;
- Mathematical Modeling;
- Data Science.

Scientists from Austria, Belarus, Bulgaria, Denmark, Germany, Great Britain, India, Iraq, Mexico, Moldova, Russia, Spain, USA, and Finland presented over 330 reports at the ITNT-2017 Conference.

These Proceedings contain both invited papers and contributed presentations, part of which was included to the special issue of the *Procedia Engineering* (Elsevier BV). As usual, topics ranged from theoretical foundation of mathematical modeling to applicably inspired problems and purely methodological advances. We hope that readers will benefit from specialized results as well as profit from exposure to new algorithms, methods of analysis, and conceptual developments.

## Guest Editors

- Sergei Sazhin, University of Brighton, UK;
- Elena Shchepakina, Samara National Research University, Samara, Russia;
- Vladimir Sobolev, Samara National Research University, Samara, Russia;
- Denis Kudryashov, Samara National Research University, Samara, Russia.

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov, Samara National Research University, Russia

## Vice-chairs

- Vladimir Bogatyrev, Samara National Research University, Russia
- Nikolay Kazanskiy, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Russia
- Eduard Kolomiets, Samara National Research University, Russia
- Alexander Kupriyanov, Samara National Research University, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Russia

# Modeling and coordinated control for the production and economic system

E.V. Orlova<sup>1</sup>

<sup>1</sup>Ufa State Aviation Technical University, K.Marx,12, 450000, Ufa, Russia

---

## Abstract

Problems of simulation and control in production and economic system (PES) under condition of economy's modernization are discussed in the paper. The developed control system considers the coordination of three systems – PES, market and taxation institutional system. We take into account the PES features as dynamism, large number of parameters, nonlinearity, nonstationarity, strong interconnectivity and hierarchical structure. The aim of the developed models and coordinated mechanisms is to improve the decisions making effectiveness at the expense of a coherent management in PES.

*Keywords:* production and economic systems; modeling; coordinated control; coordination of interests; multiple criteria for decision making

---

## 1. Introduction

To improve the economy productivity and to create the sustainable innovation development system for enterprises of different industries, organizational and legal forms, sizes and spatial location it is necessary to develop mechanisms of multi-level strategic planning and management. The control system, based on these mechanisms should include strategic planning at different level – the level of enterprise, the level of market agents and the level of state. At the enterprise level the strategic planning system must be unified in functional and management verticals. At the market agent level the development and implementation of strategic plans is to take into account the interests of all enterprise. At the state level, fiscal policy should be formed to be the most active component of economic and industrial policy, focused on the development of the capacity of enterprises as productive and economic system (PES) and its modernization.

Therefore the problem of modeling and controlling tools creating for the PES operation under a systematic modernization of the economy, taking into account the interests of agents matching the most PES, market and institutional environments is of great importance. The solution of this problem and developed application decision will increase the efficiency of controlling processes in PES.

The analysis of the investigations in the field of modeling and control in different organizational systems [1-8] is showed the following. Problems related to the economic-mathematical modeling and control of the processes of harmonization the economic interests of participants in production and economic, in market and fiscal processes are not solved. Previous studies do not deal with dynamic pricing for enterprises products in a competitive market, changing consumer preferences and behavioral strategies of competitors.

## 2. PES as a controlled object

Industrial and economic system is a complex organizational system combining production, sale and the resources reproduction. It is characterized by dynamism, a number of parameters, strong interconnectedness of parameters, hierarchical structure, the presence of the inverse of the material and information communications, nonlinearity and nonstationarity. The scheme of interaction the PES with macroeconomic systems – state, society is shown in Fig. 1.

State regulate the economy and the PES as a part of the economy, provide security, unity and territorial integrity of the national economy, get the possibility of economic development. The activity of enterprises is carried out under a number of restrictions: technological constraints (production function), financial constraints (the cost of production factors), demand (market size), competitor activity, government regulation, taxes and subsidies, ethical rules and norms (social norms of business), time.

Based on the above, the problem field can be represented as a set of interrelated issues. These questions arise to the process of interaction between the PES as the main system of the economy, producing products to the market as the sphere of activity of agents that influence the economic decisions of PES, as well as to the institutional environment of the state, which determines the rules of economic activity of the PES and a fiscal adjustment.

The following most important key issues can be identified as:

- a) ensuring the development of PES through tax incentives;
- b) sustainable functioning of the PES through the balancing of resources and results of operations;
- c) coordination of economic interests of PES;
- d) increasing the management efficiency through the PES agreed with external agents behavior in industrial market.

There are a number of contradictions accompanying production and economic activity: PES - tax system, PES - competitors, PES - consumers. These contradictions include the following.

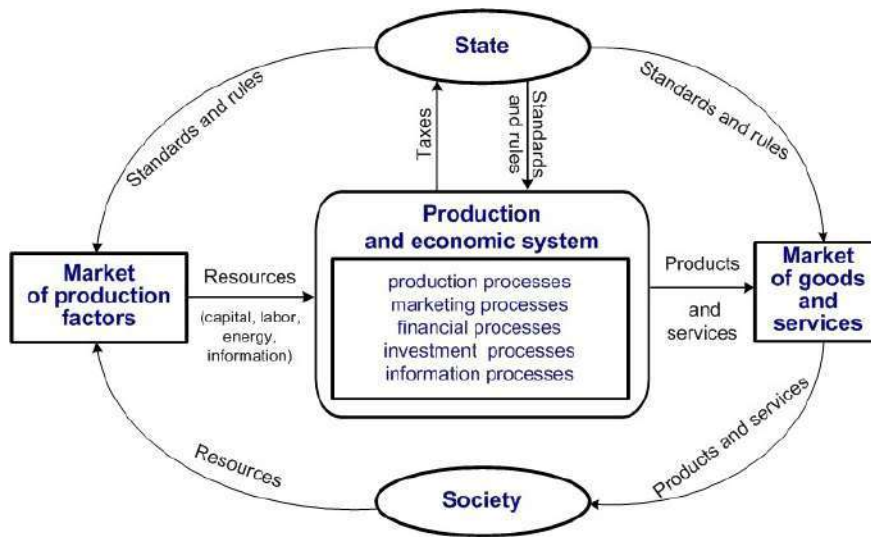


Fig. 1. Interaction of PES with macroeconomic systems.

1. The financial and economic stability of PES and its effectiveness depend on reducing tax payments and the stability of the economy connected with the growth of tax revenues into the budget. Smoothing of this contradiction of interests can be achieved by the harmonizing the tax burden on PES and the level of tax rates that provide both the PES development and required budget revenues.

2. When selling the product on the market PES aim to maximize its revenues at the expense of highest possible price in complete market. Consumers make their choice of those products which has minimum price with equal quality. Resolution of this contradiction is possible due to adaptive pricing that ensures the coordination of PES utility functions, competitors utility functions and consumers preferences.

3. When implementing strategic goals aimed at long-term growth and development PES should monitor solvency and liquidity in order to avoid risks associated with a decrease in the level of sustainability and solvency in the short term. This requires a set of measures aimed at managing financial resources and cash flows and providing coordinated management at the strategic and operational levels of the PES system management.

4. Management of the PES stability can be realized by balancing: resource flows, output volumes and products price by the criteria of profitability and profit.

These contradictions are formed with the synergetic interaction of economic systems - the enterprise as a PES, the market system and the tax system and can be resolved through the development of models, methods and mechanisms for managing the interaction of these systems on a single methodological platform. In order to solve the problem of reconciling the multidirectional interests of subsystems - PES, market and tax systems, it is necessary to create a decision support toolkit under uncertainty to ensure a PES and economy efficiency increasing. The problem of PES management features are:

- 1) complexity of control object;
- 2) synergistic impact of factors and uncertainty on the process of PES functioning and development;
- 3) necessity of PES management system efficiency increasing to ensure PES sustainability;
- 4) necessity to harmonize the multidirectional interests of interacting systems to ensure economic growth.

Therefore, the problem of conflicting interests harmonizing of PES, market and tax systems at different management levels, developing tools for supporting decision-making under conditions of uncertainty is urgent. The management system should include management at the PES level, at market agents level and at the state level.

### 3. Conceptual basics for the PES modeling and control

Models for PES process control based on complex interactions of diverse subsystems, ensuring consistency of economic interests of participants of the economic, financial and market processes. The control system structure is defined in the form of three blocks: 1 – block for providing an effective enterprise resource management system, aimed at harmonizing the strategic and operational levels of management (project management, processes); 2 – block for adaptation of market management mechanism that supports the interests of consistency of producers and consumers (control pricing processes); 2 – block for tax burden and tax rates regulation as an instrument of fiscal policy, aimed at improving economic growth in general (management of the institutional environment). Solution of these problems is based on simulation of a system having a hierarchical structure which includes the components listed below.

- a) simulation of the tax burden and tax rates, taking into account conflicting interests of industrial and economic systems (tax subjects) and the economy as a tool of fiscal policy, aimed to the development of the PES;
- b) simulation of market interactions of producers and consumers. This section is important because it leads to the formation of market prices, which are among the most important characteristics of the PES effectiveness;

- c) simulation of resource component for production and economic system. At this level, it is considered the conversion of resources into results and the aim is to ensure the most efficient allocation of resources for production and to ensure the effective development of the PES.

The coordinated control mechanism for the PES consists of inter-related technologies: the balanced efficiency; the financial planning; the prices control; the technology of synthesis of optimal tax rates. First two technologies are united in subsystem for the PES resources control. Functional diagram of coordinated control mechanism for the PES is shown in Fig. 2.

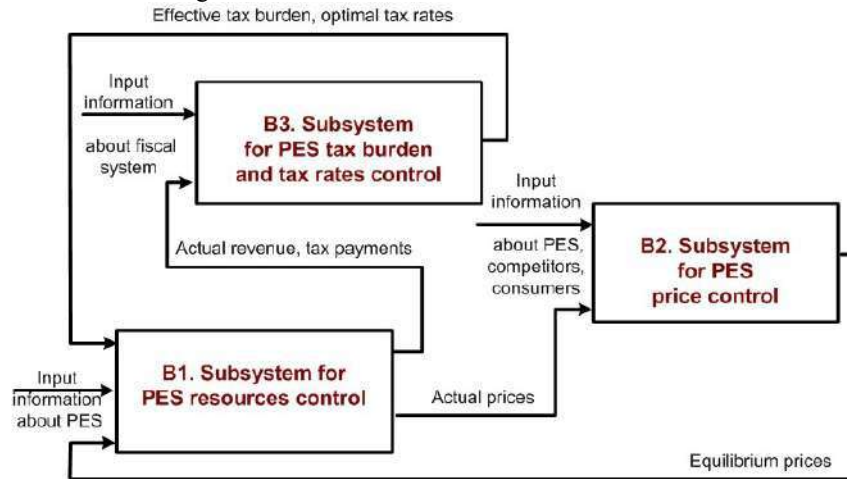


Fig. 2. Functional model for PES coordinated control mechanism.

The hierarchical simulation system is constructed as follows. First level is modeling of optimal tax system [9], includes the model for determining the effective tax burden and the model for determining the optimal level of tax rates on taxes and tax homogeneous groups of subjects [10, 13]. Simulation results at this level area are allowed tax burden in groups of similar taxation objects and the optimal tax rates that satisfy the interests of the taxation objects and the economy as a whole.

The second simulation level is the dynamic model of market pricing, designed for the PES control in conditions of nonstationarity of environment parameters and dependence spheres of production and consumption on the basis of the adaptive pricing mechanism. Prices level formation and changing are coordinated both with the strategic objectives of the manufacturers and the changing preferences of consumers [11, 12]. The result of the modeling of market processes are Nash equilibrium prices for manufactured products that achieve the maximum efficiency for PES.

The third simulation level based on production and economic system control models, designed for the optimal combinations of production resources within the constraints generated by the first and second levels of modeling [13, 14]. Modeling results are the resources costs, production quantity and prices for production and economic system.

#### 4. Models for PES control

The model of economic efficiency of production is formed as the ratio of profit and production costs:

$$\alpha = \frac{\Pi}{C}, \quad (1)$$

where the profit is

$$\Pi = pq - (C_f + q \cdot C_v). \quad (2)$$

Total production cost is defined as the sum of variable and fixed costs

$$C = C_f + q \cdot C_v, \quad (3)$$

where  $\alpha$  is the profitability of goods production in  $q$  quantity,  $\Pi$  is the gross profit generated on the sale of goods in  $q$  quantity and  $p$  price,  $C$  is the total cost, including a constant part  $C_f$  and a variable part  $qC_v$ . The profit tax payable to the budget is defined as

$$T_{PT} = t_{PT} \cdot \Pi, \quad (4)$$

where  $T_{PT}$  is the profit tax,  $t_{PT}$  is the profit tax rate.

The basis of tax burden simulation is the Laffer assumption about the nonlinear connection of output  $X$  and the level of tax burden  $\theta$ . For each group of similar taxation objects we design the dependences as the follow

$$\theta = T/X, \quad (5)$$

where  $\theta$  is the tax burden,  $T$  are the tax revenues.

We assume that the production also has non-linear relation with the tax burden. The production function is approximated by a quadratic polynomial:

$$X(\theta) = a\theta^2 + b\theta \quad (6)$$

and the tax function  $T(\theta)$  has the form:

$$T(\theta) = a\theta^3 + b\theta^2, \quad (7)$$

where  $a$  and  $b$  are function parameters. Identification of production function  $X(\theta)$  and tax function  $T(\theta)$  allow to find the first and second Laffer points in which the production and tax functions has their maximum respectively.

Harmonization of the operational financial planning system and the strategic financial planning system is provided due to coordination of special indicators of operational and strategic control levels. At the operational enterprise level it is ensured the products competitiveness, the overall status of PES and its financial and operating efficiency, at the strategic control level it is conducted the enterprise investment attractiveness, the growth of its value in the long term. Interconnection of these indicators is based on a multifactor model

$$ROE = \frac{\Pi - T_{PT}}{E} = \frac{\Pi}{A} \cdot \frac{A}{E} \cdot \frac{NI}{EBIT} = ROA \cdot LR \cdot B, \tag{8}$$

where  $ROE$  is the return on equity;  $NI$  is the net income;  $E$  is the equity capital;  $EBIT$  are the earnings before interest on loans and income tax;  $A$  is assets;  $ROA$  is return on assets;  $LR$  is the coefficient that determines the effect of financial leverage;  $B$  is the coefficient reflecting the decrease in profitability of the enterprise in the payment of interest on the capital employed and tax deductions.

The model of competitive interaction of the enterprises (in the case of a duopoly) is represented as mapping:

$$\begin{cases} p_1(t+1) = p_1(t) + k_1 \frac{(-p_1^2(t)p_2(t) + 2c_1p_1(t)p_2(t) + c_1p_2^2(t))}{(p_1^2(t) + p_1(t)p_2(t))^2}, \\ p_2(t+1) = p_2(t) + k_2 \frac{(-p_1(t)p_2^2(t) + 2c_2p_1(t)p_2(t) + c_2p_1^2(t))}{(p_2^2(t) + p_1(t)p_2(t))^2}. \end{cases} \tag{9}$$

where  $p_1(t), p_2(t)$  are the prices of products of the first and second firms, taken at discrete intervals of time  $t$ ; second terms in both equations show how the change in prices in period  $t$ , and how this change will affect the price in the next period. Parameters  $k_1$  and  $k_2$  represents the increase in prices due to changes in the pricing policy. Variables  $c_1$  and  $c_2$  represents production cost of the first and second firms respectively.

### 5. Numerical results

We consider the computational experiment implemented the designed harmonized mechanism for the tax burden and tax rates. Modeling is based on the statistical data for several years about taxable bases for individual taxes, tax revenue for five classes of PES - large enterprises, providing about 50 % of tax revenue into the region budget. Further we give modeling results for the aggregate tax burden and the tax becoming the first group of taxpayers. The first taxpayers can be described as the largest taxpayer, which has the following structure of taxable bases: the share of value-added tax is 0.576; profit taxes - 0,125; the unified social tax - 0,098; property tax - 0.08; other taxes - 0,121.

Production and tax functions, as well as their extremes are:

$$X_1(\theta) = -141.5 \cdot 10^7 \theta^2 + 400.1 \cdot 10^6 \theta, T_1(\theta) = -141.5 \cdot 10^7 \theta^3 + 400.1 \cdot 10^6 \theta^2, \theta^* = 0.13, \theta^{**} = 0.19. \tag{10}$$

Analysis of the actual total tax burden for the first group of taxation objects shows that in periods  $t$  and  $t-1$ , its value is more than two points Laffer  $\theta^*$  and  $\theta^{**}$  and is equal to 0.23 and 0.25 respectively. The insensitive area for the tax burden is from 0 to 0.09 ( $\theta_h$ ), Fig. 3.

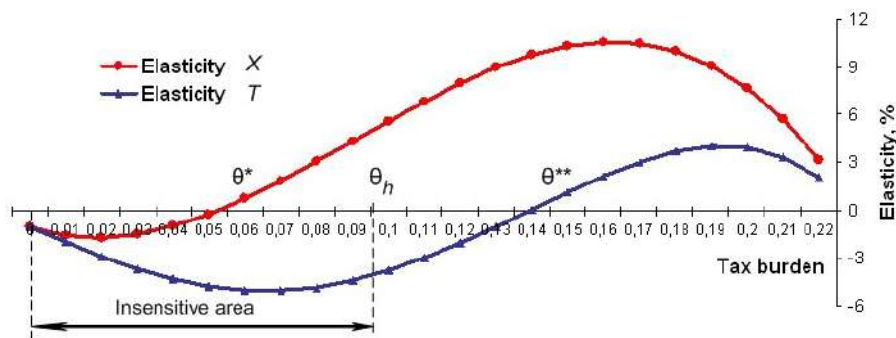


Fig. 3. Elasticity function for the production and tax revenue.

Detailed analysis of the production and tax functions for the first group of enterprises showed that the current tax burden is such that there is in the third zone. This corresponds to a situation in which the tax burden is the right of both the Laffer points. This means that fiscal policy stimulates the fall of the production function, while dissatisfaction fiscal interests. Therefore it is a great necessarily for reduce the overall tax burden, which will increase the value of production and tax functions simultaneously, table 1.

Consider the several cases to change (decrease) in the overall tax burden for first group objects of the. The most significant increase in the production function is achieved at the tax burden level of 13-14 %, but further reduction in the tax rate will not give further effect in the tax revenues growth. Elasticity's analysis of the production and tax functions showed that the area of insensitivity can be determined as  $[0; 0.09]$ , since this segment is the growth of the tax burden leads to a greater reduction in tax

revenue than that achieved for a given load in output growth. Therefore the recommended decision is to reduce the overall tax burden up to 14 % or 9 %.

Table 1. Growth rates for production and tax functions.

Tax burden changing $\Delta\theta$	New tax burden $\theta$	Growth rate for function $X$	Growth rate for function $T$
-	0,23	-	-
0,01	0,22	0,138	0,089
0,02	0,21	0,260	0,150
0,03	0,2	0,365	0,187
0,04	0,19	0,454	0,201
0,05	0,18	0,526	0,194
0,06	0,17	0,581	0,169
0,07	0,16	0,620	0,127
0,08	0,15	0,643	0,071
0,09	0,14	0,649	0,004
0,1	0,13	0,639	-0,074
0,11	0,12	0,612	-0,159
0,12	0,11	0,568	-0,250
0,13	0,1	0,508	-0,344
0,14	0,09	0,432	-0,440
0,15	0,08	0,339	-0,534
0,16	0,07	0,229	-0,626
0,17	0,06	0,103	-0,712
0,18	0,05	-0,040	-0,791
0,19	0,04	-0,199	-0,861
0,2	0,03	-0,374	-0,918
0,21	0,02	-0,566	-0,962
0,22	0,01	-0,775	-0,990
0,23	0	-1,000	-1,000

Changing the overall tax burden is possible by changing tax rates. Moreover, the conceptual analysis of the types of taxes remains within the Laffer theory that maintains the unity of methodological research. Therefore, for qualitative and quantitative analysis of the need to build on each group of objects depending on tax bases of each taxes: value added, income, profit, property. For the first taxpayers group tax production functions for each type of tax as well as the extreme points of these functions, and the actual tax burden value are as follows (for value-added tax, for profit tax, for income tax, for property tax correspondingly):

$$\begin{aligned}
 X_{11}(\theta) &= -107.3 \cdot 10^6 \theta^2 + 548.5 \cdot 10^5 \theta, T_{11}(\theta) = -107.3 \cdot 10^6 \theta^3 + 548.5 \cdot 10^5 \theta^2, \theta^* = 0.28; \theta^{**} = 0.35, \theta = 0.37, \\
 X_{12}(\theta) &= -273.2 \cdot 10^8 \theta^2 + 715.8 \cdot 10^7 \theta, T_{12}(\theta) = -273.2 \cdot 10^8 \theta^3 + 715.8 \cdot 10^7 \theta^2, \theta^* = 0.23; \theta^{**} = 0.34, \theta = 0.15 \\
 X_{13}(\theta) &= -368.7 \cdot 10^6 \theta^2 + 479.3 \cdot 10^5 \theta, T_{13}(\theta) = -368.7 \cdot 10^6 \theta^3 + 479.3 \cdot 10^5 \theta^2, \theta^* = 0.07; \theta^{**} = 0.09, \theta = 0.08, \\
 X_{14}(\theta) &= -273.2 \cdot 10^8 \theta^2 + 715.8 \cdot 10^6 \theta, T_{14}(\theta) = -273.2 \cdot 10^8 \theta^3 + 715.8 \cdot 10^6 \theta^2, \theta^* = 0.009; \theta^{**} = 0.016, \theta = 0.02. \tag{11}
 \end{aligned}$$

Many tax rates combinations that implements the single changing in total tax burden can be represented in the form of the matrix. For the analyzed companies there have been determined the particular solution, and presented in the form of a matrix A. The elements of this matrix reflects the need to transform the specific tax rates in the form of an increase / decrease in implementing the first option - reducing the overall tax burden in 1% lead to the increase in the production function in 14 % and the growth of the tax function in 9 %:

$$A = \begin{pmatrix} -0.02 & 0.05 & -0.01 & -0.01 & 0 \\ 0.07 & 0.11 & 0.05 & 0.23 & 0 \\ 0.02 & 0.33 & -0.08 & -0.53 & 0 \end{pmatrix}. \tag{12}$$

It should be noted that the proposed change in tax rates implements only one of the possible options for fiscal policy, which allows to increase the efficiency of the fiscal system. Modeling of competitive tax system and finding the optimal tax rates on different PES groups is presented in [9].

**6. Conclusion**

The proposed approach for PES modeling and control is differ from similar ones in that takes into account its properties, complex nonlinear relationships between economic agents of internal and external environment. This allow to simulate the processes of PES functioning in conjunction with the economic agents, agreeing in control their conflicting goals and criteria.

The conceptual model for the PES control defines a single system-methodological position for control structure in the form of three blocks: creation the effective PES control system that ensures the coordination of strategic and operational management processes in the PES; formation the effective pricing system for the PES as a market mechanism for strategic cooperation with competitors and consumers; formation the tax burden and tax rates as an instrument of fiscal policy, aimed to stimulating the PES development and economic growth. Conceptual model is the basis of the methodology and aims to improve the decisions making effectiveness at the expense of a coherent decision-making in PES control system.

The proposed scheme of PES control system unites disparate economic and mathematical methods and models, reflects the heterogeneous properties of PES and provides a synthesis of efficient control algorithms. Developed decision support tools for the PES control in the form of control mechanisms, methods and models has been implemented as an integrated software package.

## References

- [1] Prangishvili IV. The systems approach and system-wide patterns. M.: SINTEG, 2000.
- [2] Burkov VN, Gubko MV, Korgin NA, Novikov DA. Theory of control of organizational systems and other management science organizations. Control Problems 2012; 4.; 2–10.
- [3] Novikov DA. Theory of control of organizational systems. M.: Fizmatlit, 2012.
- [4] Kleiner GB. The system platform for the development of the economy of modern economic theory. Problems of economy, 2013; 6.
- [5] Livshits VN. On Unsteady Russian transition economy. The Control Problems of theory and practice 2014; 2; 8–13.
- [6] Zang WB. Differential Equations, Bifurcations, and Chaos in Economics. NY: World Scientific Publishing Company, 2005; 512 p.
- [7] Evstigneeva LP, Evstigneev RN. Economics as a synergetic system. M.: Lenard, 2010; 272 p.
- [8] Makarova EA. Intellectual support management decision making in multi-sector macro-economic systems based on market relations on the basis of dynamic models. dis. ... PhD. in Eng. 05.13.10. Ufa, 2011; 389 p.
- [9] Orlova EV. Economic-mathematical tools for management in economic system under uncertainty. Ufa: UGATU, 2012; 172 p.
- [10] Orlova EV, Ismagilova LA. The tax system and the real sector: optimization of interests. Journal of Economic Regulation 2014; 2; 133–142.
- [11] Orlova EV. Model for Economic Interests Agreement in Duopoly's Making Pricing Decision. Computer Research and Modeling 2015; 6: 1309–1329.
- [12] Orlova EV. Concept for Industrial and Economic Systems Management Based on Criteria Coordination of Interested Agents. Program Engineering 2016; 2: 86–96.
- [13] Orlova EV. Tools for Management of Efficiency Tax System. Economic analysis: theory and practice 2013; 43: 59–70.
- [14] Orlova EV. Modeling and Decision Support for the Firms' Pricing Policy under a Chaotic Dynamic of Market Prices. CEUR-WS, Proceedings of the Workshop on Computer Modeling in Decision Making (CMDM 2016) 2016; 1726: 81–88.

# One approach to control of a neural network with variable signal conductivity

A. Olshansky<sup>1</sup>, A. Ignatenkov<sup>2</sup>

<sup>1</sup>Railway Signalling Institute JSCo, 27 bld. 1 Nizhegorodskaya str., 109029, Moscow, Russia

<sup>2</sup>Samara State Transport University, 2 V Svobody str., 443066, Samara, Russia

---

## Abstract

The article focuses on a new extent to synthesize a control strategy of learning of a special artificial neural network with variable signal conductivity and on features of neural networks with variable signal conductivity. The purpose of the research is to create a new control strategy based on analysis of the neural network error signal. Also the article contains an approach how to compute the control strategy. The main result is contained in the necessity to realize the control strategy on the basis of preliminary training of the neural network with specific techniques. Authors also suggest a technique of computing the crucial trajectory of the network's error decreasing. The trajectory may be computed using signal processing methods.

*Keywords:* neural networks; control; signal processing; preliminary training

---

## 1. Introduction

Artificial neural networks (ANN) are well known in application to transport planning, scheduling etc. Some papers are based on Hopfield's classical neural network with minimization of the energy function [1,2]. In [3] authors propose a hybrid scheme with the Hopfield's network and some heuristic methods to create timetable of CPU load. In [4] authors solve some transport problem by the multilayer perceptron with one hidden layer. All these works are not dedicated to creating the timetable with complex analysis of transport constraints. Besides all of these networks are not being trained with rigor methods. Moreover in [5,6] an approach to consider the neural network's teaching as an optimal control problem is presented. Hence in this paper authors are going to present an attempt to set the problem of rational control of the special neural network which solves rail scheduling problems.

## 2. The object of the study

The object of our analysis is a special neural network with variable signal conductivity described in [7]. Its topology is given in fig.1

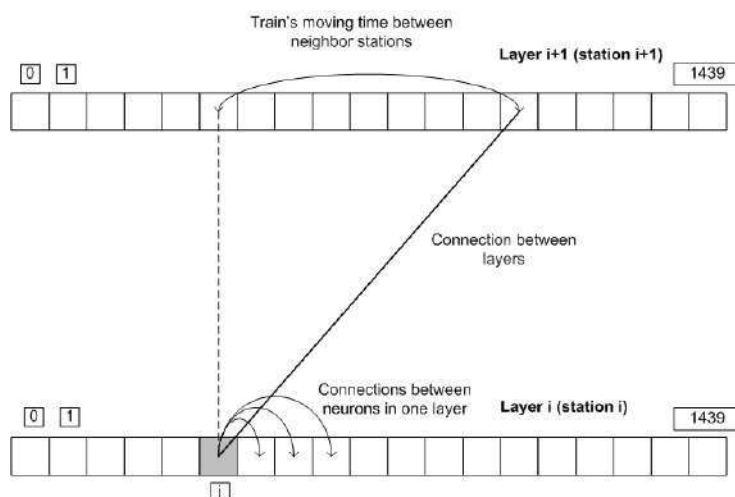


Fig.1. The architecture of the special neural network.

It is given a double-track railway section, consisting of several runs with the set times of the train moving in even and odd directions. There is an input vector  $X = \{x_0, x_1, \dots, x_{1439}\}$ , consisting of zeros and ones characterizing moments of train departure from the specified stations of the section. Also we know the desired vector of the train arrivals to the final station, which is the desired output value of the zero layer  $X^{(0)}$  of the ANN.

The number of layers is equal to the number of railway stations. Each layer has 1440 neurons, which is equal to the number of minutes in the period of twenty-four hours.



Each neuron of the  $i^{th}$  layer is connected to each neuron of the next layer (total number is 1440 links). In addition, each neuron is associated with several neurons on the left (i.e., with neurons with a smaller number) and on the right (with neurons with a larger number).

Each matrix of weights  $W$  between two layers with numbers  $i, i+1$  is a square matrix where the number of rows and columns is equal to 1440.

$$W_{i,i+1} = \begin{pmatrix} w_{0,0} & \dots & w_{0,1339} \\ \vdots & \ddots & \vdots \\ w_{1339,0} & \dots & w_{1339,1339} \end{pmatrix},$$

where  $w_{ij}$  is the weight value on the link connecting the neuron with the  $i^{th}$  layer number and the neuron with the  $j^{th}$  number of the adjacent layer.

Every neuron may be characterized by its state. Possible states of the neuron are: "active", when the input signal can be received at the input of the corresponding neuron, "sleep", when the value of the potential of the given neuron is zero, "off", when we cannot receive signals from the previous layer. The state "sleep" exists for even and odd directions.

Weights of constraints are initially specified randomly by real numbers from 0 to 0.1. Later they change during the process of the neural network's learning. The transit of the signal through the connections between the neurons of neighboring layers displays the process of traversing the train on a distance between the stations. Pay attention to minimum travel time between two stations (which is integer), all weights of links from neuron with number  $j$  from 0 to  $j+t$  (where  $t$  is minimal running time) are taken equal to minus infinity. These weights never change.

Calculation of the ANN output is performed by sequential calculation of the potentials of active neurons in adjacent layers along the signal path using the sigmoid activation function.

Link weights define the level of competition between neurons for the right to receive a signal (train) in the next layer (station). Thus, the transmitted signal from one layer of the neural network to the other can't violate the rule of minimum travel time.

The direct calculation of the network is performed as follows.

1. The vector  $X = \{x_0, x_1, \dots, x_{1439}\}$ , which is a sequence of zeros and ones, is being fed to the first layer. The ones mean that the corresponding minute is associated with the passage of the train.  $X^l$  means a vector of neurons' values from a layer with the number  $l$ .

2. For all links issuing from the layers with numbers  $n, n-1, \dots, 1$  the following holds:

$\forall k \in (0, 1, \dots, 1439)$  if  $X_k^{(l)} > 0$ , then  $\exists j \in (0, 1, \dots, 1439): w_{k,j} = \max\{w_{km}\}, X_j^{l-1} = 0$ , then we establish  $X_j^{l-1} := f(X_k^{(l)}, w_{k,j})$ ,  
where

$w_{km}$  is the weight at the connection between the neurons of the layers  $k$  and  $m$ ,  
 $k$  is the number of the neuron of the current layer,  
 $j$  is the number of the layer adjacent to the current.

The meaning of the condition described above is that: for each neuron with the number  $k$  of the current layer with a positive output value in the next layer, we search for such a neuron that the value of their connection weight is the maximum of all the neuron bonds with the number  $k$  with the next layer.

3. The activation function is  $f(x, w) = \begin{cases} 0, & \text{if } x = 0, \\ \frac{1}{1+e^{-x*w}}, & \text{if } x > 0 \end{cases}$  (1)

where

$f(x, w)$  is the value of the activation function,  
 $w$  is the weight value of the neuron with the maximum-weighted link, winning in the competition of neurons,  
 $x$  is the input to the neuron-winner.

4. With the output of the network  $Y$  we take the vector of values of the last layer  $X^{(0)}$ .

Moreover  $Y_d$  is the desired output of the network. It also consists of a sequence of zeros and ones whose meaning is identical to the sense of the input vector  $X$ .

We introduce the concept of "train number", which we denote by  $r$ . We say that the train passes through the station with the number  $l_1$  per minute  $k_1$  and the station with the number  $l_2$  per minute  $k_2$ , if the equality is fulfilled:  $r(X_{k_1}^{l_1}) = r(X_{k_2}^{l_2})$   $\forall r \forall l \exists! k: r(X_k^l) = r$ , i.e. all trains pass through each station at exactly one point-minute. By train numbers trains are tracked for moving through all stations. The train number determines the category of the train and current train classification for scheduling.

The neural network is being trained according to the following algorithm.

1. Calculation of the ANN error.

A network error is calculated using the formula:

$$E = \frac{\sum_b (k_d - k_Y)^2 + \sum_b (\tau * k'_d)^2}{b},$$
 (2)

where

$b$  is the number of trains that are required to be plotted in the schedule.

$r(Y_{k_Y}) = r(Y_{k_Y})$  is the number of the train for all trains that have reached the last layer,

$k'_d$  is the index numbers of the elements of the vector  $Y_d$ , for which the network signal did not reach the last layer,

$Y_{k_d}, Y_{k_Y}$  are the values of the elements of the target vector  $Y_d$  and the actual arrival vector  $Y_k$  respectively.

$\tau$  is the penalty coefficient for the train that has not reached the last station.

This type of error was introduced in order to emphasize the physical meaning of the network requirements (all trains must reach the layer with number 0).

2. Within the given number of learning epochs, while  $E > \Delta$ , where  $\Delta$  is the forward accuracy of the laying of trains, perform the following steps:

For all the layers with numbers  $l \in (0, 1, \dots, n-1)$ , for all neurons of the  $l^{\text{th}}$  layer with the number  $j \in (0, 1, \dots, 1439)$  it is set: if  $X_j^{(l)} > 0$ , then for the train identifier  $r(X_j^{(l)}) \exists ! i: r(X_j^{(l)}) = r(X_i^{(l+1)})$ . The meaning of this expression is searching in the layer  $(l+1)$  of the neuron with the number  $i$ , from which the train with the identifier  $r(X_j^{(l)})$  has come to the neuron  $j$  of the layer  $l$ .

We calculate the new weights of the  $l^{\text{th}}$  row of the matrix of  $W_{l+1,l}$  by recalculating values of the weights according to the follows:

- reducing the weights that are situated “on the left” from the maximum weight position according to the formula:

$$\forall m \in (0, 1, \dots, i-1): w_{i,m} = w_{i,m} - \eta \cdot x_i^l \cdot f'(x_i^{l+1})$$

- reducing the maximum weight by the formula:

$$w_{i,j} = w_{i,j} - \eta \cdot x_i^l \cdot f'(x_i^{l+1})$$

- increasing the weights that are situated “to the right” from the position of maximum weight by the formula:

$$\forall m \in (i+1, i+2, \dots, i+s): w_{i,m} = w_{i,m} + G \cdot \eta \cdot x_i^l \cdot f'(x_i^{l+1}) \cdot |w_{i,j} - w_{i,m}|$$

In the equations the following notations are accepted:

$f'$  is the derivative of the activation function (1),

$s = \sqrt{E}$  is an indicator characterizing the width of a segment within which there is a positive correction of the balance,

$G$  is the coefficient introduced for accelerated growth of weights “on the right” from the position of the maximum-weighted link,

$\eta$  is a speed of network training,

$x_i^l$  is the output of the neuron  $i$  of the layer  $l$ ,

$m$  is the position of the neuron in the layer relative to the value of the neuron with the maximum weight,

$w_{i,j}$  is the value of the maximum weight,

$w_{i,m}$  is the weight value for the neuron at position number  $m$ ,

$f'(x_i^{l+1})$  is the value of the derivative of the activation function in the subsequent layer  $(l+1)$  for the neuron numbered  $i$ ,

$x_i^l$  is the output of the neuron  $i$  of the layer  $l$ .

The physical meaning of learning is this: when the weights of connections “to the left” from the connection with the maximum weight are being decreased, the chance of the signal of the selected neuron to pass through the reduced connection became smaller at the next attempt (epoch). “On the right” of this connection, on the contrary, the values of the weights are being increased.

After each epoch (after a new calculation of the network values), the learning rate  $\eta$  increases by a value equal to  $\frac{\eta}{e}$ , i.e.

$$\eta(e) = \eta_{e-1} + \frac{\eta_e}{e}$$

where  $e$  is the number of epochs.

After selecting all the signal trajectories and printing them on the paper we'll get a variant of a train schedule created by the artificial neural network during its training process.

During its training process every ANN has its own error signal. The error signal is an important indicator of ANN's behavior. The further investigations are devoted to using the ANN error signal to improve its training process. Typically the dynamics of the error function (error signal) looks like in fig.2

### 3. Methods

In particular case when the error function could be described as sum of sinusoidal-based harmonics with different frequencies and amplitudes we may use the results obtained in [8]. In general case we are not sure in this signal error representation.

To analyze the behavior of the error function we plot its autocorrelation function (fig.3).

It gave us an assumption that it is possible to decompose the signal. The goal of this decomposition is to filter the main components of the error function. After filtering we should try to implicate the decomposition for a rational control scheme to train the network.

According to the useful practice in stochastic market signal processing we have successfully implicated LOESS techniques [9] to decompose the signal of the neural network error function. The analyzed signal, as we discovered, consists of three perceptible components: a trend part, a periodic signal and the irregular components.

The trend curve of the neural network error function provides guides for synthesis of the rational control.

### 4. Discussions about ways of control implementation

During the research authors have planned, realized and analyzed several series of computational experiments with variable key parameters of the neural network functioning including initial trainspeed, desired mean of the error, number of trains to

scheduling etc. It is set that we have stable and iterative character of error function view (fig.2), trend function view (fig.4). So we conclude that we should find rational control of the neural network based on this curves (fig.4).

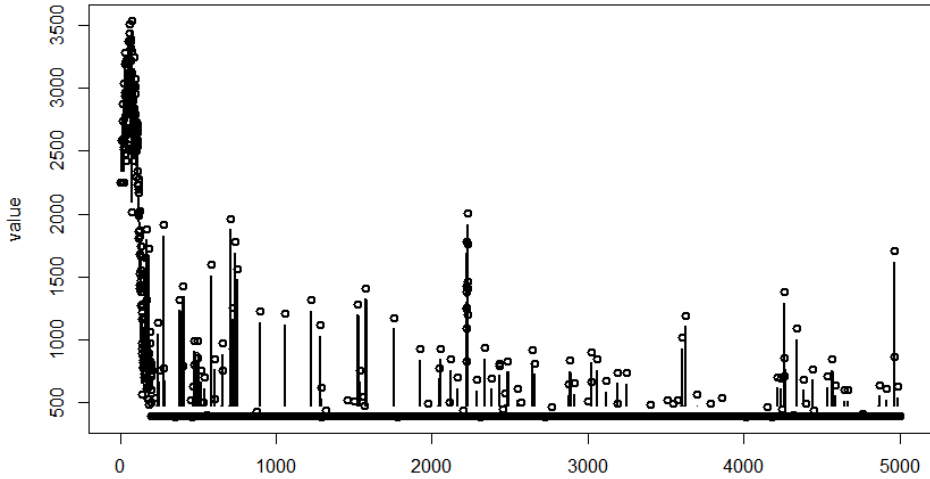


Fig.2. One example of error function.

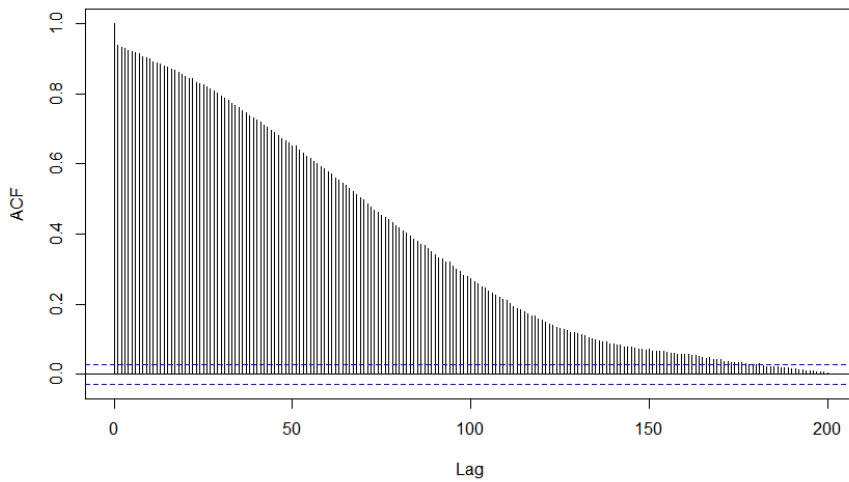


Fig.3. ACF of the error function.

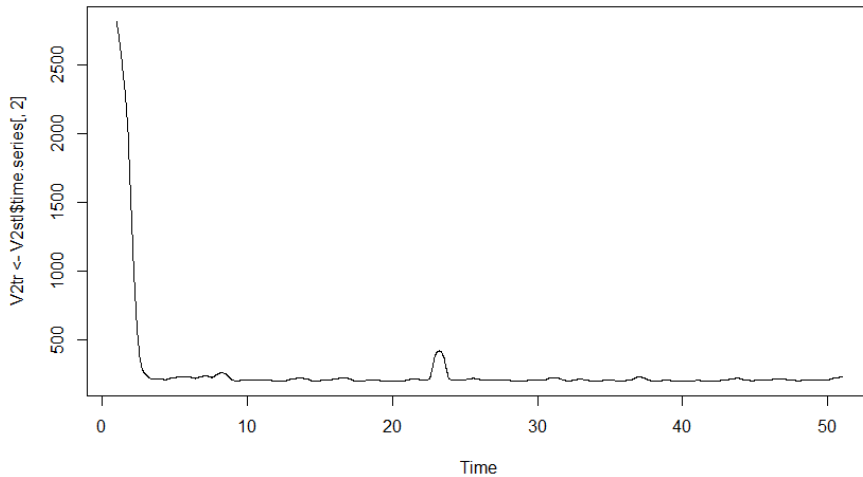


Fig.4. An example of STL-decomposition of the error function (trend).

In consideration of the enormous quantity of links between neurons it is impossible to solve the control problem in terms of dynamics of every neuron link (and the system of differential equations). So it is potentially useful to apply some special techniques to simplify and generalize control influence like:

- Pre-amplification of the bundle of links in the concrete areas of the neural network;
- Pre-dimintion of the bundle of links which are rather useless to desired trajectory of signal distribution (e.g. the links which may create too earlier or too later arrivals etc.);
- Simultaneously pre-amplification and pre-dimintion of the links;
- Mutation of links' weights in bounded area;

- Swap of weights between the links of selected pair of neurons.

Another way to invent the control strategy is implementation of inverse general neural network control.

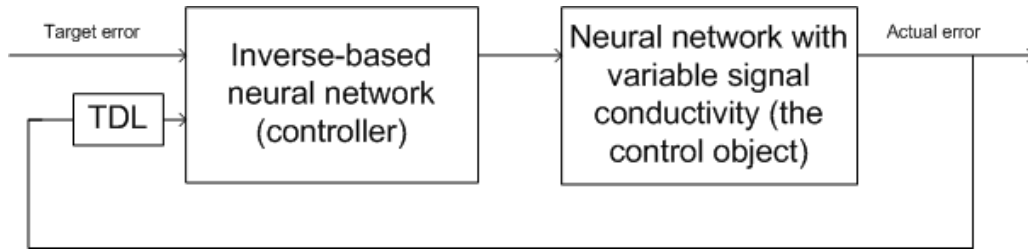


Fig.5. A principal scheme of inverse-based neural network control.

The main idea of the scheme given in fig.5 is as follows. We realize a control strategy for the neural network with variable signal conductivity (which is a controllable object) using another neural network (which is a controller). We have a target error curve (fig.4) and we know all statements of every weight per every moment so we may represent every weights change like a control step act. In epoch  $(k-1)$  we know all the picture of weights changes and hence the total of used control acts. Also we know the target level of the error function in epoch number  $k$  and the actual level of the error function in epoch  $k$ . So we may describe this situation as:

$$U(k - 1) = f(E_k, E_k^t) \tag{3}$$

where:

$k - 1$  is a number of previous epoch,

$k$  is a number of current epoch,

$E_k$  is an observable meaning of the error function

$E_k^t$  is a target meaning of the error function given by STL-decomposition

$U(k - 1)$  is the set of used control acts,

$f()$  is a reaction of the inverse-based neural network.

An equation (3) describes the inverse-based neural network training mode.

In the work circuit given in fig.5 we feed into the inversed neural network's input the target error in  $(k+1)$  epoch and actual error in current epoch. The output of the inversed neural network should be interpreted as the control signal applicable to the neural network with variable signal conductivity. The last described scheme will work properly if the inversed-based neural network is trained relevant to behavior of the control object. Let us consider some issues of the implementation of the scheme considered in fig.5.

In [8] there was an attempt to synthesize global control strategy for artificial neural network with variable signal conductivity solving Bellman optimal control feedback task. The main conclusion of investigation [8] is that a multilayer neural network with variable signal conductivity is an output-controllable system, but not a fully-controllable system. Authors got a principal control curve, but it is rather difficult to implement the founded control function. E.g., if the neural network has only 1440 neurons in a layer and every neuron has only 100 active links with non-zero weights values for only 10 layers (is equal to 10 stations), we get about  $k \times 10^6$  active weights. Hence we need to embed  $k \times 10^6$  control regulators to realize the founded in [8] the only control curve. It is potentially inconvenient and it is only theoretical result because we need to research the form of functional dependence between the error level  $E(t, u^*)$  and all set of  $k \times 10^6$  weights and regulators ( $u^*$  in this case means the founded optimal control curve).

To avoid this authors offer training a supervising neural network using a special database to store any step of the considered neural network with its all weights and neural statuses. Authors suggest a supervisor neural network which feeds desired error level and observed error level and returns one parameter of concrete weight value in the multilayer neural network with variable signal conductivity. It leads to creation of the ensemble of multilayer perceptrons which will give us every weight parameter of considerable neural network. The structure of the database is given below.

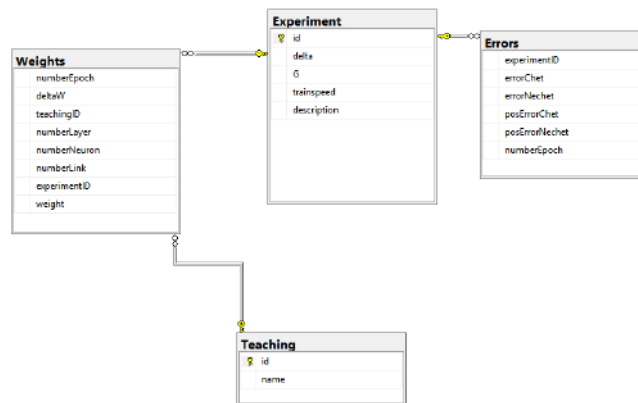


Fig.6. A structure of the database for storing the neural network statuses.

All the fields of the database are described in the table 1.

Table 1. The database fields' description.

Name of the database table	Name of the field	Data type	Description
Experiment	Id	int	Experiment identification number
	delta	int	Desired accuracy of the network
	G	int	Coefficient of the weight growth/decrease non-uniform rate of the links weights
	trainspeed	decimal(18, 5)	Initial train speed
	description	nchar(100)	comments
Errors	experimentID	int	Experiment identification number
	errorChet	decimal(18, 5)	Level of the even signals (for even trains schedule) error
	errorNechet	decimal(18, 5)	Level of the odd signals (for odd trains schedule) error
	posErrorChet	decimal(18, 5)	Positive even signal level error
	posErrorNechet	decimal(18, 5)	Positive odd signal level error
	numberEpoch	int	Number of epochs
Weights	numberEpoch	int	Number of epochs
	deltaW	float	Difference of the concrete weight level
	teachingID	int	Teaching type identifier
	numberLayer	int	Number of layer
	numberNeuron	int	Number of neuron
	numberLink	int	Number of link
	experimentID	int	Experiment identification number
Teaching	weight	float	Value of the concrete weight level
	Id	int	Teaching type identifier
	name	nchar(20)	Teaching type name

A principal control circuit is given in fig.7.

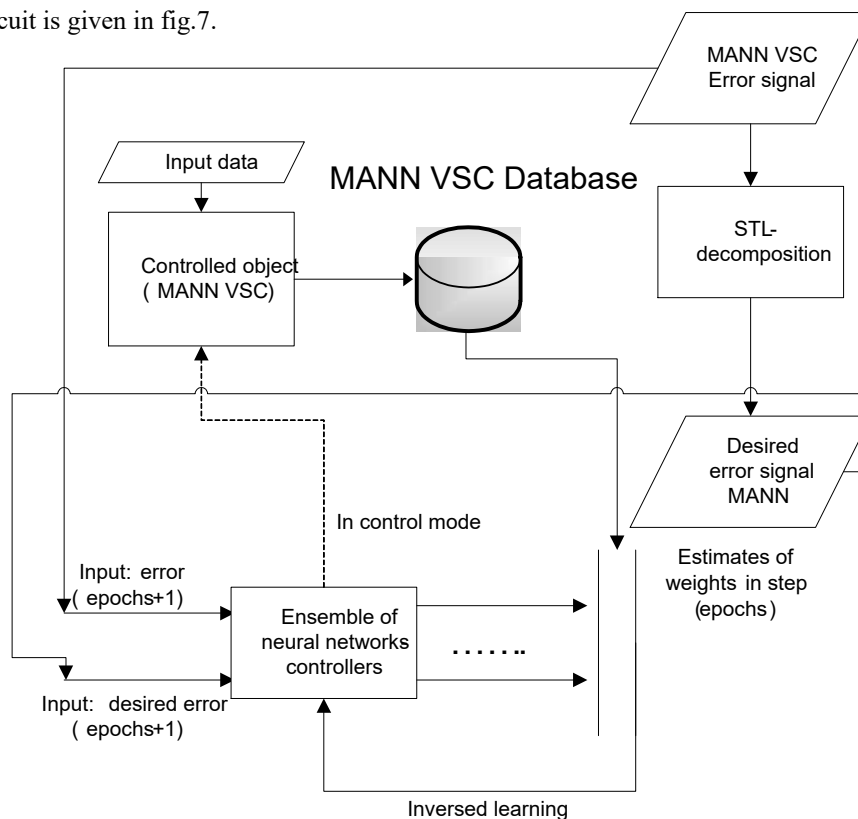


Fig.7. A structure of the control circuit and the database.

On fig.7: MANN VSC is a multilayer artificial neural network with variable signal conductivity, MANN VSC Database is a special database described in table 1, “Ensemble of neural networks controllers” is a supervisor neural network.

The main idea of fig.7 is the follows. Entrance of the controllable neural network feeds the moments of train departures and its desired arrival moments. The multilayer neural network with variable signal conductivity is functioning according to its algorithms and rules and is returning a set of error curves. These curves are being decomposed by STL to create a set of desired error signals. During this moment all statuses of the controllable neural network are being saved in the database for further storage. Values of desired error signals and real error signals are entering the entrance of the neural network supervisor ensemble, statuses of the controllable neural network are entering the output layers of the ensemble. We get inversed training of the ensemble because the order and the training procedure are implemented as in fig.5.

In current mode all trained weights for each epoch would be given to the MANN VSC entrance.

So in the present paper authors describe a new prospective approach to train the neural networks for transport scheduling.

## 5. Conclusions

1. To improve existing training methods of the neural networks with variable conductivity of signals it is possible to use some rigor mathematical disciplines. The theoretical basis for it can be synthesized of the settlements from Theory of Optimal Control and Theory of Signal Processing. It allows us to construct various transport timetables in more effective way.
2. Each neural network despite the kind of solved problems should be detected and registered. Its error signal should be processed and filtered to distinguish main component from signal series.
3. According to the trend component of the signal the system of weights and links of the neural network must be modified using one of techniques described above.
4. It worth sharing a new control scheme working with an inverse-based neural network control. The specific feature of the considered task is the neural network as a controllable object. This representation is innovative because in classical case we have a technical or a chemical systems described by its evolutionary equations as the controlled system and a neural network as a controller.

## Acknowledgements

Our gratitude to Professor Ivanov B.G. and Kopeykin S.V. (Samara State Transport University) and Professor Prokhorov S.A. (Samara State Aerospace University) for constructive critical feedback and very beneficial advice.

## References

- [1] Hopfield JJ, Tank DW. Neural Computation of Decisions in Optimization Problems. *Biological cybernetics* 1985; 52; 3: 141–152.
- [2] Chen RM, Huang YM. Competitive neural network to solve scheduling problems. *Neurocomputing* 2001; 37(1): 177–196.
- [3] Kostenko VA, Vinokurov AV. Local-optimal scheduling algorithms based on the use of Hopfield networks. *Programmirovaniye* 2003; 4: 27–40. (in Russian)
- [4] Martinelli DR, Teng H. Optimization of railway operations using neural networks. *Transportation Research Part C: Emerging Technologies* 1996; 4(1): 33–49.
- [5] Fahotimi O, Dembo A, Kailath T. Neural network weight matrix synthesis using optimal control techniques. USA, Stanford University. URL: <http://papers.nips.cc/paper/191-neural-network-weight-matrix-synthesis-using-optimal-control-techniques.pdf>.
- [6] Becerikli Y, Konar AF, Samad T. Intelligent optimal control with dynamic neural networks. *Neural networks* 2003; 16(2): 251–259.
- [7] Ignatenkov AV. Model of an artificial neural network for plotting the traffic schedule of trains on a two-track section. *International Scientific Conference Proceedings “Advanced Information Technologies and Scientific Computing”*. Samara Scientific Center of RAS Publishing, 2016; 619–623. (in Russian)
- [8] Ignatenkov A, Olshansky A. Extent of error control in neural networks. *Cornwell University Library*. URL: <https://arxiv.org/abs/1608.04682> (03.01.2017).
- [9] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics* 1990; 1(6): 3–73.

# A discrete phase problem in reconstruction of signals in space-rocket hardware

A.A. Kuleshova<sup>1</sup>, E.A. Shchelokov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

Reconstruction of information hidden in vector signal phases does not lose its relevance. Sets of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$ , called frames, in a space  $C^m(R^m)$  can be used for theoretical research of phase retrieval. The article shows that phase retrieval is equivalent to phaseless reconstruction. Examples are considered in  $R^m$  and  $C^m$ , for which sets of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  that simultaneously carry out phase retrieval and phaseless reconstruction are constructed.

*Keywords:* phase retrieval; phaseless reconstruction; frame; complement property; weak phase retrieval; generic frame

---

## 1. Introduction

A search of the fast algorithms for phaseless signal reconstruction is topical now. The main property of frames, which makes them so useful in applied tasks, is their redundancy. A well-chosen frame can provide numerical stability for signal recovery and obtaining important characteristics of the signal [1]. A family of frames recovers the signal by absolute values of frame coefficients in polynomial time.

It is shown that in the real case under certain conditions a generic frame consisting of  $(2m-1)$ -vectors can recover the signal without phases. The similar result was obtained in the complex space for  $(4m-2)$ -vectors.

Along with the "phaseless reconstruction", another version of the discrete phase problem statement – "phase retrieval" – is considered. The issue of their equivalence is raised and partially resolved.

The present work continues this line of research and gives examples of signal recovery in small-dimension spaces.

## 2. A Discrete Phase Problem in Reconstruction of Signals in Space-Rocket Hardware

Now the problem of reduction of the mass of cable systems in spacecraft is widely known. In this connection we offer for consideration an option of replacement of the cable system by a radio channel [2].

Widespread introduction of wireless devices has become possible as a result of improvement and reduction in cost of electronic components. Modern chips, which are used in construction of wireless networks, only require connection of several passive components and program setup.

In connection with the above-mentioned it is reasonable to consider the introduction of wireless technologies into space-rocket hardware as one of ways to reduce the mass and complexity of the cable system.

Let's consider, as an example, a signal transmitted by a radio channel with a modulation of an OFDM type. The main advantage of the chosen method is that the signal propagation delay is much less, than time of transfer of a symbol in auxiliary carriers as compared to other types of modulation. That allows implementing more stable transfer of information under conditions of symbols overlapping in the course of rereflections of the signal.

Figure 1 shows a distribution model of levels of signals from a different number of access points for a case when a set of blocks is arranged inside a spacecraft. The model is presented in a two-dimensional form, however, as is obvious from distribution of levels of signals, 4 access points is enough to provide communication of all blocks among themselves, including by relaying.

As is evident from figure 1, the signal levels at the border of the external and internal parts of the compartment do not exceed minus 30 dBm (blue color:-30 to -40 dBm, green color:-40 to - 50 dBm) relative to 0 dBm at the antenna exit. If we take into account that aluminum has a shielding factor of 70 dBm (for the thickness of 5 mm) we obtain the coefficient of signal attenuation outside the working zone equal to 100 dBm. If an additional protection is necessary, it is enough not to allow blocking direct visibility of the transmitting part and the compartment border, which will increase attenuation by 30-40 dB [2].

In OFDM modulation, data are distributed among a great number of auxiliary carriers, that's why it is necessary to recover the information lost in several subchannels for further data handling. A Search of the signal recovery algorithms is topical now. Sampling and quantization of the analog signal lead to consideration of the signal as an element of a finite-dimensional space  $V$ . By the orthonormalized basis (ONB)  $\{u_i\}_{i=1}^m$ , the "signal"  $v \in V$  can be uniquely represented in the form of the sum:

$$v = \sum_{i=1}^m \langle v, u_i \rangle u_i.$$
 Actual measurements prove real, and the gap between  $\langle v, u_i \rangle$  and amplitudes of measurements  $|\langle v, u_i \rangle|$  proves

insuperable in signal reconstruction [1, P. 280], [4, P.281].

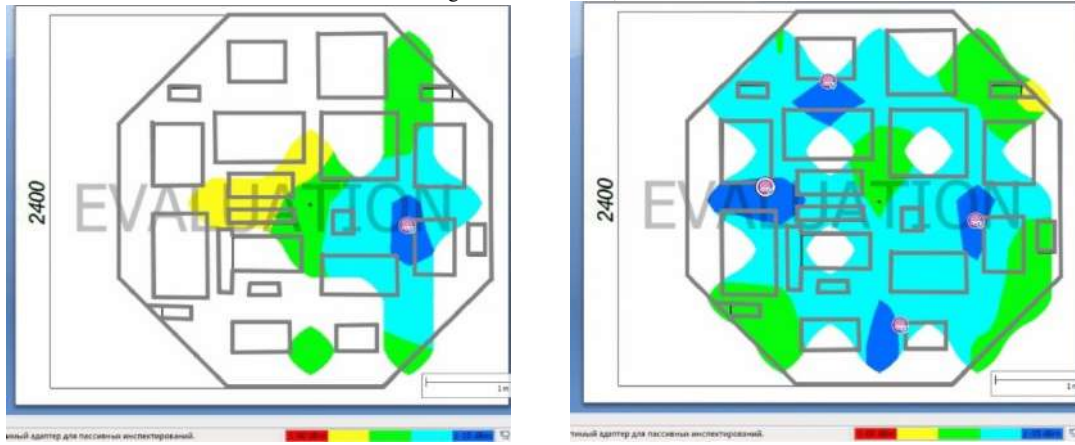


Fig. 1. Distribution of the signal from one access point (on the left) and from four access points (on the right).

In recent years, significant amount of works has been devoted to solution of the following task: to construct such systems of "measuring" vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  that allow recovering a signal  $v \in V$  by a set of real numbers  $\{|\langle v, \varphi_i \rangle|\}$ .

Such task has no decision in the ONB class.

The main problem set in [3] is still far from final solution. It is to find necessary and sufficient conditions for a system of representation vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  (so-called "measuring vectors"), which provide injectivity and stability of mapping of "amplitude measurement" of the signal  $x$

$$(A(x))(i) := |\langle x, \varphi_i \rangle|^2$$

We have proved that exact recovery of the signal (to the unimodular multiplier) is theoretically possible if complete redundant systems are used as a representation system [2, P. 354]. Frames are such redundant systems.

In 2006, Balan/Casazza/Edidin [4,5] defined one of versions of the discrete phase problem, which they called "phaseless reconstruction". It was shown that in the real case a generic frame consisting of  $(2m-1)$ -vectors can do phaseless reconstruction under certain conditions. The similar result was obtained for  $(4m-2)$ -vectors in the complex space.

### 3. Frames

Let  $H^m$  be a space  $R^m$  or  $C^m$ .

**Definition 1.** A family of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  is called a frame of a Hilbert space  $H^m$  if there are such constants  $0 < A \leq B < \infty$  that for all  $x \in H^m$  the following inequalities are achieved:

$$A \|x\|^2 \leq \sum_{i=1}^n |\langle x, \varphi_i \rangle|^2 \leq B \|x\|^2.$$

$A$  and  $B$  are called frame bounds. The greatest of the lower bounds is called the optimum lower bound, and the smallest of the upper bounds is the optimum upper bound. If  $A=B$ , then the frame is called  $A$ -tight and if  $A=B=1$ , it is called a Parseval-Steklov frame.

The numbers  $\{|\langle x, \varphi_i \rangle|\}_{i=1}^n$  are called frame coefficients.

If all frame elements have the same norm than such frames are called uniform ones.

In the finite-dimensional space the notion of a frame is equivalent to the notion of completeness of a system, that is to the equality  $span\{\varphi_i\}_{i=1}^n = H^m$  [5].

**Definition 2.** Let  $\Phi = \{\varphi_i\}_{i=1}^n$  be a frame. The linear mapping:

$$T: H^m \rightarrow H^n, \quad T(x) = \{\langle x, \varphi_i \rangle\}_{i=1}^n$$

is called an analysis operator.

**Definition 3.** The linear mapping:

$$T^*: H^n \rightarrow H^m, \quad T^*(\{c_i\}_{i=1}^n) = \sum_{i=1}^n c_i \varphi_i$$

is called a synthesis operator.

The composition of  $T$  and  $T^*$  defines a frame operator, which is a positive, self-conjugate reversible operator:

$$S = T^*T: H^m \rightarrow H^m: Sx = T^*Tx = \sum_{i=1}^n \langle x, \varphi_i \rangle \varphi_i.$$

It provides the exact formula for reconstruction:

$$x = \sum_{i=1}^n \langle x, \varphi_i \rangle S^{-1} \varphi_i.$$

**Definition 4.** A family of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  is a uniform equiangular tight frame if



- 1)  $\exists \beta > 0: \|\varphi_i\| = \beta \quad \forall i = 1, n$  ;  
 2)  $\exists c > 0$ : for any pair of frame vectors  $\varphi_j$  and  $\varphi_k, j \neq k$ , we have:

$$\langle \varphi_j, \varphi_k \rangle = c.$$

It is known that there is an upper bound for a number of vectors in the uniform equiangular tight frame  $\Phi = \{\varphi_i\}_{i=1}^n$  on the  $m$ -dimensional Hilbert space  $H$ . In the real case it is  $n \leq \frac{m(m+1)}{2}$ , in the complex case it is  $n \leq m^2$  ([6], [7]). Creation of the maximum number of vectors for the uniform equiangular tight frame is a very complex and unresolved problem in the theory of frames.

Let us consider a non-linear mapping  $P$ , which transfers the vector into a set of the absolute values of frame coefficients:

$$P: H \rightarrow l^2(I), P(x) = \{|\langle x, \varphi_i \rangle|\}_{i=1}^n$$

**Definition 5.** The frame  $\{\varphi_i\}_{i=1}^n$  is called generic if  $\{\varphi_i\}_{i=1}^n \subset L \in U$ , where  $U$  is the Zariski open set and  $U \subset Gr(m, n)$ .

**Theorem 1 [8,9].** Let  $\Phi = \{\varphi_i\}_{i=1}^n \subseteq C^m$  and the mapping  $A: C^m = C^m / T^1 \rightarrow R^m$  be defined by  $(A(x))(i) := |\langle x, \varphi_i \rangle|^2, i = 1, \dots, n$ .

Let us consider  $\{\varphi_i \varphi_i^* u\}_{i=1}^n$  as vectors of the space  $R^{2m}$ . Let  $S(u) := span_R \{\varphi_i \varphi_i^* u\}_{i=1}^n$ . The following statements are equivalent:

- (a)  $A$  is injective.  
 (b)  $\dim S(u) \geq 2n - 1$  for every  $u \in C^m \setminus \{0\}$ .  
 (c)  $S(u) = span_R \{iu\}^\perp$  for every  $u \in C^m \setminus \{0\}$ .

**Definition 6.** The family of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  in  $H^m$  has the complement property if for any  $I \subseteq \{1, \dots, n\}$ , either  $\{\varphi_i\}_{i \in I}$ , or  $\{\varphi_i\}_{i \in I^c}$  is complete in  $H^m$  [10].

**Definition 7.** The family of vectors  $\{\varphi_i\}_{i=1}^n \subseteq R^m$  is called a set with a full spark, if every its subset of  $m$  vectors is complete in  $R^m$  [10].

**Lemma 1.** Every set with the full spark  $\Phi = \{\varphi_i\}_{i=1}^n$  in  $R^m$  with  $n \geq 2m - 1$  satisfies the complement property.

**Proof.** Let's assume the contrary: there is such  $S \subseteq \{1, \dots, n\}$  that neither  $\{\varphi_i\}_{i \in S}$  nor  $\{\varphi_i\}_{i \in S^c}$  are not complete in  $R^m$ .

By definition of the full spark, from this it follows that  $|S| < m - 1$  and  $|S^c| < m - 1$ , that is  $n < 2m - 2$ , which contradicts the condition.

**Theorem 2.** In the real case if  $\Phi = \{\varphi_i\}_{i=1}^n$  in  $R^m$  and  $n \leq 2m - 2$ , then mapping  $A$  is not injective.

If  $n = 2m - 1$ , then mapping  $A$  is injective if and only if when  $\Phi = \{\varphi_i\}_{i=1}^n$  is a full spark.

**Proof.** If  $n \leq 2m - 2$ , then the set  $\{1, \dots, n\}$  can be divided into sets  $S$  and  $S^c$  such that the cardinality of each would not exceed  $m - 1$ . None of the sets  $\{\varphi_i\}_{i \in S}, \{\varphi_i\}_{i \in S^c}$  can be complete.

If  $n = 2m - 1$  and  $\Phi = \{\varphi_i\}_{i=1}^n$  is a full spark, then the injectivity of  $A$  follows from lemma 1 and theorem 1.

And vice-versa, if  $A$  is injective, then  $\Phi = \{\varphi_i\}_{i=1}^n$  is an alternatively full family. Let's take an arbitrary subset  $S \subseteq \{1, \dots, n\}$  with  $|S| = m$ . Then  $|S^c| = m - 1$  and  $\{\varphi_i\}_{i \in S^c}$  can't be full. Therefore,  $\{\varphi_i\}_{i \in S}$  is full, and  $\Phi = \{\varphi_i\}_{i=1}^n$  is a full spark.

The exact minimum bound is unknown for the complex case. Besides, in the real case there is a simple direct method for checking injectivity of the mapping  $A$  for the corresponding frame [7].

**Theorem 3 [4, 11].**

(a) If  $H \in R^m, n \geq \frac{m(m+1)}{2}$  and  $\Phi = \{\varphi_i\}_{i=1}^n$  is a generic frame, the nonlinear map  $P$  is injective. Then the vector  $x \in H$  can be reconstructed (up to a sign) from the set  $\{|\langle x, \varphi_i \rangle|\}_{i=1}^n$  of absolute values of the frame coefficients in a polynomial number ( $O(m^6)$ ) of steps.

(b) If  $H \in C^m, n \geq m^2$  and  $\Phi = \{\varphi_i\}_{i=1}^n$  is a generic frame, the nonlinear map  $P$  is injective. Then the vector  $x \in H$  can be reconstructed (up to multiplication by a root of unity) from the set  $\{|\langle x, \varphi_i \rangle|\}_{i=1}^n$  of absolute values of the frame coefficients in a polynomial number ( $O(m^6)$ ) of steps.

#### 4. About Equivalence of Phase Retrieval and Phaseless Reconstruction

Let  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$  be vectors in a space  $H^m$ .

**Definition 1:** For the phase of number  $z \in C^m$ , we take the value of the angle  $\varphi = \text{ph } z_i + 2\pi k, k \in Z$ , defining the deviation of the radius vector of the point on the plane, corresponding to  $z \in C^m$ , from the real axis in  $C^m$ . In the real case, the phase in  $R^m$  is equal to 0 or  $\pi$ .

We shall say that  $x, y$  have the same phases if:

$$ph a_i = ph b_i, i = 1, 2, \dots, m.$$

**Definition 2.** Let  $\Phi = \{\varphi_i\}_{i=1}^n$  be a family of vectors in  $H^m$  (respectively,  $\{P_i\}_{i=1}^n$  be a family of projections on  $H^m$ ) satisfying the following property: for every  $x, y$  the following condition is satisfied:

$$|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|, \text{ for all } i = 1, 2, \dots, n.$$

(Respectively,

$$\|P_i x\| = \|P_i y\|, \text{ for all } i = 1, 2, \dots, n).$$

Then

1) If there is a  $|\theta| = 1$  such that  $x$  and  $\theta y$  have the same phases, one can say  $\Phi = \{\varphi_i\}_{i=1}^n$  does phase retrieval (respectively,  $\{P_i\}_{i=1}^n$  does phase retrieval).

2) If there is a  $|\theta| = 1$  such that  $x = \theta y$ , one can say  $\Phi = \{\varphi_i\}_{i=1}^n$  does phaseless reconstruction. (Respectively,  $\{P_i\}_{i=1}^n$  does phaseless reconstruction.)

**Definition 3.** Let us call a family of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  in  $H^m$  an alternatively full one, if for any  $I \subseteq \{1, \dots, n\}$ , either  $\{\varphi_i\}_{i \in I}$ , or  $\{\varphi_i\}_{i \in I^c}$  is complete in  $H^m$ .

If  $\{\varphi_i\}_{i=1}^n$  retrieves phases in  $H^m$  then  $span\{\varphi_i\}_{i=1}^n = H^m$ . This means that  $\{\varphi_i\}_{i=1}^n$  is a frame in the space  $H^m$ . Otherwise, there exists  $0 \neq x \in H^m$  such that  $\langle x, \varphi_i \rangle = \langle y, \varphi_i \rangle = 0$ ,  $i = 1, 2, \dots, n$ , while phases of vectors  $x$  and  $0$  are not the same.

If  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$  have the same phases then  $a_i = 0$  if and only if  $b_i = 0$ , for the phase of  $0$  is not determined.

**Theorem 1.** Let  $\Phi = \{\varphi_i\}_{i=1}^n$  be a set of vectors in  $R^m$ . The mapping  $A: R^m / \{\pm 1\} \rightarrow R^n$  ( $n > m$ ) is defined by  $(A(x))(i) := |\langle x, \varphi_i \rangle|^2$ ,  $i = 1, \dots, n$ . If  $\Phi = \{\varphi_i\}_{i=1}^n$  does phaseless reconstruction, then it has a complement property. In the real case these concepts are equivalent.

**Proof.** ( $\Rightarrow$ ) Assume that  $\Phi$  fails the complement property. Then there exists  $I \subseteq \{1, \dots, n\}$  such that neither  $\{\varphi_i\}_{i \in I}$ , nor  $\{\varphi_i\}_{i \in I^c}$  is complete in  $R^m$ .

We choose nonzero vectors  $u, v \in R^m$  such that  $\langle u, \varphi_i \rangle = 0$  for all  $i \in I$  and  $\langle v, \varphi_i \rangle = 0$  for all  $i \in I^c$ . For every  $i$  we then have:

$$|\langle u \pm v, \varphi_i \rangle|^2 = |\langle u, \varphi_i \rangle|^2 \pm 2\langle u, \varphi_i \rangle \overline{\langle v, \varphi_i \rangle} + |\langle v, \varphi_i \rangle|^2 = |\langle u, \varphi_i \rangle|^2 + |\langle v, \varphi_i \rangle|^2.$$

From this it follows that  $|\langle u + v, \varphi_i \rangle|^2 = |\langle u - v, \varphi_i \rangle|^2$  for every  $i$ , and  $A(u + v) = A(u - v)$ . Moreover, since  $u$  and  $v$  are nonzero by assumption, then  $u + v \neq \pm(u - v)$ . Thus there is no phaseless reconstruction.

( $\Leftarrow$ ) Assume that  $\Phi = \{\varphi_i\}_{i=1}^n$  fails phaseless reconstruction. That means there exist vectors  $x, y \in R^m$  such that  $x \neq \pm y$  and  $A(x) = A(y)$ . Take  $I := \{i : \langle x, \varphi_i \rangle = -\langle y, \varphi_i \rangle\}$ .

We have:  $\langle x + y, \varphi_i \rangle = 0$  for every  $i \in I$ . Otherwise if  $i \in I^c$ , we have  $\langle x, \varphi_i \rangle = \langle y, \varphi_i \rangle$  and then  $\langle x - y, \varphi_i \rangle = 0$ . According to the assumption,  $x \neq \pm y$ , therefore  $x + y \neq 0$  and  $x - y \neq 0$ . Thus, neither  $\{\varphi_i\}_{i \in I}$  or  $\{\varphi_i\}_{i \in I^c}$  are complete in  $R^m$ .

In  $R^m$  the phase of a vector can be equal to  $0$  or  $\pi$ . Coordinates of the vectors have the same phases if signs of the coordinates of the vector  $x$  are the same as those of the vector  $y$ . At the same time the phase of  $0$  is not defined. That is, if  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$ , then  $x, y$  have the same phase in the following cases:

- 1) If  $a_i \neq 0 \neq b_i$ , then  $a_i b_i > 0$ .
  - 2) If  $a_i = 0$ , then corresponding to it  $b_i = 0$  (it is symmetric: if  $b_i = 0$ , then corresponding to it  $a_i = 0$ ).
- Otherwise, the vectors have different phases.

Then, if we are given two vectors, so as to define whether their phases are equal or not it is necessary:

1) To check equality of all indices of zero coordinates of the vectors. If all indexes of zero coordinates of the first vector correspond to the indexes of the second vector (and vice versa), then it is necessary to check 2), otherwise, the vectors have different phases.

2) For nonzero coordinates to check fulfillment of the following condition: if  $a_i b_i > 0 \Rightarrow$  the vectors have the same phases, and if  $a_i b_i < 0$  then the vectors have different phases.

Definition 1 in the real case will mean:

Let  $\Phi = \{\varphi_i\}_{i=1}^n$  be a set of vectors in  $R^m$ , satisfying the following property: for every  $x, y$  the following condition is fulfilled:

$$|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|, i = 1, 2, \dots, n.$$

Then,

- 1)  $\Phi = \{\varphi_i\}_{i=1}^n$  does phases reconstruction if there exist  $\theta = \pm 1$  such that
  - a) For  $\theta = 1$  the vectors  $x$  and  $y$  have the same phase.
  - b) For  $\theta = -1$  the vectors  $x$  and  $-y$  have the same phase.
- 2)  $\Phi = \{\varphi_i\}_{i=1}^n$  does phaseless reconstruction if there exists  $\theta = \pm 1$  such that
  - c) For  $\theta = 1 \Rightarrow x = y$ .
  - d) For  $\theta = -1 \Rightarrow x = -y$ .

**Theorem 2.** Let  $\Phi = \{\varphi_i\}_{i=1}^n$  be a set of vectors in  $R^m$ . The mapping  $A: R^m / \{\pm 1\} \rightarrow R^n$  ( $n > m$ ) is defined by equations  $(A(x))(i) := \langle x, \varphi_i \rangle^2$ ,  $i = 1, \dots, n$ . If  $\Phi = \{\varphi_i\}_{i=1}^n$  does phase retrieval, then it has the complement property. In the real case these concepts are equivalent.

**Proof.** Assume that  $\Phi$  does phase retrieval, but fails phaseless reconstruction. Assume that the set  $\Phi = \{\varphi_i\}_{i=1}^n$  fails complement property, that is there exists  $I \subseteq \{1, \dots, n\}$  such that neither  $\{\varphi_i\}_{i \in I}$ , nor  $\{\varphi_i\}_{i \in I^c}$  is complete in  $R^m$ .

Let us choose nonzero vectors  $x = (a_1, a_2, \dots, a_m)$ ,  $y = (b_1, b_2, \dots, b_m) \in R^m$  such that  $\langle x, \varphi_i \rangle = 0$  for all  $i \in I$  and  $\langle y, \varphi_i \rangle = 0$  for all  $i \in I^c$ . Then for some  $i$  either  $\langle x, \varphi_i \rangle = 0$ , or  $\langle y, \varphi_i \rangle = 0$ . Fix  $c \neq 0$ , so that for every  $1 \leq i \leq n$

$$\langle x + cy, \varphi_i \rangle = \langle x - cy, \varphi_i \rangle$$

Then

$$\langle x + cy, \varphi_i \rangle^2 = \langle x - cy, \varphi_i \rangle^2.$$

By assumption,  $\Phi$  does phase retrieval, and it means that there exists  $|\theta| = 1$  such that  $(x + cy)$  and  $\theta(x - cy)$  have the same phases. Let's assume that there exists  $1 \leq i_0 \leq m$  such that  $a_{i_0} \neq 0 \neq b_{i_0}$  and let  $c = \frac{-a_{i_0}}{b_{i_0}}$ . Then

$$(x + cy)_{i_0} = a_{i_0} + cb_{i_0} = a_{i_0} + \frac{-a_{i_0}}{b_{i_0}} b_{i_0} = 0$$

and

$$(x - cy)_{i_0} = a_{i_0} - cb_{i_0} = a_{i_0} - \frac{-a_{i_0}}{b_{i_0}} b_{i_0} = 2a_{i_0} \neq 0.$$

But it is impossible because if  $x$  and  $y$  have the same phases, then  $a_i = 0$  if and only if  $b_i = 0$ .

As the vectors  $x$  and  $y$  are nonzero, then the last two equalities are possible if and only if either  $a_i = 0$ , or  $b_i = 0$ ,  $1 \leq i \leq m$ . Let  $I = \{1 \leq i \leq m : b_i = 0\}$  and  $\{e_i\}_{i=1}^m$  be an orthonormalized basis in  $R^m$ . Then

$$x + y = \sum_{i \in I} a_i e_i + \sum_{i \in I^c} b_i e_i \text{ and } x - y = \sum_{i \in I} a_i e_i + \sum_{i \in I^c} (-b_i) e_i.$$

Then there exists  $|\theta| = 1$  such that  $(x + y)$  and  $\theta(x - y)$  have the same phases and they are equal. We have arrived at a contradiction.

Let's consider an example in  $R^2$ . Let  $x = (a_1, a_2) \neq (0, 0)$  and  $y = (b_1, b_2) \neq (0, 0)$ .

For  $\Phi = \{\varphi_i\}_{i=1}^3$ , we take the Mercedes-Benz frame in  $R^2$ , consisting of 3 unit vectors located at an angle of  $120^\circ$  (Fig. 2):

$$\varphi_1 = (0, 1), \quad \varphi_2 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right), \quad \varphi_3 = \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right).$$

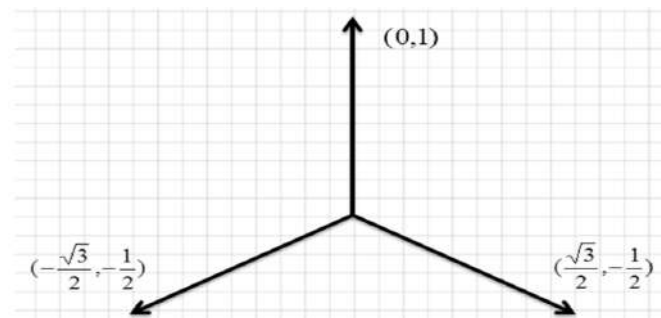


Fig. 2. The Mercedes-Benz Frame in  $R^2$ , consisting of 3 unit vectors located at an angle of  $120^\circ$ .

Then fulfillment of the condition  $|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|_{i=1}^3$  means that

$$\left\{ \begin{array}{l} |\langle (a_1, a_2), (0,1) \rangle| = |\langle (b_1, b_2), (0,1) \rangle| \\ \left| \left\langle (a_1, a_2), \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \right\rangle \right| = \left| \left\langle (b_1, b_2), \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \right\rangle \right| \\ \left| \left\langle (a_1, a_2), \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \right\rangle \right| = \left| \left\langle (b_1, b_2), \left(-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \right\rangle \right| \end{array} \right. \Rightarrow \left\{ \begin{array}{l} |a_2| = |b_2| \\ \left| \frac{\sqrt{3}}{2} a_1 - \frac{1}{2} a_2 \right| = \left| \frac{\sqrt{3}}{2} b_1 - \frac{1}{2} b_2 \right| \\ \left| -\frac{\sqrt{3}}{2} a_1 - \frac{1}{2} a_2 \right| = \left| -\frac{\sqrt{3}}{2} b_1 - \frac{1}{2} b_2 \right| \end{array} \right.$$

$$\left\{ \begin{array}{l} |a_2| = |b_2| \\ |\sqrt{3}a_1 - a_2| = |\sqrt{3}b_1 - b_2| \\ |\sqrt{3}a_1 + a_2| = |\sqrt{3}b_1 + b_2| \end{array} \right.$$

Squaring the equations of the last system we obtain that

$$a_1 a_2 = b_1 b_2 \text{ and } a_1^2 = b_1^2.$$

From this it follows that the first equality gives coincidence of the signs (to the multiplier), and the second – coincidence of the absolute values of coordinates. Moreover, from these equalities it also follows that zero coordinates are the same, if any.

We obtain that either  $x = y$ , or  $x = -y$ . Then there really exists  $\theta = \pm 1$  such that if  $\Phi = \{\varphi_i\}_{i=1}^3$  is the Mercedes-Benz frame, it does both phases reconstruction (because the vectors have the same signs, which means that the phases are the same too) and phaseless reconstruction (for  $x = \theta y$ ) at the same time.

If we know the absolute values of coordinates of the vectors  $x = (a_1, a_2)$  and  $y = (b_1, b_2)$ , then  $x$  and  $y$  can be one of 4 vectors (Fig. 3):

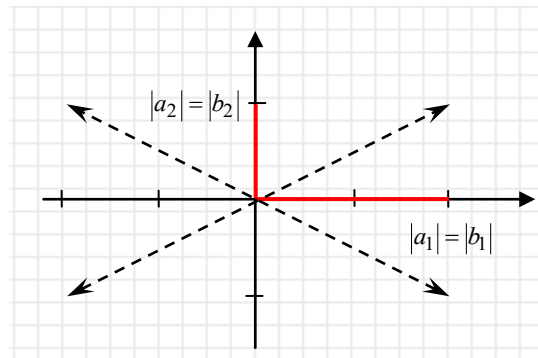


Fig. 3. Possible vectors for the known absolute values of coordinates of the vectors.  $x, y$ .

After scalar multiplication by the frame coordinates, the condition  $a_1 a_2 = b_1 b_2$  imposes restrictions on the signs of coordinates (Fig. 4):

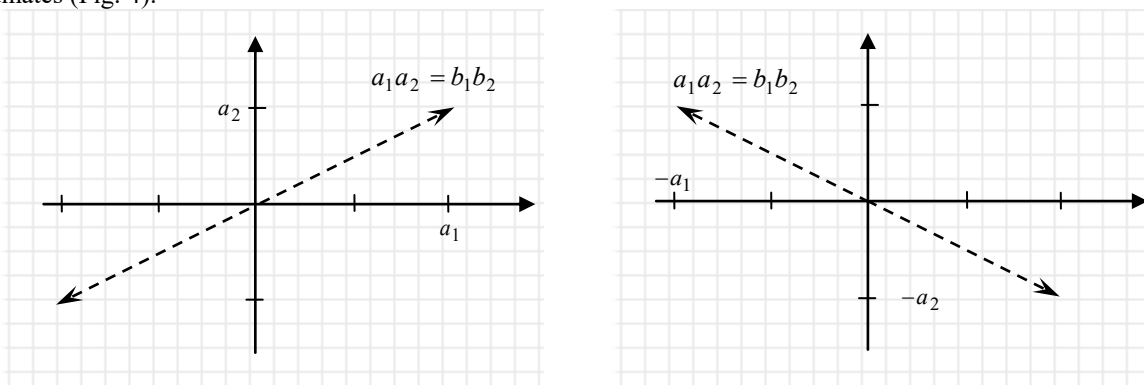


Fig. 4. Possible vectors for the known  $|\langle x, \varphi_i \rangle|, |\langle y, \varphi_i \rangle|$ .

Let's consider an example in  $C^2$ . In the complex case:

$$x = (x_1, x_2) = (a + ib, c + id) \text{ and } y = (y_1, y_2) = (e + if, g + ih).$$

As  $\Phi = \{\varphi_i\}_{i=1}^5$  let's take a frame of the following type:

$$\varphi_1 = (1, 0), \quad \varphi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right), \quad \varphi_3 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad \varphi_4 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i\right), \quad \varphi_5 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i\right).$$

Then fulfillment of the condition  $|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|_{i=1}^5$  means that

$$\left\{ \begin{array}{l} |\langle (x_1, x_2), (1, 0) \rangle| = |\langle (y_1, y_2), (1, 0) \rangle| \\ |\langle (x_1, x_2), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) \rangle| = |\langle (y_1, y_2), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) \rangle| \\ |\langle (x_1, x_2), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \rangle| = |\langle (y_1, y_2), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \rangle| \\ |\langle (x_1, x_2), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i) \rangle| = |\langle (y_1, y_2), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i) \rangle| \\ |\langle (x_1, x_2), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i) \rangle| = |\langle (y_1, y_2), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i) \rangle| \end{array} \right. .$$

Then  $(x, \varphi_i) = \langle (x_1, x_2), (\varphi_{i1}, \varphi_{i2}) \rangle = \sum_{\substack{j=1,2 \\ i=1,5}} x_j \overline{\varphi_{ij}}$  and  $|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|_{i=1}^5$  mean that:

$$\left\{ \begin{array}{l} |x_1| = |y_1| \\ |x_1 - x_2| = |x_1 - x_2| \\ |x_1 + x_2| = |x_1 + x_2| \\ |x_1 + x_2i| = |x_1 + x_2i| \\ |x_1 - x_2i| = |x_1 - x_2i| \end{array} \right. \Rightarrow \left\{ \begin{array}{l} |x_1| = |x_1| \\ |x_1 - x_2| = |y_1 - y_2| \\ |x_1 + x_2| = |y_1 + y_2| \\ |x_1 - x_2i| = |y_1 - y_2i| \\ |y_1 + y_2i| = |y_1 + y_2i| \end{array} \right. .$$

Let's rewrite the system in the following form:

$$\left\{ \begin{array}{l} a^2 + b^2 = e^2 + f^2 \\ (a-c)^2 + (b-d)^2 = (e-g)^2 + (f-h)^2 \\ (a+c)^2 + (b+d)^2 = (e+g)^2 + (f+h)^2 \\ (a+d)^2 + (b-c)^2 = (e+h)^2 + (f-g)^2 \\ (a-d)^2 + (b+c)^2 = (e-h)^2 + (f+g)^2 \end{array} \right. .$$

From the last system we obtain

$$c^2 + d^2 = g^2 + h^2 .$$

And it means that the absolute values of the second complex coordinates of the vectors  $x$  and  $y$  are equal, because:

$$c^2 + d^2 = |x_2| = g^2 + h^2 = |y_2| .$$

So, if we take a frame of the type  $\varphi_1 = (1, 0)$ ,  $\varphi_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ ,  $\varphi_3 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ ,  $\varphi_4 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i)$ ,  $\varphi_5 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i)$ ,

then the absolute values of the corresponding complex coordinates of the vectors  $x$  and  $y$  are equal, i.e.:

$$|x_1| = |y_1| = r_1 \text{ and } |x_2| = |y_2| = r_2 .$$

Now, let us write down the coordinates of the vectors in the polar form, taking into account the equality of the absolute values of the coordinates:

$$x = (x_1, x_2) = (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}) \text{ and } y = (y_1, y_2) = (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}) .$$

Then  $|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|_{i=1}^5$  will look as follows:

$$\left\{ \begin{array}{l} |\langle (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}), (1, 0) \rangle| = |\langle (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}), (1, 0) \rangle| \\ |\langle (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) \rangle| = |\langle (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}) \rangle| \\ |\langle (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \rangle| = |\langle (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \rangle| \\ |\langle (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i) \rangle| = |\langle (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}), (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}i) \rangle| \\ |\langle (r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i) \rangle| = |\langle (r_1 e^{i\psi_1}, r_2 e^{i\psi_2}), (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}i) \rangle| \end{array} \right. .$$

From this it follows that  $\varphi_1 = \psi_1$  and  $\varphi_2 = \psi_2$ . We obtain that phases of the vectors  $x$  and  $y$  are equal to  $2\pi k$ ,  $k \in \mathbb{Z}$ .

## 5. Weak phase retrieval

**Definition 1.** Two vectors  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$  do weak phase retrieval if there is a  $|\theta| = 1$  such that phase  $a_i = \theta$  phase  $b_i$ , for all  $i = 1, 2, \dots, m$ , such that  $a_i \neq 0 \neq b_i$ .

In the real case, if  $\theta = 1$ , we say that  $x, y$  have weakly like signs and if  $\theta = -1$  they have weakly opposite signs.

**Definition 2.** A family of vectors  $\Phi = \{\varphi_i\}_{i=1}^n$  in  $H^m$  does weak phase retrieval if from equalities

$$|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|, \quad i = 1, 2, \dots, n.$$

It follows that there exists a  $|\theta| = 1$ , such that

$$\text{phase } x_i = \theta \text{ phase } y_i, \quad \text{for all } i = 1, 2, \dots, m, \text{ so that } a_i \neq 0 \neq b_i.$$

The weak phase retrieval differs from the phase retrieval described in Definition 2 of Section 4 in that there can be both  $a_i = 0$  and  $b_i \neq 0$ .

Let us consider an example where the weak phase retrieval is done but the phase retrieval by Definition 2 of Section 4 is failed. Let us consider in  $R^m$  a set of vectors  $\Phi = \{\varphi_i\}_{i=1}^{m+1}$ , which coordinates form the following matrix columns:

$$A = \begin{pmatrix} 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}_{m \times (m+1)}.$$

Then for every  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$ , if  $|\langle x, \varphi_i \rangle|^2 = |\langle y, \varphi_i \rangle|^2$ , it follows that  $a_i a_j = b_i b_j$  for all  $i \neq j$ .

This set of  $(m+1)$ -vectors in  $R^m$  will do weak phase retrieval.

**Theorem 1:** Let  $x = (a_1, a_2, \dots, a_m)$  and  $y = (b_1, b_2, \dots, b_m)$  be two vectors in  $R^m$ . Then the following statements are equivalent:

- 1)  $\text{sgn}(a_i a_j) = \text{sgn}(b_i b_j)$ , for all  $1 \leq i \neq j \leq m$ .
- 2)  $x, y$  have either weakly like signs or weakly opposite signs.

## 6. Conclusion

In case the phase information is not available, the signal recovery is theoretically possible if redundant systems called frames are used as the system of representation. A well-chosen frame can provide numerical stability for recovery of the signal and obtaining important characteristics of the signal.

In the real case under certain conditions a generic frame consisting of  $(2m-1)$ -vectors can do phaseless reconstruction. In the complex space a generic frame consisting of  $(4m-2)$ -vectors can do the same under certain conditions. A family of frames recovers the signal by the absolute value of frame coefficients in polynomial time. The issue of the equivalence of phases retrieval and phaseless reconstruction is raised and partially resolved. Examples of signal recovery in small-dimension spaces are given.

A search and theoretical justification of new methods of recovery of information hidden in phases of transmitted signals and unavailable for measurements by publicly available physical instruments is conducted. The technique is based on the latest achievements in the research of complete linearly dependent systems called space frames.

## Acknowledgements

The authors thank S.Ya. Novikov for his fruitful discussions.

## References

- [1] Botelho-Andrade S, Casazza P, Van Nguyen H, Tremain J. Phase retrieval verses phaseless reconstruction. ArXiv:1507.05815 [math.FA] – 21 Jul 2015.
- [2] Shchelokov EA. Application of technologies of wireless data transmission on aerospace hardware. The Bulletin of the Ryazan state radio engineering university 2016; 56: 131–135
- [3] Bandeira A, Cahill J, Mixon D, Nelson A. Saving phase: Injectivity and stability for phase retrieval. Applied and Computational Harmonic Analysis (ACHA) 2014; 37(1.1): 106–125.
- [4] Balan R, Bodmann BG, Casazza PG, Edidin D. Fast algorithms for signal reconstruction without phase. Proceedings of SPIE-Wavelets XII, San Diego 2007; 6701: 670111920–670111932.
- [5] Balan R, Casazza P, Edidin D. On signal reconstruction without phase. Appl. Comput. Harmon. Anal. 2006; 20: 345–356.
- [6] Holmes R, Paulsen VI. Optimal frames for erasures. Lin. Alg. Appl. 2004; 377: 31–51.
- [7] Balan R, Bodman BG, Casazza PG, Edidin D. Painless reconstruction from magnitudes of frame coefficients, preprint.
- [8] Novikov SYa, Fedina ME. Complete systems in problems of signal reconstruction. Proceedings of the International Scientific and Technical Conference. Vol. 1. Perspective Information Technologies 2015; 280–283. (in Russian)
- [9] Cameron PJ, Seidel JJ. Quadratic forms over GF(2). Indag. Math. 1973; 35: 1–8.

[10] Cahill J, Mixon DG. Full Spark Frames. Available online: arXiv:1110.3548.

[11] Kuleshova A. Generic frame in problems for signal reconstruction without phase. Information Technology and Nanotechnology 2016, Ceur WS 2016.; 1638; 364–372.

# Mathematical models for forecast of geometrical parameters of assembly units

V.A. Pechenin<sup>1</sup>, M.A. Bolotov<sup>1</sup>, N.V. Ruzanov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The article describes two mathematical models for forecast of the geometric parameters of assembly units: the model of surface fitting and the combinatorial search model. The objects of modeling are a pair of mating cone rings. The first cone ring is characterized by the form deviation of the mating surface. The second cone is characterized by the runout of the outer and inner surfaces. Approval of the developed models was carried out through experimental studies of the assembly of two tailored and certified cone rings. A comparison of results of theoretical and experimental studies has shown that the developed models allow forecasting the geometric parameters of assembly units.

*Key words:* mathematical model; form deviation; filtration; assembly; geometric parameter; coordinate measurements

---

## 1. Introduction

The performance characteristics of machines are in large part determined by the geometric precision of the parts and assembly units. The geometric precision of the assembly units is standardized by the assembly parameters and depends on the deviations of the form and location of the parts surfaces. The accuracy of the assembly parameters is achieved by the correct assignment of requirements for the deviations of the form and location of the parts surfaces and the choice of the process of assembly. A variety of sources [1, 2, 3, 4] is devoted to the technology of manufacturing and assembly of parts. Development of design of aviation and space equipment is constantly increasing the requirement for quality level of items. These requirements directly affect both the manufacturing processes of production and the technological processes of assembling individual parts and assembly units. Substantial reserves of increase of accuracy and productivity of assembly process are developed by using forecast and optimization models directly to manage the process of forming a required accuracy of assembly. The implementation of such reserves is possible when performing forecasting calculations based on real geometric models [5]. It is assumed that, the possible spatial positions of the parts achieved during mating through a plurality of surfaces can be calculated immediately before the assembly. In this case, the calculations must take into account the deviation of form and location of surfaces of the parts. Forecast assembly models are also useful for assembling micromechanical elements, for example microturbines [6, 7].

The paper describes two mathematical models that allow forecasting the relative positions of parts that have geometric deviations in the assembly. A description of the developed models is provided below.

## 2. Object of research

To research the forecast of assembly parameters, the pair of rings were manufactured from tool steel Cr12. Mating surfaces of rings are conical with specified deviations. The small ring opening is a precise cylindrical surface, the outer surface of the larger ring is the exact conical surface. This feature of the design is associated with the need to determine accurately the coordinates of the centers of cones during physical experiments and it is supported by the use of the basing procedure.

The internal conical ring has a form deviation, which can be described by the sum of the harmonic functions. The external ring has precise conical surfaces and the runout of the outer and inner surfaces. The parameters of mating surfaces of the rings are presented in Table 1.

Physical experiments were conducted on assembling conical rings as a check on the adequacy of the developed models for forecast of assembly parameters. The resulting positions were obtained by measuring the base surfaces of the conical rings in the assembled form by the coordinate measuring machine DEA Global Performance 07.07.07.

Table 1. Parameters of mating rings.

Parameters	The inner ring	The outlet ring
Height, mm	36,295	39,971
An average radius of the circle of the lower section, mm	27,36	27,3487
Form deviation, mm	0,09	Absent
Runout of mating and base surfaces, mm	Absent	0,389

## 3. Methods

The following paragraph describes developed models for forecasting assembly states, as well as the procedure for generation of data on the geometry of the assembled rings using coordinate measurements.



### 3.1. Forecast of assembly parameters based on the model of fit of surfaces

To calculate and forecast parameters of mating of parts consisting of surfaces with deviations, a model was developed; it is described in detail in [8] and that allows to calculate the mating of parts without taking into account deformations during the assembly process. The model uses an iterative algorithm to define the mating state, which involves iterative displacement of one interfacing surface relative to the other in the direction of the application vector  $\overline{D}_1$  of the assembly force of the surfaces. In [8] the concept of a gap function  $G(\overline{V})$  is introduced, it characterizes the achievement of the mating state of the surfaces of parts and depends on the vector of the geometric relationship of surfaces  $\overline{V}$ . To calculate the function  $G(\overline{V})$ , the best fit of the mating surfaces is performed at each stage, for this reason this approach is called the method of surfaces fitting (MSF). To perform the best fit procedure the iterative nearest-point algorithm (ICP) is used, it is presented in [9]. According to this algorithm at each iteration by the methods of non-linear optimization search, the rotation angles and components of the displacement vector are calculated along the coordinate axes. To avoid intersections of two surfaces, the system of inequalities is used, which are presented in [10], which imposes restrictions on the gap function.

### 3.2. Combinatorial search model

In the case of mating of a conical ring having the same form deviation along the generator line and an ideal conical surface, it is possible to find the assembly center in a simpler way rather than using the ICP approach. The essence of the approach is to find the parameters of a circle of minimal radius circumscribed around one of the sections of the external cone ring. To solve this problem, a combinatorial search model was developed basing on enumeration of combinations of  $n$  defining profile points by 3 points at a time without repetitions. Each set of 3 points defines a circle, and its center and radius are calculated. The parameters of the circles are found from the system of equations:

$$\begin{cases} (P_{x1} - O_x)^2 + (P_{y1} - O_y)^2 = R_c^2, \\ (P_{x2} - O_x)^2 + (P_{y2} - O_y)^2 = R_c^2, \\ (P_{x3} - O_x)^2 + (P_{y3} - O_y)^2 = R_c^2; \end{cases} \quad (1)$$

where  $P_1, P_2, P_3$  are points of the profile of the internal cone ring;

$O$  is the center of the circle passing through the points  $P_1, P_2, P_3$ ;

$R_c$  is the radius of the circle passing through the points  $P_1, P_2, P_3$ .

To find the circles circumscribed around the search points, the distances from the center of the circle to all points of the profile are computed and compared with the radius of the circle  $R_c$ . If there are distances exceeding the radius  $R_c$ , then such combinations are not considered.

The circumscribed circles, which has the minimum radius  $R_{c\_min}$ , is chosen of all. It has a radius of the mating surface of the external ring, and the position of the center of the circle is the position of the center of the axis of the inner surface of the external ring. To find the displacement of the external ring along the z-axis (for example, focusing on the displacement of the upper end), it is necessary to calculate height of the defined section location. For this, the value of the applicator of the section of the internal ring is added to the value equal to:

$$\Delta_H = (R_{c\_min} - R_{upper\_end}) / \operatorname{tg}(\gamma / 2), \quad (2)$$

where  $R_{upper\_end}$  is the radius of the upper section of the external conical ring;

$\gamma$  is the angle at the vertex of the cone.

Thus, the position of the external ring along the z axis is calculated. It is possible to calculate the position of the centers during experimental studies graphically, taking into account the outrun of the surfaces of the external ring. The points of a circle with a diameter equal to the runout are calculated around the center of the mating surface of the external ring.

The more points of the surface of the internal ring are taken for the solution, the more accurate it will be, but the longer it will take to search through all the solutions. So, in the case of 120 setting points, the number of combinations which are to be considered equals  $C_{360}^3 = 120! / ((120-3)! \cdot 3!) = 280840$ . Accordingly, it is appropriate to use parallel computation and limit the number of search points used in the developed model.

### 3.3. Simulation of models of real parts surfaces

To calculate the assembly parameters using the models described above, information about the reference points and a mathematical description of the mating surfaces is necessary. Reverse engineering technology was used to obtain data on the surfaces of the conical rings under consideration, that involves measuring the surfaces by the coordinate measuring machine and further mathematical processing of the measurement data. The reverse engineering procedure applied to manufactured conical rings is presented in Figure 1.

During coordinate measurements of parts at the first stage their mathematical basing is made. The technology of mathematical basing of conical rings on CMM is based on measuring the surfaces by which you can specify the position of the coordinate axes of the part. Mathematical basing is a procedure consisting of calculation of the location of the part coordinate system (PCS) using the points of the base elements of the part previously measured in the machine coordinate system (MCS) to the subsequent

transformation of the coordinates of the points of other component parts from the MCS to the PCS. Thus, mathematical basing means calculating the optimal transformation matrix  $M$  which ensures the best fit of the measured points with the corresponding nominal points. The matrix contains three components of displacement along the coordinate axes  $\Delta X$ ,  $\Delta Y$  and  $\Delta Z$ , as well as three components of rotation around the coordinate axes  $\Delta\Phi$ ,  $\Delta\Theta$  and  $\Delta\Psi$ .

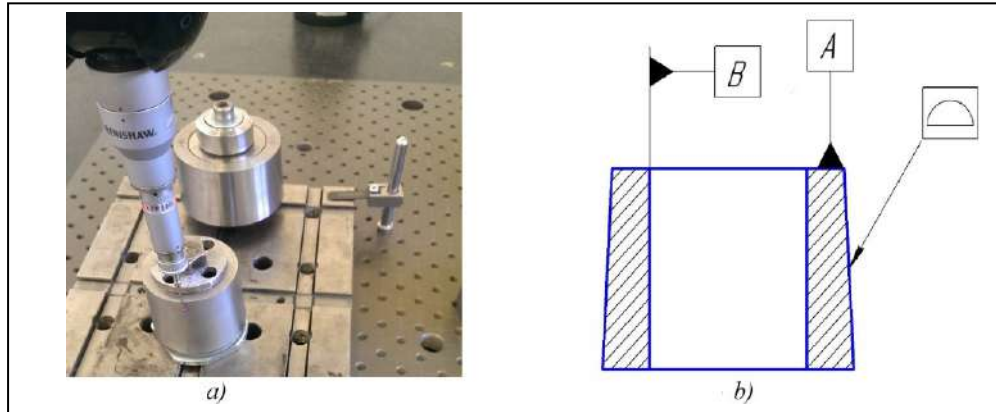


Fig. 1. Basing of the conical ring a) measurement by the CMM; b) scheme of conical ring basing.

Position of the main axis, the direction of the second axis and the center of the coordinate system are to be determined in the process of determination of the part coordinate system, the. For considering parts the direction of the rotation axis is defined by the butt end A (figure 1). The normal vector of the plane is collinear to the axis of the ring. The hole on the small ring and the outer surface of the external ring determine the centers of the axes. The hole points are measured in one section (base B). After measurement, the replacement element "circumference" is inscribed into the cloud of measured points. according to the method of least squares (LSM) The axis of the ring moves to the center of the circle. The cylinder axis is the axis of rotation. At the intersection of the end plane and the axis, the origin point is specified.

The basing procedure described above is carried out in two stages. At the first stage, the elements are measured manually by the smaller number of points (draft basing). At the second stage, more points are measured automatically that is clean basing.

Nevertheless, the coordinate system constructed on surfaces A and B may be out of the coordinate system of the mating surface, due to processing errors (non perpendicularity to the end, displacement of the center). To eliminate this error, the best comparison of the measured points of the mating surface with the mathematical model of the surface, along which coordinate system was processed and transformed, is performed. The mating surface is measured at list in three sections throughout the height and there are 300-360 points in each section.

The obtained parameters of matrix transformation  $M$  are used in further experiments for basing of the part, i.e. it is not required to produce the best combination of surfaces each time.

The coordinates of the measured section points of the mating surface  $P_{meas}$  are loaded into the MATLAB system.

The measured coordinates of the points contain various random components of errors caused by measurement errors, as well as by emissions from various factors atypically for the entire surface. For modern contact CMM such error is 0.5-4  $\mu\text{m}$  but when measuring complex curved surfaces or if condition is non-compliance with normal conditions this measurement error can be greater. During the processing of the measured data it is often necessary to remove the noise representing the inaccuracy of the measuring instrument [11, 12]. The problems associated with the fall in the image processing occur [13, 14, 15]. There are an extensive research and reduction in the processing data, [16, 17]. It is necessary to filter such errors for more accurate measurement results and obtaining adequate estimates of the parameters of the measured rings. One of the most effective tools for filtering random errors is smoothing splines [18].

The smoothed spline  $\hat{\varepsilon}$  of the set of deviations  $\varepsilon$  minimizes the equation:

$$p \sum_{i=1}^n w_i (\varepsilon(u_i) - \hat{\varepsilon}(u_i))^2 + (1-p) \int (D^2 \hat{\varepsilon}(u))^2 du \rightarrow \min, \quad (3)$$

where  $(u, \varepsilon)$  are approximated data (the point and magnitude of the deviation in it);

$p$  is the smoothing parameter  $p \in [0, 1]$ ;

$W$  is a vector of weights (it is taken as the vector of units).

Filtering with a parameter  $p$  equal to 0.99 was used to filter the errors of the measured ring point arrays. To carry out the filtration, it is necessary to calculate the radius vectors of the measured points in each section. The radius vector of the  $i$ -th point is calculated by the equation:

$$r_{Pi\_meas} = \sqrt{P_{xi\_meas}^2 + P_{yi\_meas}^2}. \quad (4)$$

The deviation of the form at the points of the section  $\delta_F$  is calculated as the difference between the radius vector of the point  $r_p$  and the nominal radius of the cone in the cross section  $r_{cone}$ :

$$\delta_F = r_p - r_{cone}. \quad (5)$$

After filtration, the average deviation of the form at each point of each measured section is calculated. The values of calculated deviations at points were used to calculate the points of a new surface (taking into account manufacturing errors). In the MATLAB system the NURBS surface was automatically constructed and saved as the an \*.igs file. The built surface was loaded in CAD system and the it was changed taking in account a new surface of the 3D model of the ring.

At the final stage, a remeasurement was carried out according to the procedure described above, in order to reduce the measurement errors that arise due to the discrepancy of the measured part to the reference model [19]. The measurements were re-processed according to foregoing sequence and the surface parameters were saved for modeling the assembly process.

#### 4. Results

To obtain experimental data on the assembly parameters, 20 measurements of the rings in the assembled form were made by a coordinate measuring machine. The external ring in each measurement was rotated by an angle about the axis. To find the center of the inner mating surface of the external ring, a least-squares circle was inscribed in 20 measured points (Figure 2).

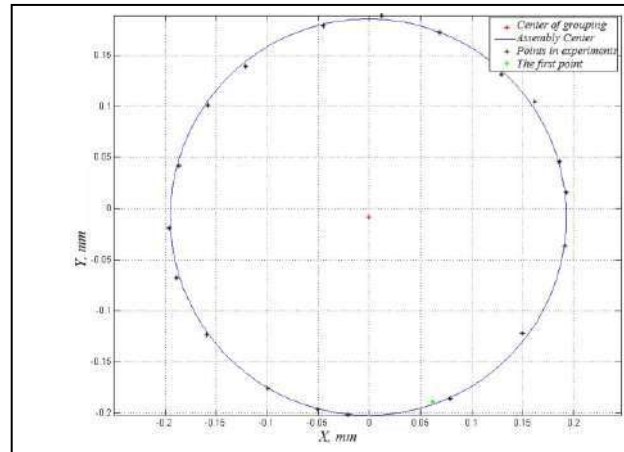


Fig. 2. A graphical solution for finding the center of mating and outrun.

The diameter of the circle characterizes the outrun of the external cone ring relative to the internal ring (Table 1). The nonparallelism of the axes of the mating cone rings results in additional error, which during the experiments was 0.005-0.014 mm.

Thus, the parameters characterizing the mating of the conical ring with the deviation of the form and the cone ring with the deviation of the location (outrun of the conical surfaces) are obtained.

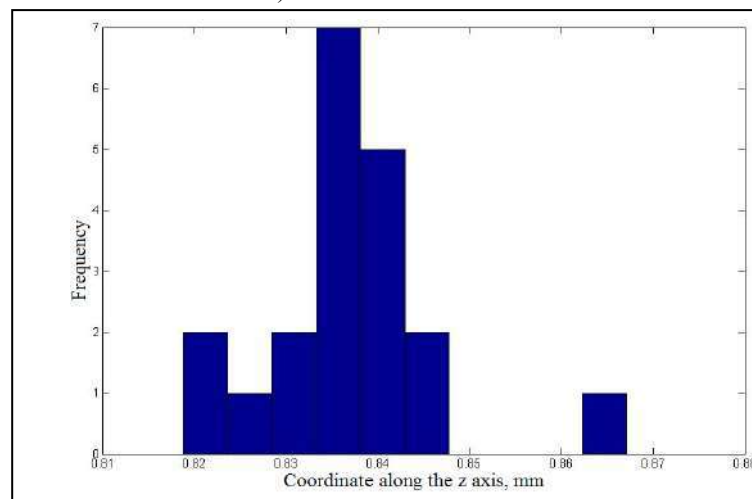


Fig. 3. Histogram of the coordinates along z.

At the same time, during assembly process various positions of the centers along the z axis were performed, that is explained by the action in the mating of rings of frictional forces, which are characterized by non-parallel axes during assembly process. Consequently, the action of friction forces leads to the appearance of an error relative to the calculated mating center. A histogram of the distribution of center values along the z axis is shown in Figure 3.

Further, due to fluctuations along the z axis during assembly, the calculated positions of the centers have difference from the ideal circle inscribed in them, which also means the possible magnitude of the errors in determination of the assembly center. Figure 4 shows histograms of the distances of oscillations of the radius vectors of points relative to the radius of an ideal circle determined by the method of least squares.

After experiments the ring assembly parameters were calculated using the MSF models and combinatorial search. To calculate the assembly parameters, the points of the surface of the internal ring calculated using reverse engineering procedures were used. The points of the complementary mating surface of the external ring were set as points of an ideal conical surface.

A comparison of the resulting positions of the external conical ring in the assembly determined by MSF, the combinatorial search model and in physical experiments, is given in Table 2.

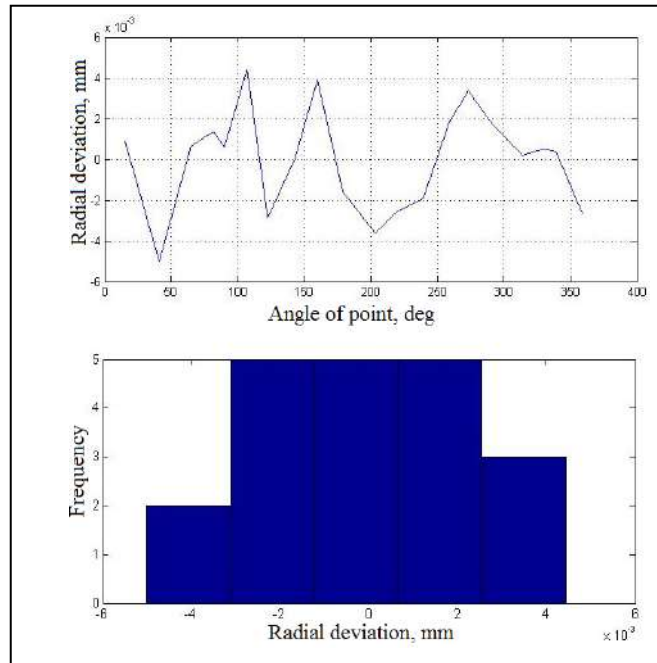


Fig. 4. Oscillations of the radius of the measured points of the centers.

Table 2. Comparison of the resulting coordinates of the position of the parts centers during assembly.

Coordinates of mating center	Experimental data	MSF	Combinatorial search model
Along X, mm	-0,0009	-0,0018	-0,0016
Along Y, mm	-0,0083	-0,0059	-0,0069
Along Z, mm	0,8188-0,8671	0,9471	0,9353

Comparing the results of the experimental solution and the solutions by simulation, it can be noted that both methods have rather high accuracy of the search for a solution along  $X$  and  $Y$  axes, and deviations don't exceed  $2 \mu\text{m}$  in the case of SMEs and  $1 \mu\text{m}$  using the combinatorial search model. Coordinates along the axis  $Z$  have larger deviations, since deviations of micrometers in the plane  $XOY$  result in deviations of tens of micrometers along  $Z$ . The discrepancies are connected, first of all, with the fact that the frictional force of the surfaces was not taken into account in modeling and there is no misalignment of axes of cones during assembly. Deformations are also possible on contact of surfaces, but for such details as conical rings they are not significant.

## 5. Conclusion

The paper describes two mathematical models: model for combining surfaces and the combinatorial search model, that allow to forecast the geometric parameters of assembly units. The developed models are useful because they may forecast the relative positions of parts, which have geometric deviations. Comparison of simulation results with experimental data on the example of the assembly of conical rings showed fine precision in determination of the position of the center of the external ring axis. It is shown that the calculation of the errors of the assembly parameters can be made from the results of measurements of geometry by modern CMMs, which allow obtaining data on the points of surfaces in electronic form. The developed models can be used to forecast the assembly of critical parts of aviation equipment, such as shafts, discs, blades.

## Acknowledgement

This work was supported by the Ministry of Education and Science of the Russian Federation in the framework of the implementation of the Program "Research and development on priority directions of scientific-technological complex of Russia for 2014–2020.

## References

- [1] Ermakov AI, Shklovets AO, Popov GM, Kolmakova DA. Investigation of the effect of the gas turbine compressor supports on gas flow circumferential nonuniformity. *Research Journal of Applied Sciences* 2014; 9(10):684–690.
- [2] Popov GM, Goryachkin ES, Baturin OV, Kolmakova DA. Development of recommendations on building of the lightweight calculation mathematical models of the axial turbines of gas turbine engines. *International Journal of Engineering and Technology* 2014; 6(5): 2236–2243.
- [3] Smelov VG, Sotov AV, Agapovichev AV. Research on the possibility of restoring blades while repairing gas turbine engines parts by selective laser melting. *IOP Conference Series: Materials Science and Engineering* 2016; 140: 012019.

- [4] Alexeev VP, Balyakin AV, Khaimovich AI. Influence of the direction of selective laser sintering on machinability of parts from 316L steel. IOP Conference Series: Materials Science and Engineering 2017;177(1): 012120.
- [5] Kolmakova DA, Baturin OV, Popov GM. Effect of manufacturing tolerances on the turbine blades. ASME 2014 Gas Turbine India Conference, GTINDIA 2014; 1–10.
- [6] Morozov AA, Skidanov RV. Rotation of microturbine in complex vortex beams. Computer Optics 2013; 37(2): 203–207. (in Russian)
- [7] Ganchevskaya SV, Skidanov RV. The microturbine rotation by not circular light beam formed by vortex Axicon . CEUR Workshop Proceedings 2016; 1638: 24–31.
- [8] Bolotov MA, Pechenin VA, Murzin SP. Method for uncertainty evaluation of the spatial mating of high-precision optical and mechanical parts. Computer Optics 2016; 40(3): 360–369. DOI: 10.18287/2412-6179-2016-40-3-360-369.
- [9] Besl PJ, McKay ND. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 1992; 14(2): 239–256.
- [10] Pierce RS, Rosen D. Simulation of mating between nonanalytical programming formulation . Journal of Computing and Information Science in Engineering 2007; 7(4): 314–321.
- [11] Soifer VA, Kupriyanov AV. Analysis and recognition of the nanoscale images: conventional approach and novel problem statement. Computer Optics 2011; 35(2):136–144.
- [12] Myasnikov V, Popov SB, Sergeev VV, Soifer VA. Computer Image Processing. Part I: Basic concepts and theory. VDM Verlag, 2009; 296 p.
- [13] Kazanskiy NL, Khonina SN, Skidanov RV, Morozov AA, Kharitonov SI, Volotovskiy SG. Formation of images using multilevel diffractive lens. Computer Optics 2014; 38(3): 425–434.
- [14] Borodin SA, Volkov AV, Kazanskiy NL. Device for analyzing nanoroughness and contamination on a substrate from the dynamic state of a liquid drop deposited on its surface. Journal of Optical Technology 2009; 76(7): 408–412.
- [15] Kopenkov VN, Sergeev VV, Timbai EI. Regression restoration methods as applied to solve the problem of multidimensional indirect measurements. Pattern Recognition and Image Analysis 2011; 21(3): 501–504.
- [16] Fleishman S, Drori I, Cohen-Or D. Bilateral mesh denoising. ACM Transactions on Graphics 2003; 22(3): 950–953.
- [17] Abdul-Rahman HS, Scott PJ, Jiang XJ. Freeform surface filtering using the lifting wavelet. Precision Engineering 2013; 37(1): 187–202.
- [18] De Boor C. A Practical Guide to Splines (Revised Edition). New York: Springer, 2001; 348 p.
- [19] Mayer J, Mir Y, Trochu F, Vafaeseefat A, Balazinski M. Touch probe radius compensation for coordinate measurement using kriging interpolation. Proceedings of the Institution of Mechanical Engineers. Part B: Journal of Engineering Manufacture 1997; 211(1): 11–18.

# Probability-theoretical model for product assembly parameters assessment

N.V. Ruzanov<sup>1</sup>, M.A. Bolotov<sup>1</sup>, V.A. Pechenin<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

The proposed work provides a model for estimating the assembly parameters of the products by the example of the cone ring assembly process. There was carried out the simulation of the process of parts mating along the surfaces with form deviation. As a result of Monte Carlo simulation, there were obtained statistical characteristics of the parameters of the assembled units depending on various initial positions of the parts being assembled. The proposed model could be used to assess the accuracy of the products assembled in the aircraft industry.

*Keywords:* assembly parameters; mating of components; Monte Carlo method; probabilistic assessment; probability density

---

## 1. Introduction

The result of manufactured products assembly depends on the actual shape of the parts used, their initial location and the assembly process. The errors of the actuating mechanisms used during parts manufacture have a significant effect on the accuracy of the manufactured products [1], so that even the parts of the same type do not have actual identical shape. Moreover, even for a single batch of products it is hardly achievable to ensure the exact matching of assembly conditions for various items manufacture.

Assembly parameters assessment will allow determining an achievable accuracy of product manufacturing, which, in turn, will enable to solve a number of problems:

1. Determine the percentage of products meeting process requirements [2].
2. Determine rational tolerances for product parameters [3].
3. Identify critical factors affecting the quality of product assembly (assembly deviation, parts prepositioning or tool selection for parts assembly) [4-7]

An assessment of the assembly parameters of products can be obtained through various approaches. One of them is based on accumulation and analysis of results of the manufactured products inspection phase. Another approach is based on numerical simulation of the product assembly process and subsequent analysis of the obtained results. The use of production statistics requires considerable human and material resources, and therefore is hardly feasible. The first approach being rather difficult to realize is the cause why math modeling methods are widely used to solve the specified problem.

In this paper, the authors present a model for estimation of product assembly parameters, based on assembly process simulation by numerical methods. Moreover, there is given an example of using the proposed model for estimation of assembly parameters of cone rings that are widespread products in the aircraft industry.

## 2. Model description

Product characteristics depend not only on parts comprising it but also on assembly procedures. Assembling of complex products consisting of many parts is a multi-criteria task that takes into account the size of all the components, their mutual arrangement in the finished product, and the alignment procedure [8].

Today, the parts' quality issues are understood deeply, so in order to improve quality characteristics of the product further, the researches aimed at studying the product assembly procedure are becoming increasingly popular. The complexity of implementing multiple product assembly leads to the fact that the majority of such researches is based on the employment of numerical simulation methods in the assembly process.

One of the features of parts assembly in the aircraft industry is the deformation of parts of the product. So as to determine the condition of parts of the product assembled, ANSYS application is used, which enables us to calculate the strength of the parts and assemblies, to solve the problems of gas and hydrodynamics [9]. Such approaches have high computational costs, so in order to simplify the solution of the assembly problem, many researchers consider the mating parts as absolutely rigid bodies.

The authors [10] consider mating two parts along plane surfaces, having form deviation. Researchers proposed a mathematical model simulating the assembly along the planar surfaces; and the result of using of such model is calculation of the clearance between the surfaces in product assembled.

Paper [11] is also devoted to the problem of parts mating along the planar surfaces. The authors of this work suggested a model for describing the deviation of the shape of the specified surfaces and considered the result of the parts mating with various shape deviations of the planar surfaces.

Most of the works focus on the simulation of the assembly process without setting up a formal problem. In paper [12] the authors formalize the product assembly problem and suggest using such concepts as the initial assembly conditions, the assembly quality assessment function, and the assembly sequence function.

The proposed work considers the assessment of the assembly parameters of the product obtained from various assembly process models.

The mutual arrangement of the parts is a fundamental assembly parameter of the product as other geometric characteristics of the finished product can be derived from it (for example, out-of-true running, out-of-flatness, uneven clearance, etc.). The

methods for positioning coordinate systems that are bound to these parts are well suited to describe the mutual arrangement of the parts.

Let us consider a product that is assembled from two parts  $K_1$  and  $K_2$ . The product coordinate system  $R$  is compatible with the design coordinate system  $R_1$  of the part  $K_1$ . With such a transformation, the part  $K_1$  will be stationary relative to the coordinate system  $R$ , and the assembly will be carried out by moving the part  $K_2$ . The assembled state of the product can be described by the position of the design coordinate system  $R_2$ . The errors in the assembly process will result in the onset of a set of assembled states  $\Omega$ .

$$\Omega = \left\{ \omega : \omega = \begin{pmatrix} dx \\ dy \\ dz \\ \alpha \\ \beta \\ \gamma \end{pmatrix} \right\}, \quad (1)$$

where  $dx, dy, dz$  is the displacement of the coordinate origin of system  $R_2$  relative to the origin of system  $R_1$ ;

$\alpha, \beta, \gamma$  are the angles of turn of basis vectors of system  $R_2$  relative to vectors  $R_1$ .

The elements of this set describe the mutual arrangement of the parts in the assembled product. To solve the problem of evaluating the assembly parameters of the products, it is necessary to get the parameters for this set. The use of the Monte Carlo method enables us to determine the approximate value of this parameters. According to this method, the first step is to perform a multiple numerical simulation of the assembly of the product and save the simulation results. The second step is to investigate the obtained simulation results and calculate the required parameters.

We will use the following values to estimate the set of assembled states: the mean value, root-mean-square deviation and probability density.

To calculate the average value, let us use the following formula (2):

$$\bar{\omega} = \begin{pmatrix} \bar{dx} \\ \bar{dy} \\ \bar{dz} \\ \bar{\alpha} \\ \bar{\beta} \\ \bar{\gamma} \end{pmatrix}, \quad \bar{a} = \frac{1}{n} \sum_{i=1}^n a_i, \quad (2)$$

The root-mean-square deviation is calculated with the formula (3):

$$\sigma = \begin{pmatrix} \sigma_{dx} \\ \sigma_{dy} \\ \sigma_{dz} \\ \sigma_{\alpha} \\ \sigma_{\beta} \\ \sigma_{\gamma} \end{pmatrix}, \quad \sigma_a = \sqrt{\frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})^2}, \quad (3)$$

Obtaining an analytic expression for a probability density function is a complicated task, so we can use numerical methods to evaluate the state with the approximate values of the function with the required accuracy, based on the distribution histograms and their possible approximation.

Distribution histogram method provides empirical estimates  $K_2$  of the density of distribution of the random value [13]. The algorithm for obtaining the distribution density histogram is shown in Fig. 1.

To create a histogram, the observed range of the random variable is divided into several intervals, and then the number of the random value hits in each interval is calculated. The Sturges' rule (4) is used to determine the number of the intervals:

$$n = 1 + \lceil \log_2 N \rceil, \quad (4)$$

The next step is standardization of the received values to meet the condition (5)

$$\int_{\Omega} f(\omega) d\omega = 1, \quad (5)$$

The final step is to approximate the probability density function on the basis of the midpoint of the intervals and values calculated in the previous step.

The mean and root-mean-square deviation values, the approximate value of the probability density function describe a set of assembled states and can be used to solve further tasks.

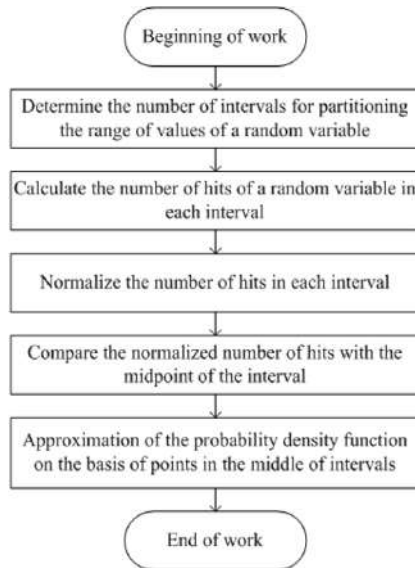


Fig. 1. Algorithm for numerical calculation of probability density. Algorithm for numerical definition of the probability density.

### 3. Simulation results

To test the proposed methodology, there has been carried out the evaluation of the assembly parameters of the parts that have the cone surfaces. Mating of such surfaces is widely used in the aviation industry and the characteristics of the entire product depend on the assembly quality of these parts.

The mathematical model of the cone surface of a part can be represented in a parametric form (6):

$$\left\{ \begin{array}{l} \Delta F = \Delta F(u, v) \\ x(u, v) = \left( \frac{R_2 - R_1}{H} * \left( \frac{R_1 * H}{R_2 - R_1} + H - u \right) + \Delta F(u, v) \right) * \cos(v) \\ y(u, v) = \left( \frac{R_2 - R_1}{H} * \left( \frac{R_1 * H}{R_2 - R_1} + H - u \right) + \Delta F(u, v) \right) * \sin(v) \\ z(u, v) = u \end{array} \right. , \quad (6)$$

where  $\Delta F$  is the function of deviation of the shape of the actual part from its nominal value;

$R_1, R_2, H$  are the radii of the cone surface and its height;

$u, v$  - surface parameters

$x(u, v), y(u, v), z(u, v)$  - surface point coordinates

A model of a part that has a cone surface is shown in fig. 2.

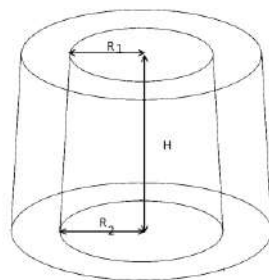


Fig. 2. A model of a part with a cone surface.

Fig. 3 shows a mechanical system model consisting of two parts that have cone surfaces. For parts  $K_1$  and  $K_2$  the local design coordinate systems  $R_1$  and  $R_2$  are set. Parts mating is performed along the surfaces  $B_1$  and  $B_2$ . Each surface is set in the local coordinates of the part and is described by the formula (6).

As an assembly procedure, let us consider a translational movement of the second part. This procedure simulates the process of cone rings assembling under press-in technology. The part  $K_2$  is lowered onto part  $K_1$  until the parts are in contact. For simplicity, let us consider the parts to be absolutely rigid, so that their deformation can be left out of account. In the course of the work, 1000 experiments were carried out on the modeling of the mating of cone surfaces and the amount of data required to carry out the evaluation was collected. Fig. 4 demonstrates the assembled states of the mechanical system.



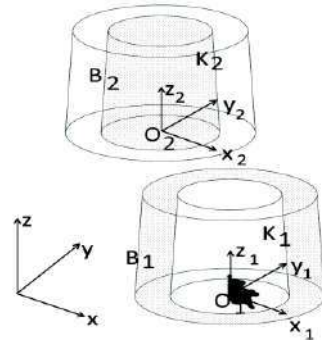


Fig. 3. A mechanical system model for assembling two cone rings.

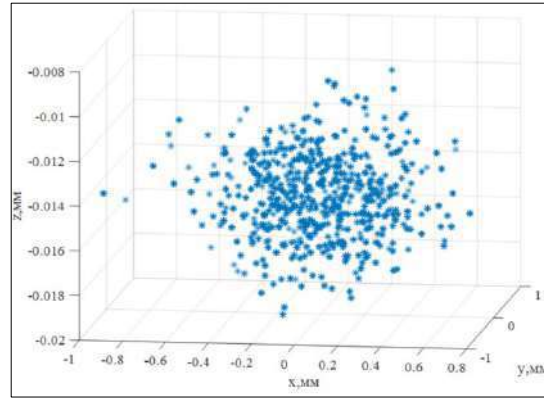


Fig. 4. The assembled states of the mechanical system for the various initial simulation conditions.

The statistical characteristics of the obtained set are specified in Table 1

Table 1. Parameters of the set of assembled states $\Omega$ .			
Parameter	X	Y	Z
Mean value	-0.0066	-0.0087	-0.0147
Root-mean-square deviation	0.2576	0.2557	0.0018

Fig. 5, 6 and 7 demonstrate histograms of the distribution of assembled states along the corresponding coordinates.

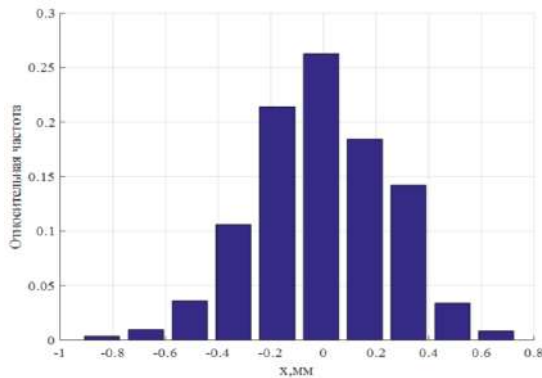


Fig. 5. Distribution density of the x coordinate of the system assembled state.

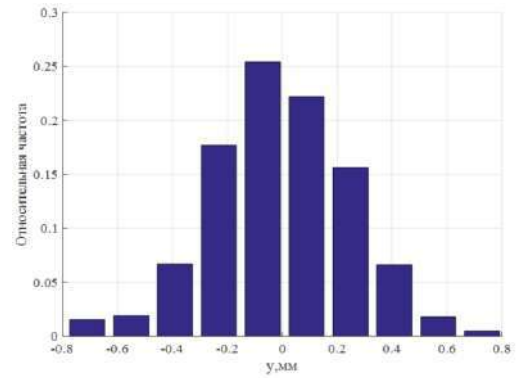


Fig. 6. Distribution density of the y coordinate of the system assembled state.

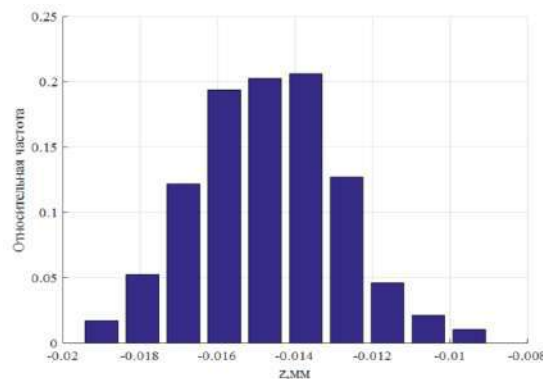


Fig. 7. Distribution density of the z coordinate of the system assembled state.

Based on the simulation results shown in Fig. 4, there was obtained a probability density histogram for the distribution of the assembled states of the system. The obtained chart is an approximate value of the probability density function for the assembled state of the mechanical system. Fig. 8, 9, 10 demonstrate the approximate value of a section of this function for various  $z$  heights.

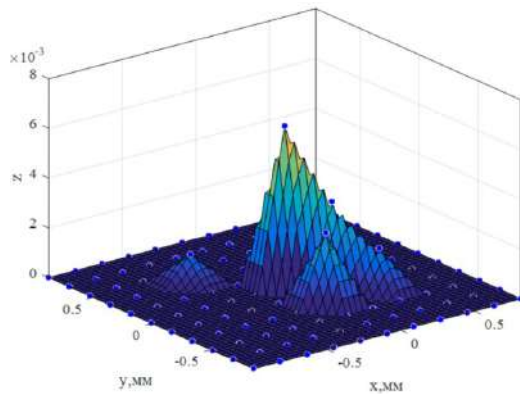


Fig. 8. Section of an approximate probability density value at height  $z = -0.0189$ .

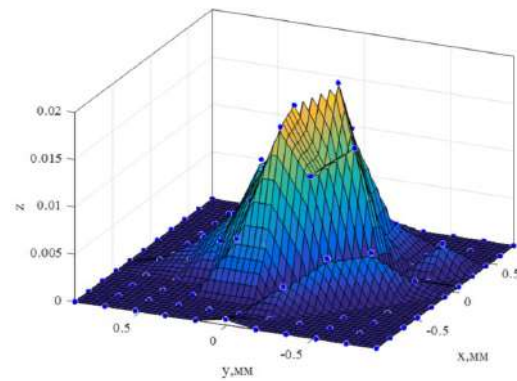


Fig. 9. Section of an approximate probability density value at height  $z = -0.0154$ .

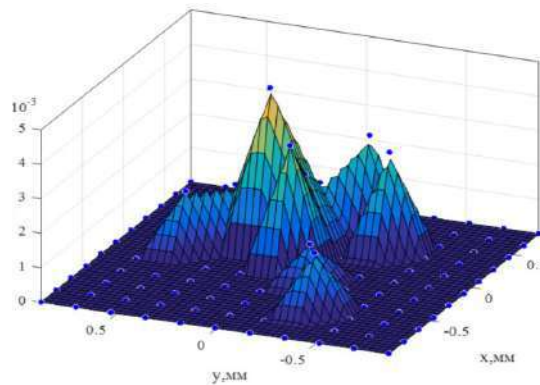


Fig. 10. Section of an approximate probability density value at height  $z = -0.0108$ .

The obtained values describe a set of assembled states of the product and can be used to solve further tasks.

#### 4. Conclusion

Assessment of the assembly parameters of the products allows to solve a number of important production tasks of the aircraft industry related to the efficiency and the quality of the manufacturing process of the parts. Such an estimation is difficult to implement without processing the product assembly results. One way to get the geometric parameters of an assembled product is numerical simulation of the process of its assembly. The results of multiple simulations can be used to evaluate some assembly parameters.

In the framework of this research, there was proposed a model for estimating the mutual arrangement of parts of the product, which is one of the main assembly parameters. It enables us to obtain many other geometric parameters, such as out-of-true running, out-of-flatness, uneven clearance, etc. The proposed estimation is based on the calculation of the parameters of a set of assembled products: mean and root-mean-square deviation values, approximate probability density. These parameters may be useful for other production tasks. Mean and root-mean-square deviation values for the mutual arrangement of parts can be used to solve the problem of determining rational tolerances for product parameters. The probability density function can be applied to determine the percentage of products that meet the technology requirements. All of these parameters can be used to increase the efficiency of the technological process by identifying the most critical factors influencing the final product quality.

The proposed model was used to assess the mutual arrangement of the parts that have cone surfaces. The next step is to test the results of the numerical simulation in practice.

#### Acknowledgments

This work was supported by the Ministry of Education and Science of the Russian Federation in the framework of the implementation of the Program “Research and development on priority directions of scientific-technological complex of Russia for 2014–2020”.

**References**

- [1] Ganchevskaya SV, Skidanov RV. A technique for optimizing the structure of an optical trap to rotate multiple microobjects. *Optical Memory and Neural Networks (Information Optics)* 2016; 25(3): 160–167.
- [2] Grechnikov FV, Khaimovich AI. Development of the requirements template for the information support system in the context of developing new materials involving big data. *CEUR Workshop Proceedings* 2015; 1490: 364–375. DOI: 10.18287/1613-0073-2015-1490-364-375.
- [3] Demin FI, Pronichev ND, Shitarev IL. *Technology of turbine engines basic parts manufacture*. Samara: Samara State Aerospace University Publishing, 2012; 324 p.
- [4] Vdovin RA, Smelov VG, Bolotov MA, Pronichev ND. Paths of Improving the Technological Process of Manufacture of GTE Turbine Blades. *IOP Conference Series: Materials Science and Engineering* 2016; 142: 1–8. DOI: 10.1088/1757-899X/142/1/012073.
- [5] Smelov VG, Sotov AV, Agapovichev AV. Recovery technology features of aerospace parts by layering synthesis. *Key Engineering Materials* 2016; 684: 316–322. DOI: 10.4028/www.scientific.net/KEM.684.316.
- [6] Kolmakova DA, Baturin OV, Popov GM. Effect of manufacturing tolerances on the turbine blades. *ASME 2014 Gas Turbine India Conference, GTINDIA, 2014*; 1–10.
- [7] Kazansky NL, Stepanenko IS, Haimovich AI, Kravchenko SV, Byzov EV, Moiseev MA. Injectional multilens molding parameters optimization. *Computer Optics* 2016; 40(2): 203–214. DOI: 10.18287/2412-6179-2016-40-2-203-214.
- [8] Kirilin A, Shakhmatov E, Soifer V, Akhmetov R, Tkachenko S, Prokofev A. Small satellites "aIST" constellation - design, construction and program of scientific and technological experiments. *Procedia Engineering* 2015; 104: 43–49. DOI: 10.1016/j.proeng.2015.04.095.
- [9] Zubanov V, Shabliy L, Krivcov A. Centrifugal kerosene pump CFD-modeling. *Research Journal of Applied Sciences* 2014; 9(100): 629–634. DOI: 10.3923/rjasci.2014.629.634.
- [10] Pierce RS, Rosen D. Simulation of mating between nonanalytical programming formulation. *Journal of Computing and Information Science in Engineering* 2007; 7(4): 314–321. DOI: 10.1115/1.2795297.
- [11] Samper S, Adragna P-A, Favreliere H, Pilllet M. Modeling of 2D and 3D assemblies taking into account form errors of plane surfaces. *Journal of Computing and Information Science in Engineering* 2009; 9(40): 1–12.
- [12] Bolotov MA, Pechenin VA, Murzin SP. Method of estimating the indeterminacy of the space mating of precision optical and mechanical parts. *Computer Optics* 2016; 40(3): 360–369. DOI: 10.18287/2412-6179-2016-40-3-360-369.
- [13] Eliseeva II, Iuzbashev MM. *General theory of statistics*. M.: Finance and Statistics, 2004; 656 p.

# About the attractor-repeller points during the descent of an asymmetric spacecraft in the atmosphere

V.V. Lyubimov<sup>1</sup>, V.S. Lashin<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The aim of this study is to analyze the resonant attractor-repeller points during the atmospheric descent of a spacecraft with small asymmetry. The mathematical simulation of spacecraft rotational motion uses an approximate non-linear system of equations obtained by the method of integral manifolds. Application of the averaging method and the Lyapunov method makes it possible to obtain realization conditions of attractor-repeller points on non-resonance parts of the motion. By analyzing of the said conditions, we have identified specific cases when the principal resonance is either an attractor point or a repeller point.

*Keywords:* resonance; attractor; repeller; averaging; spacecraft; atmosphere; asymmetry

---

## 1. Introduction

Various resonance phenomena in the problem of uncontrolled descent of a spacecraft with a small asymmetry in the atmosphere are explored in [1-2, etc.]. In particular, the disturbing moments of mass-aerodynamic asymmetry can lead to the evolution of angular velocity of the spacecraft to the resonance values [3-5]. In this case, the non-resonance evolution of the angular velocity of the asymmetric spacecraft is the secondary resonance effect [6]. The external stability of resonance is considered in the problems of perturbed rotational motion of the asymmetric spacecraft in the atmosphere or satellite in orbit [5,7]. The realization of the main resonance leads to significant increase of the angle of attack. In practice, this can lead to emergency situation during the deployment of the parachute system of a spacecraft. It is known that the external stability of the resonance contributes to the evolution of the variables of the dynamic system to resonant values (resonant attractor). Therefore, the study of the phenomenon of external stability of resonance is an important practical task. This phenomenon arises when the condition of the external stability of the resonance is satisfied. It should be noted that the investigation of the resonant attractor and repeller is supposed to be performed in a more general form in comparison with the results presented in [5].

## 2. Problem statement

Let us assume that spacecraft is a solid body in the form of a cone combined with a spherical surface. Let the axis  $OX$  is the symmetry axis of the cone. In the process of entry into atmosphere, a spacecraft is directed its conical surface along of the air flow. During the atmospheric descent, the spacecraft engages in the precessional motion. It is known that a spacecraft receives small angular momentum when undocked from the base orbital module [8]. In this case, the angular momentum lead to the formation of the components of the angular velocity of the spacecraft  $\omega_x(0)$ ,  $\omega_y(0)$ ,  $\omega_z(0)$ . These components of the angular velocity are recorded of the spacecraft body-fixed coordinate system  $OXYZ$ . Suppose that these angular velocities are initial when the spacecraft enters into the atmosphere. The origin of the coordinate system  $O$  is located at the center of mass of the spacecraft. In [9] it is shown that the resonance values of the angular velocity  $\omega_x$  can be determined on the basis of the method of integral manifolds [10]. The values of angular velocity  $\omega_x$  corresponding to the principal resonance are defined as follows [5]:  $\omega_x^r = \pm \omega / (1 - I_x)^{1/2}$ . Here  $\omega = (-m_{zn} q S L c t g \alpha / I)^{1/2}$  is the angular velocity,  $m_{zn}$  is the restoring moment coefficient for the angle of attack  $\alpha$ ,  $q$  is the dynamic pressure,  $S$  is the area of the maximum cross section of a spacecraft,  $L$  is the length of a spacecraft,  $\bar{I}_x = I_x / I$ ,  $I = (I_y + I_z) / 2$ ,  $I_x, I_y, I_z$  are the principal moments of inertia of a spacecraft. It is known that the principal resonance has the greatest influence on the evolution of slow variables on the non-resonant parts of the motion, compared with resonances of higher orders. The aim is to study the realization conditions of resonant attractor and resonant repeller in case of atmospheric descent of a spacecraft with small aerodynamic-inertial asymmetry. Let the attack angle take arbitrary values. We apply the method of averaging and the Lyapunov method to the research of attractors and repellers.

## 3. Methods

### 3.1. Mathematical model

The approximate non-linear system of equations of motion of a spacecraft with small aerodynamic-inertial asymmetry, describing the motion of a spacecraft relative to the center of mass has the form [5]:

$$\bar{I}_x \frac{d\omega_x}{dt} = \varepsilon m^\Delta \omega_{1,2}^2 t g^2 \alpha \cos(2\theta + 2\theta_3), \quad (1)$$

$$\begin{aligned} \frac{F_a}{4\omega_a^2} \frac{d\alpha}{dt} &= -\varepsilon \frac{\omega \operatorname{tg} \alpha}{2\omega_a^2} \frac{d\omega}{dt} - \varepsilon \frac{m^A}{2\omega_a} \cos(\theta + \theta_1) - \\ &- \varepsilon \frac{\omega_{1,2} \operatorname{tg} \alpha}{4\omega_a^2} \left[ (10 + \bar{I}_x) \omega_x \omega_{1,2} - 2(2 + \bar{I}_x) \omega_x^2 \right] m^A \cos(2\theta + 2\theta_3) - \\ &- \varepsilon \frac{\omega_{1,2} \operatorname{tg} \alpha}{4\omega_a^2} \left[ (\operatorname{tg}^2 \alpha - 4) \omega_{1,2}^2 \right] m^A \cos(2\theta + 2\theta_3), \end{aligned} \quad (2)$$

$$\frac{d\theta}{dt} = \omega_x - \omega_{1,2}, \quad (3)$$

$$\frac{d\omega}{dt} = \varepsilon \frac{\omega}{2q} \frac{dq}{dt}. \quad (4)$$

Here  $\varepsilon$  is the small parameter,  $\theta = \varphi - \pi/2$ ,  $\varphi$  is aerodynamic roll angle;  $m^A$ ,  $m^\Delta$ ,  $\theta_1$ ,  $\theta_3$  are functions that characterize the values and relative positions of the aerodynamic and inertial asymmetries of the spacecraft,

$$\begin{aligned} m^A &= \sqrt{(m_1^A)^2 + (m_2^A)^2}, \quad m_1^A = -\frac{(1 + \bar{I}_x) \omega_x - 3\omega_{1,2}}{2\omega_a} \frac{\omega^2}{m_{zn}} (m_y^f - C_x \bar{\Delta z}) \operatorname{tg} \alpha - \frac{\omega_{1,2} \omega^2 \operatorname{tg}^2 \alpha}{2\omega_a m_{zn}} (C_{yn} \bar{\Delta z}) \\ &+ \frac{\bar{I}_{xz}}{2\omega_a} \left[ \omega_x \omega_{1,2} (\omega_x + \omega_{1,2} \operatorname{tg}^2 \alpha) - \omega_x^2 (\omega_x - 2\omega_a) \right], \quad m_2^A = -\frac{(1 + \bar{I}_x) \omega_x - 3\omega_{1,2}}{2\omega_a} \frac{\omega^2}{m_{zn}} (m_z^f + C_x \bar{\Delta y}) \operatorname{tg} \alpha + \\ &+ \frac{\omega_{1,2} \omega^2 \operatorname{tg}^2 \alpha}{2\omega_a m_{zn}} (C_{yn} \bar{\Delta y}) \pm \frac{\bar{I}_{xy}}{2\omega_a} \left[ \omega_x \omega_{1,2} (\omega_x + \omega_{1,2} \operatorname{tg}^2 \alpha) - \omega_x^2 (\omega_x - 2\omega_a) \right], \quad \sin \theta_1 = m_1^A / m^A, \quad \cos \theta_1 = -m_2^A / m^A, \quad \omega_a = \sqrt{\bar{I}_x \omega_x^2 / 4 + \omega^2}; \\ m^\Delta &= \sqrt{\bar{I}_{yz}^2 + \bar{\Delta I}^2}, \quad \sin 2\theta_3 = \bar{\Delta I} / m^\Delta, \quad \cos 2\theta_3 = -\bar{I}_{yz} / m^\Delta, \quad \bar{I}_{xy} = I_{xy} / I, \quad \bar{I}_{xz} = I_{xz} / I, \quad \bar{I}_{yz} = I_{yz} / I, \quad \bar{\Delta I} = \Delta I / I \end{aligned}$$

are dimensionless moments of inertia of a SC,  $\omega_{1,2} = \frac{\bar{I}_x \omega_x}{2} \pm \omega_a$ ;  $\omega_x - \omega_{1,2}$  is the resonant ratio of frequencies;  $C_x, C_{ym}$  are the aerodynamic coefficients;  $m_y^f, m_z^f$  are the coefficients of small moments caused by asymmetric shape of the spacecraft;  $\bar{\Delta y} = \Delta y / L$ ,  $\bar{\Delta z} = \Delta z / L$ ;  $\Delta y, \Delta z$  are small displacements of the center of mass of the spacecraft;  $F_a = F_a(\omega_x, \alpha, \omega)$  is the known function of slow variables [5]. In equations (1)-(4) we consider the principal resonance, corresponding to the following condition:  $\Delta = \omega_x - \omega_{1,2} \cong 0$ . There are signs “ $\pm$ ” and “ ” in the equations (1)-(3). We assume in the said equations that the upper sign is selected when  $\omega_x > 0$ , and the lower sign is selected when  $\omega_x < 0$ . In the numerical simulation of spacecraft motion, the system of equations (1)-(4) should be considered together with the system of three differential equations for slowly varying of the center of mass parameters: the local flight-pass inclination angle  $\vartheta(t)$ , the spacecraft airspeed  $V(t)$  and the spacecraft altitude  $H(t)$  [1].

### 3.2. Averaging and analysis of resonant attractor

After using the method of averaging on non-resonant parts of a spacecraft motion we obtain [5]:

$$\begin{aligned} \left\langle \frac{d\omega_x}{dt} \right\rangle &= \varepsilon^3 \left\{ \frac{\bar{m}^A g_3 g_1}{\Delta^3} \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \frac{\partial \Delta}{\partial \alpha} \right) - \frac{\bar{m}^A g_3}{\Delta^2} \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \frac{\partial g_2}{\partial \alpha} \right) + \right. \\ &+ \left. \frac{3(\bar{m}^A g_3)^2}{\Delta^4} \left( \Delta \frac{\partial \Delta}{\partial \alpha} \frac{\partial g_2}{\partial \alpha} - g_2 \left( \frac{\partial \Delta}{\partial \alpha} \right)^2 \right) \right\} \frac{m^A \cos(2\theta_1 - 2\theta_3)}{8}, \end{aligned} \quad (5)$$

$$\begin{aligned} \left\langle \frac{d\alpha}{dt} \right\rangle &= \varepsilon^3 \left\{ \frac{\bar{m}^A g_3 g_1}{\Delta^3} \left[ \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \right) \frac{\partial \Delta}{\partial \alpha} - \bar{m}^A g_3 \frac{\partial^2 \Delta}{\partial \alpha^2} \right] - \frac{\bar{m}^A g_3}{\Delta^3} \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \Delta \right) \frac{\partial g_1}{\partial \alpha} + \frac{g_1}{\Delta^2} \left[ \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \right) \right]^2 - \right. \\ &- \frac{\bar{m}^A g_2}{\Delta^3} \frac{\partial g_3}{\partial \omega_x} \left[ \frac{\partial}{\partial \alpha} \left( \bar{m}^A g_3 \right) \Delta + 2\bar{m}^A g_3 \frac{\partial \Delta}{\partial \alpha} \right] - \frac{(\bar{m}^A)^2 g_3}{\Delta^2} \frac{\partial g_1}{\partial \alpha} \frac{\partial g_3}{\partial \alpha} + \frac{(\bar{m}^A)^2 g_3}{2\Delta^4} \frac{\partial \Delta}{\partial \omega_x} \left[ 7g_2 \frac{\partial \Delta}{\partial \alpha} - 4\Delta \frac{\partial g_2}{\partial \alpha} \right] + \\ &+ \left. \frac{(\bar{m}^A)^2 g_3^2}{2\Delta^4} \left[ 2\Delta^2 \frac{\partial^2 g_1}{\partial \alpha^2} - g_1 \left( \frac{\partial \Delta}{\partial \alpha} \right)^2 \right] \right\} \frac{m^A \cos(2\theta_1 - 2\theta_3)}{8} + \varepsilon^3 g_4. \end{aligned} \quad (6)$$

$$\text{Here } g_1 = \frac{2\omega_a \omega_{1,2} \sin \alpha}{F_a} \left( \omega_x + \frac{\omega_{1,2}^2 \sin^2 \alpha}{2\omega_a} \right), \quad g_2 = \frac{\omega_{1,2}^2 \sin^2 \alpha}{\bar{I}_x}, \quad g_3 = \frac{2\omega_a \omega^2}{F_a}, \quad g_4 = \frac{4m_{zn} S L \omega_a^2}{IF_a} \frac{dq}{dt}.$$

Equations (4) and (5) describe the non-resonant evolution of the angular velocity  $\omega_x$  and the angle of attack  $\alpha$  caused by the effect of the principal resonance  $\Delta = 0$ . At the positive values of  $\omega_x$  resonant ratio is equal to  $\Delta = (1 - \frac{\bar{I}_x}{2})\omega_x - \omega_a$ . Let us assume that the spacecraft has the following ratio of the moments of inertia:  $\bar{I}_x = 2$ . Here the resonant ratio is equal to

$$\Delta = -\sqrt{\omega_x^2 + \omega^2}. \quad (7)$$

We introduce the function  $V(\omega_x, \omega) = \Delta^2$ . This Lyapunov function can be written as:

$$V(\omega_x, \omega) = \omega_x^2 + \omega^2. \quad (8)$$

Here  $\omega_x, \omega$  are determined from equations (5) and (4) respectively. Given the expression of (7), we see that the principal resonance  $\Delta = 0$  is realized at

$$\begin{cases} \omega_x = 0, \\ \omega = 0. \end{cases} \quad (9)$$

Thus, the condition of the external stability of the principal resonance [5] has the following form:

$$\frac{dV}{dt} = 2\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle + 2\omega \frac{d\omega}{dt} < 0. \quad (10)$$

The condition (10) is a condition of asymptotic stability of a trivial solution (9). The fulfillment of the condition (10) provides for realization of a resonant attractor (9). On the contrary, the condition of external instability of the principal resonance is the following:

$$\frac{dV}{dt} = 2\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle + 2\omega \frac{d\omega}{dt} > 0. \quad (11)$$

The condition (11) is the condition of instability of the trivial solution (9). The fulfillment of the condition (11) ensures the realization of the resonant repeller (9). Let us assume that  $\{\omega_x > 0, \omega > 0\}$ . In this case, asymptotic analysis of conditions (10), (11) makes it possible to distinguish the following twelve typical cases of resonant attractor or resonant repeller realization: 1) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt < 0$ ,  $\omega_x(0) > \max \omega_x^r > 0$ , condition (10) is fulfilled and resonant attractor (9) is realized; 2) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt > 0$ ,  $\max \omega_x^r > \omega_x(0) > \omega_x^r(0) > 0$ , condition (11) is fulfilled and resonant repeller (9) is realized; 3) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $\omega_x^r(0) > \omega_x(0) > 0$ , condition (10) is fulfilled and attractor (9) is realized; 4) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $\omega_x^r(0) > \omega_x(0) > 0$ , condition (11) is fulfilled and repeller (9) is realized; 5) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) > \max \omega_x^r > 0$ , condition (10) is fulfilled and attractor (9) is realized; 6) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $-\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < \omega \frac{d\omega}{dt}$ ,  $\omega_x^r(0) > \omega_x(0) > 0$ , condition (11) is fulfilled and repeller (9) is realized; 7) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) > \max \omega_x^r > 0$ , condition (10) is fulfilled and repeller (9) is realized; 8) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt > 0$ ,  $0 < \omega_x(0) < \omega_x^r(0)$ , condition (11) is fulfilled and resonant repeller (9) is realized; 9) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) > \max \omega_x^r > 0$ , condition (11) is fulfilled and repeller (9) is realized; 10) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt < 0$ ,  $0 < \omega_x(0) > \omega_x^r(0)$ , the condition (10) is fulfilled and the resonant attractor (9) is realized; 11) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $-\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > \omega \frac{d\omega}{dt}$ ,  $\omega_x^r(0) > \omega_x(0) > 0$ , condition (11) is fulfilled

and repeller (9) is realized; 12) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) > \max \omega_x^r > 0$ , condition (11) is fulfilled and attractor (9) is realized.

Similarly we consider the case  $\{\omega_x < 0, \omega < 0\}$ . In this case, the resonant ratio  $\Delta = (1 - \frac{\bar{I}_x}{2})\omega_x - \omega_a$  at  $\bar{I}_x = 2$  is

$$\Delta = \sqrt{\omega_x^2 + \omega^2}. \quad (12)$$

In this case, the Lyapunov function is (8). Similar typical twelve cases are following: 13) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt > 0$ ,  $\omega_x(0) < \min \omega_x^r < 0$ , condition (10) is fulfilled and resonant attractor (9) is realized; 14) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt < 0$ ,  $\min \omega_x^r < \omega_x(0) < \omega_x^r(0) < 0$ , condition (10) is fulfilled and repeller (9) is realized; 15) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $0 > \omega_x(0) > \omega_x^r(0)$ , condition (10) is fulfilled and attractor (9) is realized; 16) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $0 > \omega_x(0) > \omega_x^r(0)$ , condition (11) is fulfilled and repeller (9) is realized; 17) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) < \omega_x^r(0) < 0$ , condition (10) is fulfilled and attractor (9) is realized; 18) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $0 > \omega_x(0) > \omega_x^r(0)$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ , condition (11) is fulfilled and resonant repeller (9) is realized; 19) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) < \omega_x^r(0) < 0$ , condition (10) is fulfilled and repeller (9) is realized; 20) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt < 0$ ,  $0 > \omega_x(0) > \omega_x^r(0)$ , condition (11) is fulfilled and resonant repeller (9) is realized; 21) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle > -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) < \min \omega_x^r < 0$ , condition (11) is fulfilled and repeller (9) is realized; 22) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt > 0$ ,  $0 > \omega_x(0) > \omega_x^r(0)$ , condition (10) is fulfilled and resonant attractor (9) is realized; 23) if  $\langle d\omega_x/dt \rangle > 0$ ,  $d\omega/dt < 0$ ,  $-\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < \omega \frac{d\omega}{dt}$ ,  $\omega_x^r(0) < \omega_x(0) < 0$ , condition (11) is fulfilled and repeller (9) is realized; 24) if  $\langle d\omega_x/dt \rangle < 0$ ,  $d\omega/dt > 0$ ,  $\omega_x \left\langle \frac{d\omega_x}{dt} \right\rangle < -\omega \frac{d\omega}{dt}$ ,  $\omega_x(0) < \min \omega_x^r < 0$ , condition (11) is fulfilled and attractor (9) is realized.

#### 4. Numerical results

Numerical results obtained from solve of the equations (1)-(4) confirmed fulfillment of the twenty-four cases discussed above. In particular, Fig. 1 shows the dependence of the Lyapunov function on slow variables  $\omega_x$  and  $\omega$  when realization of resonant attractor. This numerical result corresponds to a typical case 5). Fig. 2 shows the dependence of the Lyapunov function on slow variables  $\omega_x$  and  $\omega$  when realization of resonant repeller. The numerical result shown in Fig. 2 corresponds to case 2). The following parameters of the spacecraft and initial conditions of motion were used in the construction of Figs. 1-2:  $m = 70$  kg;  $S = 0.1$  m<sup>2</sup>,  $L = 0.54$  m,  $\Delta m = 0.02$ ,  $\bar{m}^A = 0.05$ ,  $\theta_1 - \theta_3 = \pi$ ,  $I = 1$  kgm<sup>2</sup>,  $I_x = 0.3$  kgm<sup>2</sup>,  $V(0)$  is the initial value of the spacecraft velocity,  $V(0) = 3400$  m/s,  $\mathcal{G}(0)$  is the initial value of the local flight-pass inclination angle,  $\mathcal{G}(0) = -0.087$  rad,  $H(0)$  is the initial value of spacecraft altitude,  $H(0) = 100$  km,  $\varphi(0) = 0$ ,  $\alpha(0) = 0.05$  rad,  $\omega_x = 10$  s<sup>-1</sup> (Fig.1);  $\Delta m = 0.005$ ,  $\bar{m}^A = 0.05$ ,  $\theta_1 - \theta_3 = 0$ ,  $\omega_x = 15$  s<sup>-1</sup> (Fig. 2). Direction of non-resonant evolution of the corresponding variables is indicated in Figs. 1-2 by arrows.

#### 5. Conclusion and results

Thus, the use of the method of averaging and Lyapunov's second method made it possible to carry out an asymptotic analysis of the non-resonant evolution of slow variables during the atmospheric descent of the spacecraft with small aerodynamic-inertial asymmetry. By doing so, we obtained conditions for realization of the resonant attractor and resonant repeller at arbitrary angles of attack. In addition, we identified ten typical cases of resonant attractor realization and fourteen typical cases of resonant repeller realization. The approximate analytical results of the study correspond to the results of the numerical simulation. The conditions presented in this study indicate that the resonant attractor can become the resonant repeller. It is also possible for a reverse transition. These transitions can occur due to the change of sign of the angular velocity  $\omega_x$ . By analyzing of the stability conditions, we assumed that the asymmetry parameters take constant values. It should be noted that the descent of a spacecraft with variable asymmetry into the atmosphere presents of a practical interest. For example, the variable asymmetry in

the considered dynamical system can lead to a transition from the resonant attractor to the resonant repeller. Research of such transient modes falls beyond the scope of this study and may be detailed in the following papers.

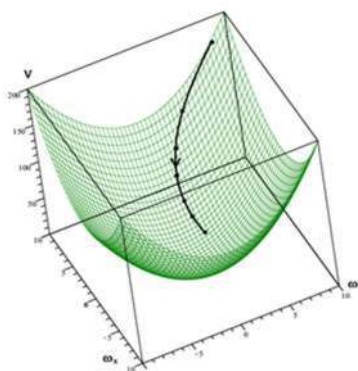


Fig.1. Lyapunov's function and angular velocities when resonant attractor.

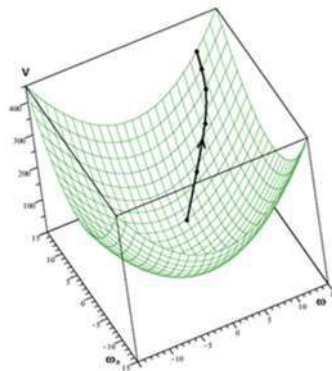


Fig.2. Lyapunov's function and angular velocities when resonant repeller.

## References

- [1] Yaroshevskiy VA. Atmospheric reentry of spacecraft. Moscow: Nauka, 1988; 336 p.
- [2] Shilov AA, Goman MG. Resonance modes of spatial uncontrolled movement of spacecraft at entry into the atmosphere. Proceedings of TsAGI 1975; 1624: 44 p.
- [3] Zabolotnov YuM, Lyubimov VV. Secondary resonance effect in the motion of a spacecraft in the atmosphere. Cosmic Research 1998; 36(2): 194–201.
- [4] Lyubimov VV. Asymptotic analysis of the secondary resonance effects in the rotation of a spacecraft with a small asymmetry in the atmosphere. Russian Aeronautics 2014; 57(3): 245–252.
- [5] Lyubimov VV, Lashin VS. External stability of a resonance during the descent of a spacecraft with a small variable asymmetry in the martian atmosphere. Advances in Space Research 2017; 59(6): 1607–1613.
- [6] Sadov YA. Secondary resonance effects in mechanical systems. Mechanics of Solids 1990; 4: 20–24.
- [7] Lyubimov VV. External stability of resonances in the motion of an asymmetric rigid body with a strong magnet in the geomagnetic field. Mechanics of Solids 2010; 45(1): 10–21.
- [8] Kalaev MP, Lyubimov VV, Semkin ND. Seminatural modeling and numerical simulation for the process of the small satellite separation. Gyroscopy and Navigation 2014; 2(85): 52–60.
- [9] Zabolotnov YuM, Lyubimov VV. Application of the method of integral manifolds for construction of resonant curves for the problem of spacecraft entry into the atmosphere. Cosmic Research 2003; 41(5): 453–459.
- [10] Shchepakina E, Sobolev V, Mortell MP. Singular perturbations: Introduction to system order reduction methods with applications. Springer Lecture Notes in Mathematics, 2014; 212 p.



# Control of a one rigid-link manipulator in the case of non-smooth periodic trajectory

N. Aksenova<sup>1</sup>, V. Sobolev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

Mathematical model of a single-link manipulator is considered. It describes the motion of the manipulator in the case of non-smooth path. Interpolation of the trajectory of motion is used, which makes it possible to reduce the amount of calculations and allows you to take into account the restrictions on the movement of the manipulator. Integral manifold method is used for the system order reduction. As a result, the reduced system of the investigated object is obtained, and the control function for the manipulation robot model in the case of a non-smooth periodic trajectory is constructed.

*Keywords:* mathematical model; manipulation robot; integral manifold; singular perturbations; periodic trajectory

## 1. Introduction

In this paper, we consider a mathematical model of a robotic manipulator that describes its motion along a non-smooth periodic trajectory. After path definition of the manipulator movement control function is selected. It allows implementing the required movement accurately. To solve the problem, we use the method of integral manifolds [1-3]. As applied to control problems, this method was considered in [4-7].

## 2. Single-link manipulator model

The equations of motion of a single-point manipulator have the form [7-8]:

$$J_1 \ddot{q}_1 + Mgl \sin q_1 + c(\dot{q}_1 - \dot{q}_m) + k(q_1 - q_m) = 0, \quad (1)$$

$$J_m \ddot{q}_m - c(\dot{q}_1 - \dot{q}_m) - k(q_1 + q_m) = u,$$

where  $J_m$  is the jet second moment;  $J_1$  is the link second moment;  $M$  is the link mass;  $l$  is the link length;  $c$  is the attenuation factor;  $k$  is the hardness. Let  $q_1$  is the link angular displacement;  $q_m$  is the output angle, and  $u$  is the control circuit. In Fig.1 the image of the single-link manipulator is presented.

Variables in the system are changed in the following manner:

$$x_1 = \frac{J_1 q_1 + J_m q_m}{J_1 + J_m}, \quad x_2 = \dot{x}_1, \quad y_1 = q_1 - q_m, \quad y_2 = \varepsilon \dot{y}_1, \quad (2)$$

Then system (1) is transformed to:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = \frac{Mgl}{J_1 + J_m} \sin \left( x_1 + \frac{J_m}{J_1 + J_m} y_1 \right) + \frac{u}{J_1 + J_m}, \quad (3)$$

$$\varepsilon \dot{y}_1 = y_2, \quad \varepsilon \dot{y}_2 = - \left( \frac{1}{J_1} + \frac{1}{J_m} \right) y_1 - \varepsilon c \left( \frac{1}{J_1} + \frac{1}{J_m} \right) y_2 - \varepsilon^2 \frac{Mgl}{J_1} \sin \left( x_1 + \frac{J_m}{J_1 + J_m} y_1 \right) - \varepsilon^2 \frac{u}{J_m}. \quad (4)$$

This system is singularly perturbed with slow subsystem (3) and fast subsystem (4). Omitting all terms of  $O(\varepsilon^2)$  order in the right hand side of the last equation the independent subsystem is obtained.

$$\varepsilon \dot{y}_1 = y_2, \quad \varepsilon \dot{y}_2 = - \left( \frac{1}{J_1} + \frac{1}{J_m} \right) y_1 - \varepsilon c \left( \frac{1}{J_1} + \frac{1}{J_m} \right) y_2,$$

The solutions of system are characterized by quite high frequency  $\frac{\sqrt{\left(\frac{1}{J_1} + \frac{1}{J_m}\right)}}{\varepsilon}$  and relatively low damping factor  $c\left(\frac{1}{J_1} + \frac{1}{J_m}\right)/2$ , and differential system has a characteristic equation

$$\varepsilon^2 \lambda^2 + c \left( \frac{1}{J_1} + \frac{1}{J_m} \right) \lambda + \left( \frac{1}{J_1} + \frac{1}{J_m} \right)$$

with complex roots

$$\lambda_{1,2} = -\frac{c}{2} \left( \frac{1}{J_1} + \frac{1}{J_m} \right) \pm \frac{i}{2} \sqrt{\left( \frac{1}{J_1} + \frac{1}{J_m} \right) - \varepsilon^2 \frac{c^2}{4} \left( \frac{1}{J_1} + \frac{1}{J_m} \right)^2} \quad (5)$$

As far as a real part is negative, slow invariant manifold can be used for model analysis of the concerned manipulator.

## 3. Integral manifold construction

To calculate the slow integral manifold for the system (3)-(4) we use asymptotic expansions and obtain, within the accuracy of  $O(\varepsilon^3)$ ,  $y_1 = \varepsilon^2 Y + O(\varepsilon^3)$  и  $y_2 = O(\varepsilon^3)$  (5), where

$$Y = - \left[ \frac{Mgl}{J_1} \sin(x_1) + \frac{u_0}{J_m} \right] \left( \frac{1}{J_1} + \frac{1}{J_m} \right)^{-1}$$

Here the representation  $u = u_0 + \varepsilon^2 u_1 + O(\varepsilon^3)$  is used.

Movement on the manifold is described by the following equations

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -\frac{Mgl}{J_1+J_m} \sin\left(x_1 + \varepsilon^2 \frac{J_m}{J_1+J_m} Y\right) + \frac{u_0 + \varepsilon^2 u_1}{J_1+J_m} + O(\varepsilon^3) \quad (6)$$

Manipulator angular displacement  $q_1$  is expressed using new variables

$$q_1 = x_1 + \frac{J_m}{J_m+J_1} y_1, \quad (7)$$

where  $y_1 = \varepsilon^2 Y + O(\varepsilon^3)$ . This allows to rewrite the system (5) on the slow integral manifold as

$$\ddot{q}_1 - \varepsilon^2 \frac{J_m}{J_m+J_1} \ddot{Y} = -\frac{Mgl}{J_1+J_m} \sin(q_1) + \frac{u_0 + \varepsilon^2 u_1}{J_1+J_m} + O(\varepsilon^3). \quad (8)$$

#### 4. Control function

Let  $q_d(t)$  be the required trajectory of the manipulator movement. Slow control function term is in the form

$u_0 = (J_1 + J_m)u_d + Mgl \sin q_1$ , где  $u_d = \ddot{q}_d - a_1(x_1 + q_d) - a_2(\dot{x}_1 + \dot{q}_d)$ . Using (8) and  $u_0$  and  $u_d$  we obtain within the accuracy of the order  $O(\varepsilon^2)$

$$\ddot{q}_1 - \ddot{q}_d + a_2(\ddot{q}_1 + \ddot{q}_d) + a_1(q_1 + q_d) = 0 \quad (9)$$

for  $q_1 - q_d$ , and  $q_1 = x_1 + O(\varepsilon^3)$  on the slow integral manifold.

Equation (9) gives the possibility to select control function  $u_d$  coefficients in such a way that the relevant control affords to achieve the required trajectory. Assume, for instance,  $M=1, k=100, l=1, J_1=1, J_m=1, g=9.8, c=2$ , at that  $a_1=3, a_2=4$ , and the required trajectory is of the form  $q_d = \sin t$ , then we obtain the following original variables control law

$$u = 2u_d + 9.8 \sin(q_1) = 2[-\sin t - 4(\dot{q}_1 - \cos t) - 3(q_1 - \sin t)] + 9.8 \sin(q_1)$$

The first stage of the control construction is to determine the desired trajectory of motion of the manipulator in the form of some analytically described function. In most cases, the manipulators do not move along smooth trajectories, so that its trajectory is a sectionally smooth line. For smoothing the interpolation of the chosen trajectory is used by polynomials of a certain class approximating the segments of the desired trajectory of the manipulation robot between the node points (for example, lines, arcs, parabolas, etc.). But there is a possibility that there will be a problem associated with the difficulty of calculating a polynomial of high degree. In this regard, to perform interpolation of the trajectory from the given nodal points, it is necessary to choose polynomials of low degrees or to break the trajectory of the manipulator's movement into separate sections.

In Fig. 1 there is a displacement-time diagram in case the required path  $q_d$  is written as

$$q_d = \begin{cases} x, & 0 < x < \delta - 1 \\ a(x-1)^4 + b(x-1)^2 + 1, & \delta - 1 < x < \delta + 1 \\ -x + 2, & \delta + 1 < x < 2 \end{cases}$$

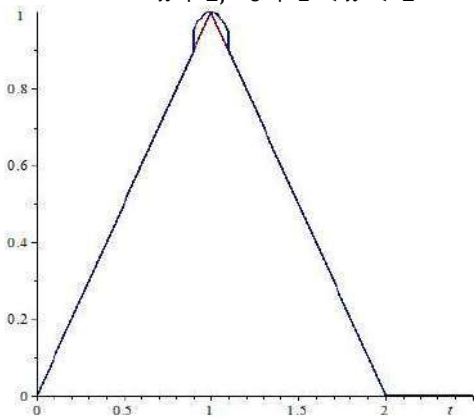


Fig. 1. Trajectory  $q_d$ .

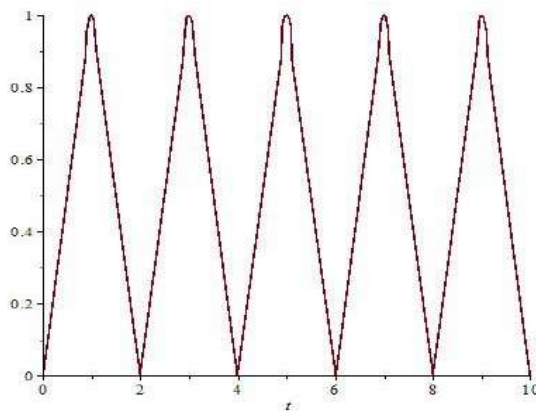


Fig.2. Periodic trajectory.

When the trajectory  $q_d$  is substituted in the system of equations of motion of the manipulation robot (1), the trajectory of motion will look as follows (fig. 3).

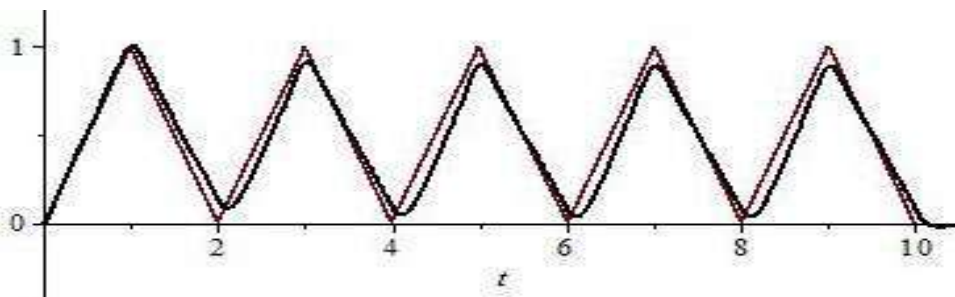


Fig. 3. Trajectory of motion of the manipulator in the case of a periodic trajectory.

#### Conclusion

The object of research is a manipulator model describing the manipulator motion in a non-smooth path. The interpolation of the trajectory of motion by polynomials is used that approximates the segments of the desired trajectory of the manipulation

robot between the nodal points, which makes it possible to reduce the amount and time of calculations, and allows us to take into account the restrictions on the movement of the manipulator. Integral manifold method is used for the system order reduction.

As a result of the work done the reduced system of the object is obtained and the control function for a diagrammatic formulation of the manipulator model motion. It is established that manifold control provides the motion of the system along the trajectory near to the effective one.

### Acknowledgements

This study was supported by the Russian Foundation for Basic Research and Samara region (grant 16-41-630524-p) and the Ministry of Education and Science of the Russian Federation as part of a program of increasing the competitiveness of SSAU in the period 2013–2020.

### References

- [1] Shchepakina E, Sobolev V, Mortell MP. Singular Perturbations: Introduction to system order reduction methods with applications, 2014; 121 p.
- [2] Sobolev VA, Tropkina EA. Asymptotic expansions of slow invariant manifolds and reduction of chemical kinetics models. *Comput. Mathematics and Math. Physics* 2012; 52(1): 75–89.
- [3] Strygin VV, Sobolev VA. Effect of geometric and kinetic parameters and energy dissipation on orientation stability of dual-spin satellites. *Cosmic Research* 1976; 14(3): 331–335.
- [4] Smetannikova E, Sobolev V. Regularization of Cheap Periodic Control Problems. *Automation and Remote Control* 2005; 66(6): 903–916.
- [5] Mikheev YuV, Sobolev VA, Fridman EM. Asymptotic analysis of digital control systems. *Autom. Remote Control* 1988; 49(9): 1175–1180.
- [6] Sobolev VA. Singular perturbations in linearly quadratic optimal control problems. *Autom. Remote Control* 1991; 52(2): 180–189.
- [7] Ghorbel F, Spong MW. Integral manifolds of singularly perturbed systems with application. to rigid-link flexible-joint multibody systems. *Int. J. of Non-Linear Mechanics* 2000; 35: 133–155.
- [8] Spong MW. Modeling and control of elastic joint robots. *J. of Dynamic Systems, Measurement and Control* 1987; 109(4): 310–319.

# On stabilizability of the manifold of steady states in a model of the spread of a mutating viruses

Ju. Ermoshkina<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

A system of semilinear parabolic equations with a manifold of steady states is considered and the conditions of stabilizability of this manifold are obtained in the paper.

*Keywords:* bifurcation; parabolic equation; the manifold of equilibrium states; the model of interaction of viruses

## 1. Introduction

Consider the system of differential equations:

$$\begin{aligned} \frac{da}{dt} &= A(a, y, z), \\ \frac{dy}{dt} &= By + Y(a, y, z), \\ \frac{dz}{dt} &= Cz + Z(a, y, z), \end{aligned} \quad (1)$$

where  $a, A \in R^l; y, Y \in R^k; z, Z \in R^m$ . Assume that  $A(a, 0, 0) \equiv 0, Y(a, 0, 0) \equiv 0, Z(a, 0, 0) \equiv 0$ . Then the system (1) has a manifold of equilibrium states  $\mathfrak{M} = \{(a, 0) | a \in R^l, 0 \in R^k \times R^m\}$ .

Following [1, 2], let say that the manifold  $\mathfrak{M}$  is stable with respect to variable  $x = (y, z)$ , if for any point  $a \in R^l$  and any neighborhood of zero  $W$  in phase space  $R^k \times R^m$  we can specify such a neighborhood of zero  $W_0 \subset R^k \times R^m$ , that for any point  $x_0 = (y_0, z_0) \in W_0$  the corresponding solution  $a = a(t, a_0, x_0), x = x(t, a_0, x_0)$  ( $a(0, a_0, x_0) = a_0, x(0, a_0, x_0) = x_0$ ) satisfies the ratio  $x = x(t, a_0, x_0) \in W$  when  $t \geq 0$ .

Let say that  $\mathfrak{M}$  is asymptotically stable with respect to variable  $x = (y, z)$ , if it is stable with respect to variable  $x$  and, moreover,  $\lim_{t \rightarrow \infty} x(t, a_0, x_0) = 0$  for all  $x_0 \in W_0$ .

Let say that  $\mathfrak{M}$  is stabilized, if it is asymptotically stable with respect to variable  $x$  and when  $t \rightarrow \infty \{a(t, a_0, x_0), x(t, a_0, x_0)\}$  converge to some point of diversity  $\mathfrak{M}$ , if  $x_0 \in W_0$ .

M.A. Ayzerman and F.R. Gantmakher established that the state of equilibrium of nonholonomic system is stable, if all roots of the characteristic equation, except for the zero roots, the number of which equals the number of equations of nonholonomic connections, have negative real parts [3, 4]. Each perturbed motion, which is close enough to unperturbed motion, is converge to one of the possible established motions, belong to a given manifold, when  $t \rightarrow \infty$ . [5]

## 2. Model description

Let consider the model of interaction of two populations of microorganisms in one-dimensional case. This system is based on the equations of Fisher-Kolmogorov-Petrovsky-Piskunov. Let  $u(x, t)$  and  $v(x, t)$  be concentrations of the two sub-types of a virus at a point  $x$  and a time  $t$ . Consider the problem on the interval  $x \in [0; 1]$ . The system has the form:

$$\begin{cases} \frac{\partial u(x,t)}{\partial t} = D_1 \frac{\partial^2 u(x,t)}{\partial x^2} + a_1 u(x,t)(1 - q_1 v(x,t))(1 - u(x,t) - v(x,t)); \\ \frac{\partial v(x,t)}{\partial t} = D_2 \frac{\partial^2 v(x,t)}{\partial x^2} + a_2 v(x,t)(1 - q_2 u(x,t))(1 - u(x,t) - v(x,t)), \end{cases} \quad (2)$$

where  $a_1, a_2$  - the replacement rates for populations  $u$  and  $v$  accordingly,  $D_1, D_2$  - the coefficients of diffusion,  $q_1, q_2$  - the coefficients of the interaction between individuals of different populations.

The condition of impermeability at the ends of the considered interval are considered as the boundary conditions in this problem. They look like:

$$\begin{aligned} \left. \frac{\partial u(x,t)}{\partial x} \right|_{x=0} &= \left. \frac{\partial u(x,t)}{\partial x} \right|_{x=1} = 0; \\ \left. \frac{\partial v(x,t)}{\partial x} \right|_{x=0} &= \left. \frac{\partial v(x,t)}{\partial x} \right|_{x=1} = 0. \end{aligned} \quad (3)$$

Continuous functions are chosen as the initial conditions. They have the form:

$$\begin{aligned} u(x, 0) &= \begin{cases} 0,9(-5(x-1)^2 + 1), u > 0, \\ 0, u \leq 0; \end{cases} \\ v(x, 0) &= \begin{cases} 0,9(-5x^2 + 1), v > 0, \\ 0, v \leq 0. \end{cases} \end{aligned} \quad (4)$$

### 3. Analysis of the model

Let find the conditions of stability for the model (2). First of all, let find the equilibrium states of the system. These stationary solutions are obtained by equating all partial derivatives to zero in equations of model. Introduce functions  $f_1, f_2$ , which defined by the following equations:

$$\begin{aligned} f_1 &= a_1 \bar{u}(1 - q_1 \bar{v})(1 - \bar{u} - \bar{v}) = 0; \\ f_2 &= a_2 \bar{v}(1 - q_2 \bar{u})(1 - \bar{u} - \bar{v}) = 0. \end{aligned} \tag{5}$$

From equations (5) it is easy to obtain the equilibrium states of the system:

$$(\bar{u}_1, \bar{v}_1) = (0, 1); \quad (\bar{u}_2, \bar{v}_2) = (1, 0); \quad (\bar{u}_3, \bar{v}_3) = (0.5, 0.5). \tag{6}$$

To obtain the closest linear system, where  $(u, v)$  is close to  $(\bar{u}, \bar{v})$ , let introduce the infinitesimal perturbations:

$$\xi(x, t) = u(x, t) - \bar{u}; \quad \eta(x, t) = v(x, t) - \bar{v}. \tag{7}$$

Consider the approximation of functions  $f_1(u, v), f_2(u, v)$  near any equilibrium states  $(\bar{u}, \bar{v})$ . Multivariable calculus may be used to obtain the following approximations:

$$\begin{aligned} f_1(u, v) &\approx f_1(\bar{u}, \bar{v}) + \frac{\partial f_1}{\partial u} \xi + \frac{\partial f_1}{\partial v} \eta; \\ f_2(u, v) &\approx f_2(\bar{u}, \bar{v}) + \frac{\partial f_2}{\partial u} \xi + \frac{\partial f_2}{\partial v} \eta. \end{aligned} \tag{8}$$

Members of the second and higher orders can be neglected since the perturbations are infinitely small. Taking into consideration equations (5), let receive:

$$\begin{aligned} f_1(u, v) &\approx \frac{\partial f_1}{\partial u} \xi + \frac{\partial f_1}{\partial v} \eta; \\ f_2(u, v) &\approx \frac{\partial f_2}{\partial u} \xi + \frac{\partial f_2}{\partial v} \eta. \end{aligned} \tag{9}$$

Finally, substituting equations determining the perturbations (7) into the equations defining the model (2), leads to a set of equations showing how the perturbation will develop in time:

$$\begin{aligned} \frac{\partial \xi}{\partial t} &= D_1 \frac{\partial^2 \xi}{\partial x^2} + \frac{\partial f_1}{\partial u} \xi + \frac{\partial f_1}{\partial v} \eta, \\ \frac{\partial \eta}{\partial t} &= D_1 \frac{\partial^2 \eta}{\partial x^2} + \frac{\partial f_2}{\partial u} \xi + \frac{\partial f_2}{\partial v} \eta. \end{aligned} \tag{10}$$

Let consider the Jacobian matrix for the system (10). The signs of the eigenvalues of this matrix will give the conditions of stability of the stationary solutions.

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial u} & \frac{\partial f_1}{\partial v} \\ \frac{\partial f_2}{\partial u} & \frac{\partial f_2}{\partial v} \end{pmatrix}. \tag{11}$$

Accounting that  $f_1 = a_1 u(1 - q_1 v)(1 - u - v), f_2 = a_2 v(1 - q_2 u)(1 - u - v)$ , let calculate the partial derivatives of these functions on variables  $u, v$ . Then, the Jacobian matrix A takes the form:

$$A = \begin{pmatrix} a_1(1 - q_1 v)(1 - 2u - v) & a_1 u(q_1(u + 2v - 1) - 1) \\ a_2 v(q_2(v + 2u - 1) - 1) & a_2(1 - q_2 u)(1 - u - 2v) \end{pmatrix}. \tag{12}$$

Following the research conducted by Juan Carlos Cantero and Andrei Korobeinikov [6], consider the position of equilibrium  $(\bar{u}_3, \bar{v}_3) = (0.5, 0.5)$ .

Substitute  $(\bar{u}_3, \bar{v}_3) = (0.5, 0.5)$  in (12):

$$A = \begin{pmatrix} -0.5a_1(1 - 0.5q_1) & -0.5a_1(1 - 0.5q_1) \\ -0.5a_2(1 - 0.5q_2) & -0.5a_2(1 - 0.5q_2) \end{pmatrix}. \tag{13}$$

The determinant of the Jacobian matrix equals to zero, and the stability of the solution will depend on the trace of the matrix A. If the trace of the matrix is negative, then the solution is stable. Then if  $a_1 q_1 + a_2 q_2 < 2(a_1 + a_2)$ , the stationary solution is stable. And if  $a_1 q_1 + a_2 q_2 > 2(a_1 + a_2)$ , the solution is not stable.

### 4. Numerical modeling

To solve the problem (2)-(4) let make an explicit finite-difference scheme. To do this, replace the differential operators of their mesh analogues. Receive:

$$\begin{cases} \frac{u_i^{k+1} - u_i^k}{\tau} = D_1 \frac{u_{i+1}^k - 2u_i^k + u_{i-1}^k}{h^2} + a_1 u_i^k (1 - q_1 v_i^k) (1 - u_i^k - v_i^k); \\ \frac{v_i^{k+1} - v_i^k}{\tau} = D_2 \frac{v_{i+1}^k - 2v_i^k + v_{i-1}^k}{h^2} + a_2 v_i^k (1 - q_2 u_i^k) (1 - u_i^k - v_i^k). \end{cases} \tag{14}$$

The boundary conditions will take the form:

$$\begin{aligned} \frac{u_1^{k+1} - u_1^k}{h} &= 0; \\ \frac{v_1^{k+1} - v_1^k}{h} &= 0. \end{aligned} \tag{15}$$

Define the initial conditions as follows:

$$\begin{aligned} u_i^0 &= \begin{cases} 0, 9(-5(x_i - 1)^2 + 1), u_i^0 > 0, \\ 0, u_i^0 \leq 0; \end{cases} \\ v_i^0 &= \begin{cases} 0, 9(-5x_i^2 + 1), v_i^0 > 0, \\ 0, v_i^0 \leq 0. \end{cases} \end{aligned} \tag{16}$$

Their graphs are presented in figure 1.

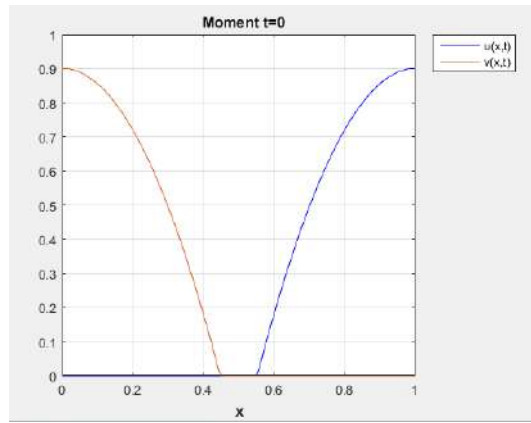


Fig. 1. Graph of the initial conditions for  $u$  and  $v$ .

To solve the problem (14)-(16) the program was realised in Matlab, which calculates the values of the grid functions on the time interval  $0 \leq t \leq 600$ .

**5. Different cases**

Consider the case when the coefficients of the first and the second equations are equal, i. e.  $a_1 = a_2 = 1$ ,  $D_1 = D_2 = 0.001$ ,  $q_1 = q_2$ . Separating the variables and solving the task on eigenvalues, find the value of parameters  $q_1 = q_2 = 2$ , in the transition through which the bifurcation happens in the system. To illustrate this phenomenon, consider the three different cases:

1.  $q_1 = q_2 < 2$
2.  $q_1 = q_2 \approx 2$
3.  $q_1 = q_2 > 2$

In the first case, the trajectories of system converge to the equilibrium  $(0,5;0,5)$ , belonging to the manifold of equilibrium states of the system. By Ayzerman-Gantmacher's theorem, the state of equilibrium of system is stable. Thus, manifold is stabilized. In the second case, there is a soft loss of stability of the system when passing through the critical value, and in the third case, it is possible to observe a complete loss of stability.

*5.1. Case, when  $q_1 = q_2 < 2$ .*

For the first case, when  $q_1 = q_2 = 1.5$ , the dynamics of function  $u(x,t)$  is presented in figure 2. The dynamics of function  $v(x,t)$  is presented in figure 3. In figure 4 a solution in a finite time  $t=600$  is presented.

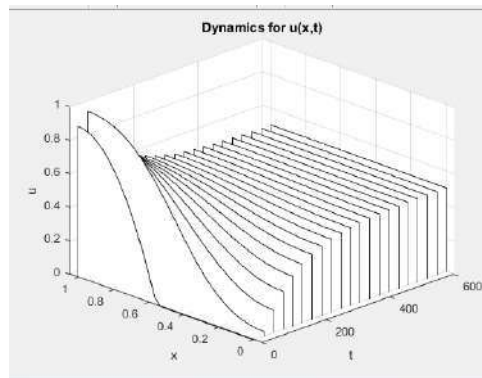


Fig. 2. The dynamics of function  $u(x, t)$  for case 1.

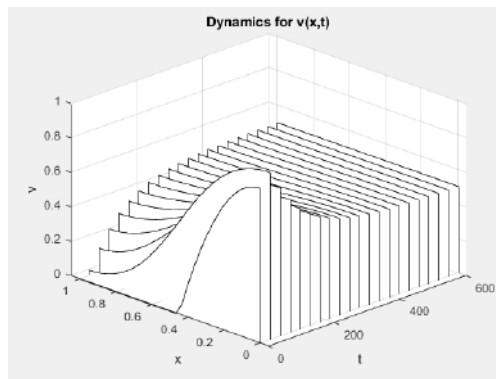


Fig. 3. The dynamics of function  $v(x, t)$  for case 1.

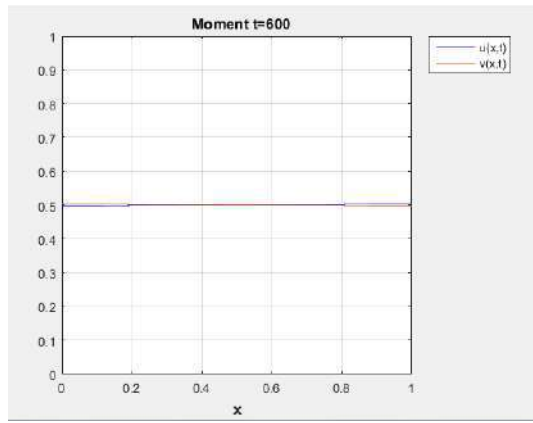


Fig. 4. A solution in a finite time  $t=600$  for case 1.

5.2. Case, when  $q_1 = q_2 \approx 2$ .

For the second case, when  $q_1 = q_2 = 2.05$ , the dynamics of function  $u(x,t)$  is presented in figure 5. The dynamics of function  $v(x,t)$  is presented in figure 6. In figure 7 a solution in a finite time  $t=600$  is presented.

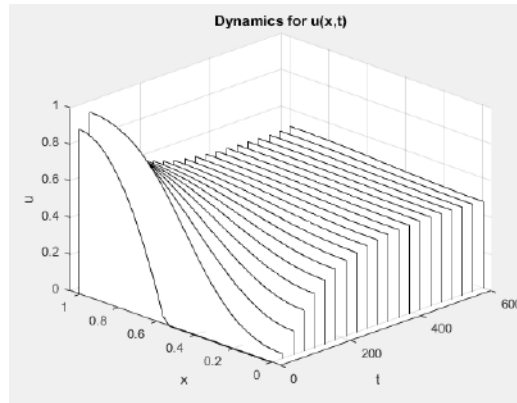


Fig. 5. The dynamics of function  $u(x, t)$  for case 2.

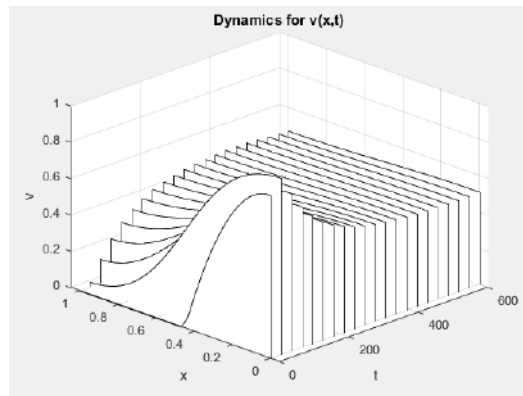


Fig. 6. The dynamics of function  $v(x, t)$  for case 2.

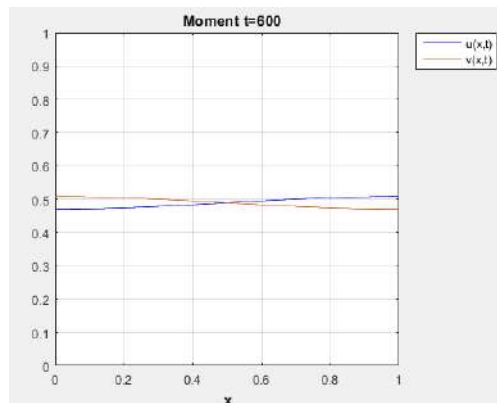


Fig. 7. A solution in a finite time  $t=600$  for case 2.

5.3. Case, when  $q_1=q_2>2$ .

For case 3, when  $q_1 = q_2 = 2.5$ , the dynamics of function  $u(x,t)$  is presented in figure 8. The dynamics of function  $v(x,t)$  is presented in figure 9. In figure 10 a solution in a finite time  $t=600$  is presented.

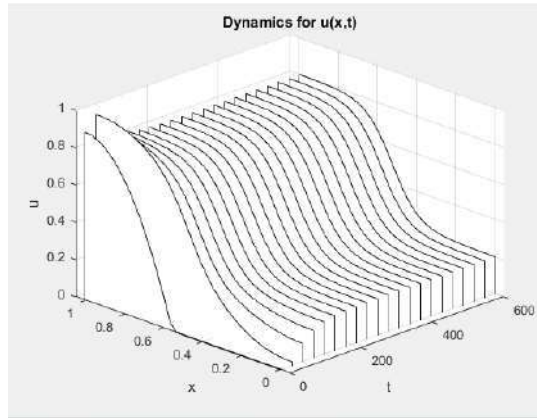


Fig. 8. The dynamics of function  $u(x, t)$  for case 3.

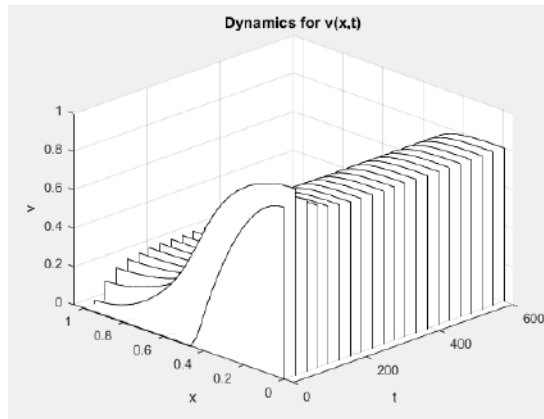


Fig. 9. The dynamics of function  $v(x, t)$  for case 3.

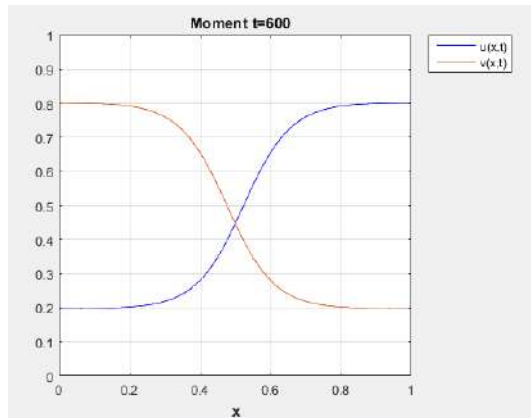


Fig. 10. A solution in a finite time  $t=600$  for case 3.

**6. Conclusion**

Hence, it is shown that for  $q_1 = q_2 < 2$  the manifold of equilibrium states of the system is stabilized, and when passing through the value of the coefficients of the interaction  $q_1 = q_2 = 2$  loss of stability occurs in the system.

**Acknowledgements**

The paper was supported by the Russian Foundation for Basic Research and the government of the Samara region in the framework of a research project № 16-41-630529.



**References**

- [1] Strygin VV, Sobolev VA. Razdelenie dvizheniy metodom integral'nykh mnogoobraziy. Moscow: Nauka, 1988; 256 p.
- [2] Strygin VV, Sobolev VA. Effect of geometric and kinetic parameters and energy dissipation on orientation stability of satellites with double spin. Cosmic Research 1976; 14: 331–335.
- [3] Ayzerman MA, Gantmakher FR. Stabilität der Gleichgewichtslage im einem nicht-holonomen System. Z. angew. Math und Mech. 1957; 37(1/2):74–75.
- [4] Neymark YuI, Fufaev NA. Dinamika negolonomnykh sistem. Moscow: Nauka, 1967; 520 p.
- [5] Kalenova VI, Karapetyan AV, Morozov VM, Salmina MA. Negolonomnye mekhanicheskie sistemy i stabilizatsiya dvizheniya. Fundamental'naya i prikladnaya matematika 2005; 11: 7: 117–158.
- [6] Cantero JC, Korobeinikov A. The spread of Two Viral Strains on a Plant Leaf . Workshop on Virus Dynamics and Evolution.
- [7] Murray JD. Lectures on Nonlinear Differential-Equation Models in Biology. Oxford: Clarendon Press, 1977; 379 p.
- [8] Murray JD. Mathematical Biology I. An Introduction. New York: Springer, 2001; 576 p.

# Conditions for the loss of stability of equilibrium manifold in satellite model

E. Shchepakina<sup>1</sup>, V. Sobolev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The problem of stabilizing a spin satellite by means of passive dampers is considered. The application of the method of integral manifolds allows us to find conditions for the loss of stability in the analytical form.

*Keywords:* stability; stabilization; manifold of steady states; satellite

---

## 1. Introduction

A lot of work has been devoted to the study of dynamic models of stabilization of satellites with the help of gyroscopic forces. As the main apparatus, the Lyapunov function method and the stability criteria applied to first approximation systems are used. In addition to gyroscopic forces for stabilization, damping devices are used in a number of models to ensure the asymptotic stability of the required modes of satellite motion. In a number of works, passive dampers are considered as such devices. For the case of two co-axial bodies, on each of which one damper is installed, the stabilization problem was considered, for example, in [1-3]. In this paper, we confine ourselves to the study of a model of a satellite consisting of two bodies, on one of which a damper with a relatively small coefficient of viscous friction is installed. The damper is modeled by a particle of relatively small mass placed in a tube filled with a viscous liquid and attached by a spring. To analyze the system of differential equations, the method of integral manifolds [3, 4] is applied, which allows to significantly reduce the dimensionality of the model and simplify the analysis.

## 2. Equations

To study the conditions and the mechanism of loss of stability for a satellite stabilized by rotation, consider a dynamic model that is a system of ordinary differential equations for dimensionless variables and parameters of the form [3]:

$$\begin{aligned} q\dot{\omega} - \varepsilon\dot{x}_1 &= \varepsilon[2x_1 v_1 - \omega x_2 u_1], \\ [1 - 2Lu_1]\dot{x}_1 - \varepsilon u_2 \dot{\omega} &= \\ &= -\Lambda x_2 + \varepsilon[-u_1 2L\omega x_2 + \varepsilon x_1 x_2 u_1 + 2Lx_1 v_1], \\ [1 - 2Lu_1]\dot{x}_2 - \varepsilon\dot{v}_1 &= \Lambda x_1 + \varepsilon[\omega^2 u_1 - 2Lu_1 \omega x_1 + 2Lx_2 v_1 - \varepsilon x_1^2 u_1] \\ \dot{u}_1 &= v_1, \\ -\varepsilon\dot{x}_2 + \varepsilon(1 - \varepsilon\rho_1)\dot{v}_1 &= \\ &= -K_1 u_1 - \varepsilon\beta_1 v_1 + \varepsilon(x_1^2 + x_2^2)(u_1 - L) - \varepsilon\omega x_1. \end{aligned}$$

Variables  $\omega, x_1, x_2$  play the role of projections of the absolute angular velocity of the main body on the axis of the coordinate system associated with it with the origin at the center of mass of this body. The variable  $u_1$  characterizes the deviation of a particle moving inside the damper from its nominal position. In these equations, the nonlinear terms containing the factors  $\varepsilon^2 u_1$  are omitted. The value of  $\varepsilon$ , which characterizes the moment of inertia of the mass moving in the damper, plays the role of a small parameter. Some details can be found in [5].

## 3. Manifold of steady states

The system of differential equations under consideration has a manifold of steady states:

$$\mathfrak{M} = \{\omega = \Omega = \text{const}, \quad x_1 = x_2 = u_1 = v_1 = 0\}.$$

Following [6], we say that this manifold is stable with respect to variables

$$x_1, x_2, u_1, v_1,$$

If for any  $\omega = \Omega$  and any neighborhood of zero  $W$  in the space of variables  $x_1, x_2, u_1, v_1$  we can find a neighborhood of zero  $W_0$  of this space such that for any point of this neighborhood the corresponding solution belongs to  $W$  for  $t \geq 0$ .

We will say that  $\mathfrak{M}$  is asymptotically stable with respect to variables

$$x_1, x_2, u_1, v_1,$$

if it is stable with respect to these variables and, in addition, the variables  $x_1, x_2, u_1, v_1$  tend to zero with unlimited increase of  $t$ .

We will say that  $\mathfrak{M}$  is stabilizable if it is asymptotically stable with respect to variables  $x_1, x_2, u_1, v_1$  under  $t \rightarrow \infty$  the solution tends to some point of the manifold  $\mathfrak{M}$ .

It follows from the results of [5, 6] that the manifold of steady states  $\mathfrak{M}$  is stabilizable if all the roots of the characteristic equation, except for one zero root, have negative real parts. Any perturbed motion, sufficiently close to the unperturbed motion, tends to one of the possible steady motions belonging to the indicated manifold if  $t \rightarrow \infty$ .

#### 4. Model reduction

The differential system under consideration is singularly perturbed one and has a three-dimensional manifold of slow motions:

$$u_1 = \varepsilon f(\omega, x_1, x_2), \quad v_1 = \varepsilon g(\omega, x_1, x_2),$$

the motion along which is described by a system of three scalar differential equations of the form:

$$\begin{aligned} q\dot{\omega} &= \varepsilon[2x_1 g - (\Lambda + \omega)x_2 f], \\ \dot{x}_1 &= -\Lambda x_2 + \varepsilon[x_2(x_2 - 2L\omega(\Lambda + \omega)) + f + 2Lx_1 g], \\ \dot{x}_2 &= \Lambda x_1 + \varepsilon[(-K_1 f - x_1(x_1 - 2L\omega(\Lambda + \omega)x_2))f + x_1^2 + x_2^2 - (1 + \rho_1)K_1 + \omega^2)f + \\ &\quad (-\beta_1 + 2Lx_2)g + (\Lambda - \omega)x_1 - L(x_1^2 + x_2^2)] + \\ &\quad \varepsilon^2\{[\omega^2 - (1 + \rho_1)^2 K_1]f - (1 + \rho_1)\beta_1 g + (1 + \rho_1)\omega x_1 - (1 + \rho_1)(x_1^2 + x_2^2)\} + \varepsilon^3(1 + \rho_1)^2(\Lambda - \omega)x_1. \end{aligned}$$

The functions  $f, g$  are computed in the usual way [5]. Restricting ourselves linearly in  $x_1, x_2$  terms to the third order and nonlinear - up to the second order in  $\varepsilon$  inclusive, we write the equations of motion with respect to the integral manifold in the form

$$\begin{aligned} q\dot{\omega} &= \frac{\varepsilon^2}{K_1}[-(\Lambda - \omega)(3\Lambda + \omega)x_2 x_1 + (\Lambda + \omega)Lx_2(x_1^2 + x_2^2)], \\ \dot{x}_1 &= -\Lambda x_2 + \frac{\varepsilon^2}{K_1}[(\Lambda - \omega)x_1^2 x_2 - 2L(\Lambda - \omega)(2\Lambda + \omega)x_2 x_1 + \\ &\quad 2L^2(\Lambda + \omega)x_2(x_1^2 + x_2^2) - Lx_1 x_2 x_1(x_1^2 + x_2^2)], \\ \dot{x}_2 &= \Lambda x_1 + \varepsilon^2[-\frac{1}{K_1}(\Lambda + \omega)(\Lambda - \omega)^2(1 - \frac{\varepsilon L^2}{K_1})x_1 - \frac{\varepsilon}{K_1^2}(\Lambda(\Lambda + \omega)(\Lambda - \omega)^2 x_2 \beta_1) + \frac{1}{K_1}2L(\Lambda - \omega)((\Lambda + \omega)x_1^2 - \Lambda x_2^2) \\ &\quad - 2L(\Lambda + \omega)x_1(x_1^2 + x_2^2) + L(x_1^2 - \omega^2)(x_1^2 + x_2^2)]. \end{aligned}$$

After linearizing the equations on an integral manifold for variables  $x_1, x_2$  we obtain the linear with respect to  $x_1, x_2$  subsystem

$$\begin{aligned} \dot{x}_1 &= -\Lambda x_2, \\ \dot{x}_2 &= \Lambda x_1 + \varepsilon^2[-\frac{1}{K_1}(\Lambda + \omega)(\Lambda - \omega)^2(1 - \frac{\varepsilon L^2}{K_1})x_1 - \frac{\varepsilon}{K_1^2}(\Lambda(\Lambda + \omega)(\Lambda - \omega)^2 x_2 \beta_1)]. \end{aligned}$$

The condition of asymptotic stability with respect to variables  $x_1, x_2$  is

$$-\Lambda(\Lambda + \omega)(\Lambda - \omega)^2 < 0.$$

For the integral manifold of slow motions, the following principle is valid: the variety of stationary states of the initial system is stable (unstable, asymptotically stable with respect to some of the variables, is stabilizable) if and only if it is stable (unstable, asymptotically stable with respect to a part of the variables, stabilizable) the variety of stationary states of a system describing the motion on an integral manifold. It is clear that a violation of the resulting inequality entails a loss of stability. This is confirmed by the results of numerical experiments. In the figures below, one can see oscillations with increasing amplitude for the variables  $x_1, x_2$  and  $\omega$ .

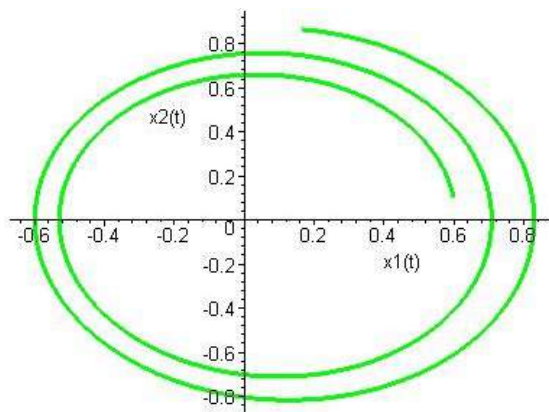
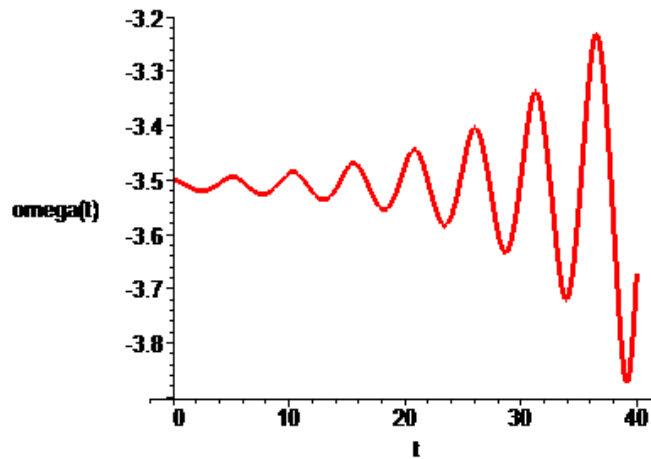


Fig. 1. Projection of the trajectory on the plane of variables  $x_1, x_2$  (the movement is made counter-clockwise).

Fig 2. Solution graph for variable  $\omega$ .

## 5. Conclusion

In the present work, the mathematical model of a satellite stabilized by rotation has been studied by the methods of the geometric theory of singular perturbations. A reduction of the system was carried out, as a result of which, instead of the original system of five differential equations, its projection onto a three-dimensional slow integral manifold was investigated. It should be noted that, due to the validity of the reduction principle for a slow integral manifold, the reduction is carried out correctly, and the reduced system of three differential equations preserves the basic qualitative properties of the original model. An inequality is obtained, in violation of which the satellite loses the required orientation in space.

## Acknowledgements

The study was carried out with the financial support of the RFBR and the Government of the Samara Region within the framework of the scientific project No. 16-41-630524 and the Ministry of Education and Science of the Russian Federation as part of the Samara University's competitiveness increase program (2013-2020).

## References

- [1] Teixeira-Filho DR, Kane DR. Spin stability of torque free systems. Part I, II. *AIAA Journal* 1973; 11(6): 862–867.
- [2] Mingori D. Effect of energy dissipation on the attitude stability of dualspin satellites. *AIAA Journal* 1969; 7(7): 862–867.
- [3] Strygin VV, Sobolev VA. Effect of geometric and kinetic parameters and energy dissipation on orientation stability of satellites with double spin. *Cosmic Research* 1976; 14: 331–335.
- [4] Shchepakina E, Sobolev V, Mortell MP. *Singular Perturbations. Introduction to system order reduction methods with applications*, 2014; 121 p.
- [5] Strygin BB, Sobolev VA. *Decomposition of motions by the Integral Manifolds Method*. Moscow: Nauka, 1988. (in Russian)
- [6] Aizerman MA, Gantmakher FR. Stabilität der Gleichgewichtslage im einem nicht-holonomen System. *Z. angew. Math. und Mech.* 1957; 37(1/2): 74–75.

# Viral evolution model with several time scales

A.A. Archibasov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

In this paper a viral evolution model with specific immune response is considered. By introducing of dimensionless variables and parameters this model can be modified to the singularly perturbed system of partial integro-differential equations with two small parameters. The transition from the initial-boundary value problem of the initial system to the generating problem makes it possible to reduce the dimension of the system and, as a consequence, to reduce the computational volume. The theorem on the passage to the limit is also represented.

*Keywords:* viral dynamics; immune response; specific immunity; singular perturbations; initial-boundary value problem; degenerate system; passage to the limit

---

## 1. Introduction

The presence of several time scales in the models of evolution biology is more a rule than an exception. This is due to the fact that an extremely slow biological evolution process proceeds against the background of significantly faster interactions of different nature. To model such processes with several time scales, systems of differential equations with a small parameter for a part of the derivatives (the so-called singularly perturbed systems of differential equations) are usually used. Numerical analysis of such systems involves a large amount of computation due to the presence of variables that vary with significantly different velocities. Therefore, it becomes relevant to construct simplified (reduced) models of lower dimensionality, but with a high degree of accuracy reflecting the behavior of the original processes.

One of the reduction methods for singularly perturbed systems are the integral manifold method, developed in [1-3], and the passage to the limit to the solution of the degenerate system, used in present paper. In this case, the dimension of the systems under consideration is reduced. Below, this approach is used to reduce the dimension in the initial-boundary value problem for a system describing the dynamics of populations of healthy and infected cells and cytotoxic T-lymphocytes.

## 2. Biological background

A virus is a small infectious agent that is basically composed of a coat of protein, which covers a genetic code (DNA or RNA). A remarkable feature of viruses is inability to replicate themselves. Thus a virus particle attaches to a host cell and injects the genetic material into the cell. Further new virus particles released from the host cell and move around in the infected organism and infect new host cells. When a virus is replicated, mutations happen randomly. A mutant can be seen as a new strain of virus, where a viral strain is a genetic variant or subtype of a virus.

The immune system attacks a virus in order to stop it from growing or to kill it all together. The two main branches of the immune system are humoral and cell-mediated responses. The latter is composed of killer T-cells (also called cytotoxic T-lymphocytes, CTL). Specific immune cells can recognize the physical structure of a pathogen. When discovering the pathogen, the immune cells multiply rapidly in order to kill off the pathogen. Killer T-cells fight infected cells. CTL response is generally considered to be the most effective response of the immune system.

## 3. Model

Let us consider the model of viral dynamics with specific immune response [4]:

$$\begin{aligned}\frac{du(t)}{dt} &= b - u(t) \int_0^{\infty} \beta(s) v(t, s) ds - cu(t), \\ \frac{\partial v(t, s)}{\partial t} &= \beta(s) u(t) v(t, s) - mv(t, s) + \mu \frac{\partial^2 v(t, s)}{\partial s^2} - \xi v(t, s) z(t, s), \\ \frac{\partial z(t, s)}{\partial t} &= qv(t, s) + \gamma z(t, s) \left( 1 - \frac{z(t, s)}{p} \right)\end{aligned}\tag{1}$$

Each virus phenotype is described by a set of parameters and all possible values of these parameters form a phenotype space, which is assumed to be a one-dimensional and continuous:  $s \in [0, +\infty)$  ( $s$  is a dimensionless quantity). Variables  $v(t, s)$ ,  $cell/mm^3$ , and  $z(t, s)$ ,  $cell/mm^3$ , are the population of infected cells of phenotype  $s$  at a time  $t$ , *day*, and the population of CTL-cells, able to kill infected cells of phenotype  $s$  at a time  $t$ , respectively. Uninfected target cells with concentration  $u(t)$ ,  $cell/mm^3$ , are produced at constant rate  $b$ ,  $cell/(mm^3 \cdot day)$ , and have a natural death at a rate  $c$ ,  $1/day$ . Uninfected cells become infected at a rate  $\beta$ ,  $mm^3/(virion \cdot day)$ . The quantity  $IF(t) = \int_0^{\infty} \beta(s) v(t, s) ds$  is called the infective force. Infected cells

die naturally at a rate  $m$ ,  $1/day$ , and are eliminated by CTL response at a rate  $\xi$ ,  $mm^3/(virion \cdot day)$ . The activation term of CTL response is assumed to be proportional to  $v(t, s)$  with a coefficient  $q$ ,  $1/day$ , since the number of infected cells has to be different from zero in order to activate the growth of  $z(t, s)$ . After activation of CTL response the activated cells will multiply by cloning (the so-called ‘‘clonal expansion’’). To model this phenomenon, a logistic term is employed. Random mutations are described by the dispersion with a coefficient  $\mu$ ,  $1/day$ . Since  $v(t, s)$  is a distribution, it is natural to assume that  $v(t, +\infty) = 0$ .

The boundary condition at  $s = 0$  is the non-flux condition  $\frac{\partial v}{\partial t}(t, 0) = 0$  for convenience. Non-negative initial conditions at  $t = 0$  are  $u(0) = u^0$ ,  $v(0, s) = v^0(s)$  and  $z(0, s) = z^0(s)$  (it is assumed that a host is already infected by a virus).

Without loss of generality, for simplicity, we assume that only  $\beta$  depends on  $s$  and that  $m$ ,  $\xi$ ,  $q$ ,  $\gamma$  are constant and have common values for all phenotypes.

Although the model is stated for  $s \in [0, +\infty)$ , the parameters  $s$  is usually assumed to belong a finite interval  $[0, \bar{s}]$ , and the boundary condition  $v(t, +\infty) = 0$  is replaced by the condition  $\frac{\partial v}{\partial s}(t, \bar{s}) = 0$ .

#### 4. The dimensionless system

Let us introduce the following notations  $t = T\bar{t}$ ,  $s = S\bar{s}$ ,  $u(T\bar{t}) = \bar{U}\bar{u}(\bar{t})$ ,  $v(T\bar{t}, S\bar{s}) = \bar{V}\bar{v}(\bar{t}, \bar{s})$ ,  $z(T\bar{t}, S\bar{s}) = \bar{Z}\bar{z}(\bar{t}, \bar{s})$ , and assume that  $\mu T/S^2 = 1$ ,  $\bar{U} = b/c$ ,  $\bar{V} = (\gamma/q)\bar{Z}$ ,  $\bar{Z} = p$ . Then the initial-boundary value problem for model (1) takes the form

$$\begin{aligned} \varepsilon \frac{d\bar{u}(\bar{t})}{d\bar{t}} &= 1 - \bar{u}(\bar{t}) \int_0^{\bar{s}} \bar{\beta}(\bar{s}) \bar{v}(\bar{t}, \bar{s}) d\bar{s} - \bar{u}(\bar{t}), \\ \frac{\partial \bar{v}(\bar{t}, \bar{s})}{\partial \bar{s}} &= \frac{\partial^2 \bar{v}(\bar{t}, \bar{s})}{\partial \bar{s}^2} - \bar{m}\bar{v} + \bar{d}\bar{\beta}(\bar{s})\bar{u}(\bar{t})\bar{v}(\bar{t}, \bar{s}) - \bar{\xi}\bar{v}(\bar{t}, \bar{s}), \\ \varepsilon \theta \frac{\partial \bar{z}(\bar{t}, \bar{s})}{\partial \bar{t}} &= \bar{v}(\bar{t}, \bar{s}) + \bar{z}(\bar{t}, \bar{s})(1 - \bar{z}(\bar{t}, \bar{s})), \end{aligned} \quad (2)$$

$$\bar{u}(0) = \bar{u}^0, \quad \bar{v}(0, \bar{s}) = \bar{v}^0(\bar{s}), \quad \frac{\partial \bar{v}}{\partial \bar{s}}(\bar{t}, 0) = 0, \quad \frac{\partial \bar{v}}{\partial \bar{s}}(\bar{t}, \bar{s}) = 0, \quad \bar{z}(0, \bar{s}) = \bar{z}^0(\bar{s}), \quad (3)$$

where  $\varepsilon = \frac{1}{cT}$ ,  $\theta = \frac{c}{\gamma}$ ,  $\bar{\beta}(\bar{s}) = \frac{pS\gamma}{cq}\beta(S\bar{s})$ ,  $\bar{m} = mT$ ,  $\bar{d} = \frac{bqS}{p\gamma\mu}$ ,  $\bar{\xi} = pT\xi$ ,  $\bar{u}^0 = \frac{cu^0}{b}$ ,  $\bar{v}^0(\bar{s}) = \frac{q}{p\gamma}v^0(S\bar{s})$ ,  $\bar{z}^0(\bar{s}) = \frac{z^0(S\bar{s})}{p}$ ,  
 $\bar{s} = \frac{\bar{s}}{S}$ .

The parameter  $T$  must be taken so that the inequality  $\varepsilon \ll 1$  holds. For example,  $T = 1/\mu$ , then  $S = \sqrt{\mu T} = 1$ . The parameter  $\mu$  is proportional to the mutation probability. For HIV  $\mu$  does not exceed  $10^{-7} - 10^{-9}$   $1/day$ , and HIV is known as one of the most rapidly mutating RNA-viruses, so that  $\varepsilon$  is substantially smaller for more slowly mutating RNA-viruses (and so much the more DNA-viruses). As  $c \ll \gamma$ , then  $\theta \ll 1$ . Thereby system (2) is a singularly perturbed system with two small parameters and as result has three time scales. It should be noted that a system with several time scales was considered in the original work [5]. Further to simplify the notation, we omit the bar.

#### 5. Reduction of dimension

Setting  $\theta = 0$  in (2), we obtain the so-called first-order degenerate system

$$\begin{aligned} \varepsilon \frac{du}{dt} &= 1 - u \int_0^{\bar{s}} \beta v ds - u, \\ \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial s^2} - mv + d\beta uv - \xi v z, \\ 0 &= v + z(1 - z). \end{aligned} \quad (4)$$

The third equation is algebraic and has two roots  $z_{1,2} = (1 \pm \sqrt{1 + 4v})/2$ . For the first-order associated system

$$\frac{\partial \hat{z}}{\partial \tau} = \hat{z}(1 - \hat{z}) + v, \quad (5)$$

where  $v$  enters as a parameter, only one of the roots, namely  $z = \varphi(v) = (1 + \sqrt{1 + 4v})/2$ , is the asymptotically stable (in the sense of Lyapunov) stationary point, because  $\left. \frac{\partial}{\partial \hat{z}}(\hat{z}(1 - \hat{z}) + v) \right|_{\hat{z}=\varphi(v)} = -\sqrt{1 + 4v} < 0$ .

At the initial value of the parameter  $v$ , i.e., at  $v = v^0(s)$ , the system (5) with the initial condition  $\hat{z}(0, s) = z^0(s)$  has a unique solution  $\hat{z}(\tau, s)$  for  $\tau \geq 0$ , and besides  $\lim_{\tau \rightarrow +\infty} \hat{z}(\tau, s) = \varphi(v^0(s)) \quad \forall s \in [0, \bar{s}]$  (see Appendix). Thereby the initial point  $z^0(s)$  of the

first-order associated system (5) belongs to the domain of attraction of the stable stationary point  $\varphi(v^0(s))$ . Thus, for sufficiently small  $\theta$ , problem (2), (3) has a unique solution and, for some  $t_1$ , the following limiting equalities hold [6]:

$$\begin{aligned} \lim_{\theta \rightarrow +0} u(t, \varepsilon, \theta) &= u_0(t, \varepsilon) \quad \text{for } 0 \leq t \leq t_1, \\ \lim_{\theta \rightarrow +0} v(t, s, \varepsilon, \theta) &= v_0(t, s, \varepsilon) \quad \text{for } 0 \leq t \leq t_1, 0 \leq s \leq \cdot, \\ \lim_{\theta \rightarrow +0} z(t, s, \varepsilon, \theta) &= \varphi(v_0(t, s, \varepsilon)) \quad \text{for } 0 < t \leq t_1, 0 \leq s \leq \cdot, \end{aligned} \tag{6}$$

where  $u(t, \varepsilon, \theta)$ ,  $v(t, s, \varepsilon, \theta)$ ,  $z(t, s, \varepsilon, \theta)$  are the solutions of the system (2) and  $u_0(t, \varepsilon)$ ,  $v_0(t, s, \varepsilon)$  are the solutions of the system (4). Note that the third limiting equality holds for  $t \neq 0$ , as the solution  $z = \varphi(v)$  of reduced system (4), generally speaking, does not satisfy initial condition for this variable in (3). The boundary layer phenomenon occurs. Equation (5) is also called the boundary layer equation. Naturally, there is no boundary layer if the initial point falls on the slow surface [7-9]. The system (4) has a dimension one less in comparison with (2).

Let  $\varepsilon = 0$  in (4). Then we obtain the second-order degenerate system

$$\begin{aligned} 0 &= 1 - u \int_0^1 \beta v ds - u, \\ \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial s^2} - mv + d\beta uv - \xi v z, \\ 0 &= v + z(1 - z), \end{aligned} \tag{7}$$

first equation in which is algebraic with respect to  $u$  and has a root  $u = \psi(v) = 1 / \left( 1 + \int_0^1 \beta v ds \right)$ . This root is the asymptotically stable (in the sense of Lyapunov) stationary point of the second-order associated system

$$\frac{d\hat{u}}{d\tau} = - \left( 1 + \int_0^1 \beta v ds \right) \hat{u} + 1. \tag{8}$$

The latter equation with the initial condition  $\hat{u}(0) = u^0$  at the initial value of the parameter  $v = v^0(s)$  has a unique solution  $\hat{u}(\tau) = (u^0 - 1/f) e^{-f\tau} + 1/f$ ,  $f = \psi(v^0(s)) = 1 + \int_0^1 \beta(s) v^0(s) ds$ , for all  $\tau \geq 0$  and  $\lim_{\tau \rightarrow +\infty} \hat{u}(\tau) = 1/f$ . Thus, the initial point  $u^0$  of the second-order associated system (8) belongs to the domain of attraction of the stable stationary point  $\psi(v^0(s))$ . Consequently, for some  $t_2$

$$\begin{aligned} \lim_{\varepsilon \rightarrow +0} u_0(t, \varepsilon) &= \psi(v_{00}(t, s)) \quad \text{for } 0 < t \leq t_2, \\ \lim_{\varepsilon \rightarrow +0} v_0(t, s, \varepsilon) &= v_{00}(t, s) \quad \text{for } 0 \leq t \leq t_2, 0 \leq s \leq \cdot, \\ \lim_{\varepsilon \rightarrow +0} z_0(t, s, \varepsilon) &= \varphi(v_{00}(t, s)) \quad \text{for } 0 < t \leq t_2, 0 \leq s \leq \cdot, \end{aligned} \tag{9}$$

where  $v_{00}(t, s)$  is the solution of the second equation in (7) with boundary and initial conditions for variable  $v$  in (3). The passage to the limit for  $u_0$  is not carried out at point  $t = 0$ . As a result, a system of three integro-differential equations reduces to one integro-differential equation. The existence and uniqueness of the solution of the initial value problem for integro-parabolic equation in (7) can be justified with the use of the approach outlined in the monograph [10].

### 6. Admissibility of the passage to the limit

In work [6] the theorem, that connects the solutions of the singularly perturbed system of partial integro-differential equations with one small parameter, is proved. Generalize this theorem to the case of two small parameters.

Consider the singularly perturbed system of partial integro-differential equations

$$\begin{aligned} \varepsilon \frac{du}{dt} &= f \left( u, \int_0^1 g(s, v) ds \right), \\ \varepsilon \theta \frac{\partial z}{\partial t} &= h(z, v), \\ \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial s^2} + q(s, u, z, v) \end{aligned} \tag{10}$$

with the initial and boundary conditions

$$u(0) = u^0, \quad z(0, s) = z^0(s), \quad v(0, s) = v^0(s), \quad \frac{\partial v}{\partial s}(t, 0) = 0, \quad \frac{\partial v}{\partial s}(t, \cdot) = 0, \tag{11}$$

where  $u, z, v \in R$ ,  $t \in R$ ,  $0 < \varepsilon, \theta \ll 1$  are the small positive parameters.

We assume that system (10) satisfies the following conditions.

I. The functions  $f(u, x)$ ,  $g(s, v)$ ,  $h(z, v)$ , and  $q(s, u, z, v)$ , together with their partial derivatives with respect to all variables, are uniformly continuous and bounded in the respective domains  $\Omega_1 = \{ |u| \leq a, |x| \leq b \}$ ,  $\Omega_2 = \{ 0 \leq s \leq \cdot, |v| \leq c \}$ ,  $\Omega_3 = \{ |z| \leq d, |v| \leq c \}$ ,  $\Omega_4 = \{ 0 \leq s \leq \cdot, |u| \leq a, |z| \leq d, |v| \leq c \}$ .

II. The equation  $h(z, v) = 0$  has an isolated root  $z = \varphi(v)$  in the domain  $\{|v| \leq c\}$  and in this domain function  $z = \varphi(v)$  is continuously differentiable.

III. The inequality  $h_z(\varphi(v), v) \leq -\alpha < 0$  holds for  $|v| \leq c$ . This condition implies that the stationary point  $\hat{z} = \varphi(v)$  of the first-order associated system

$$\frac{\partial \hat{z}}{\partial \tau} = h(\hat{z}, v), \quad (12)$$

which contains  $v$  as a parameter, is Lyapunov asymptotically stable as  $\tau \rightarrow +\infty$  uniformly with respect to  $v$ ,  $|v| \leq c$ . If assumption III is satisfied, then we say for brevity that the zero of the function  $\varphi(v)$  is stable.

IV. There exist a solution  $\hat{z}(\tau)$  of the problem

$$\frac{\partial \hat{z}}{\partial \tau} = h(\hat{z}, v^0(s)), \quad \hat{z}(0, s) = z^0(s), \quad (13)$$

for  $\tau \geq 0$ ,  $0 \leq s \leq \cdot$ . Further, this solution tends to the stationary point  $\varphi(v^0(s))$  as  $\tau \rightarrow +\infty$ , i.e.  $z^0(s)$  belongs to the domain of attraction of the stable stationary point  $\varphi(v^0(s))$ .

V. The equation  $f(u, x) = 0$  has an isolated root  $u = \psi(x)$  in the domain  $\{|x| \leq b\}$  and in this domain function  $u = \psi(x)$  is continuously differentiable.

VI. The inequality  $f_u(\psi(x), x) \leq -\beta < 0$   $\left( x = \int_0^{\cdot} g(s, \varphi(v)) ds \right)$  holds for  $|v| \leq c$ , i.e. the stationary point  $\hat{u} = \psi(x)$  of the second-order associated system

$$\frac{d\hat{u}}{d\tau} = f\left(\hat{u}, \int_0^{\cdot} g(s, \varphi(v)) ds\right), \quad (14)$$

which contains  $v$  as a parameter, is Lyapunov asymptotically stable as  $\tau \rightarrow +\infty$  uniformly with respect to  $v$ ,  $|v| \leq c$ . If assumption VI is satisfied, then we say for brevity that the zero of the function  $\psi(x)$  is stable.

VII. There exist a solution  $\hat{u}(\tau)$  of the problem

$$\frac{d\hat{u}}{d\tau} = f\left(\hat{u}, \int_0^{\cdot} g(s, \varphi(v^0(s))) ds\right), \quad \hat{u}(0) = u^0, \quad (15)$$

for  $\tau \geq 0$ . Further, this solution tends to the stationary point  $\psi\left(\int_0^{\cdot} g(s, \varphi(v^0(s))) ds\right)$  as  $\tau \rightarrow +\infty$ , i.e.  $u^0$  belongs to the domain of attraction of the stable stationary point.

VIII. The truncated system

$$\begin{aligned} \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial s^2} + q(s, \psi(x), \varphi(v), v), \\ u &= \psi(x), \\ z &= \varphi(v), \end{aligned} \quad (16)$$

$$x = \int_0^{\cdot} g(s, \varphi(v)) ds,$$

$$v(0, s) = v^0(s), \quad \frac{\partial v}{\partial s}(t, 0) = 0, \quad \frac{\partial v}{\partial s}(t, \cdot) = 0, \quad (17)$$

has a unique solution  $\bar{v}(t, s)$ ,  $\bar{u}(t) = \psi\left(\int_0^{\cdot} g(s, \varphi(\bar{v}(t, s))) ds\right)$ ,  $\bar{z}(t, s) = \varphi(\bar{v}(t, s))$ .

**Theorem.** If conditions I-VII are satisfied, then, for sufficiently small  $\varepsilon$ ,  $\theta$ , problem (10), (11) has a unique solution  $u(t, \varepsilon, \theta)$ ,  $z(t, s, \varepsilon, \theta)$ ,  $v(t, s, \varepsilon, \theta)$ , which is related to the solution  $\bar{u}(t)$ ,  $\bar{z}(t, s)$ ,  $\bar{v}(t, s)$  of the truncated problem (16), (17) by the limit formulas

$$\begin{aligned} \lim_{\substack{\varepsilon \rightarrow +0 \\ \theta \rightarrow +0}} u(t, \varepsilon, \theta) &= \bar{u}(t) = \psi\left(\int_0^{\cdot} g(s, \varphi(\bar{v}(t, s))) ds\right), \quad 0 < t \leq T, \\ \lim_{\substack{\varepsilon \rightarrow +0 \\ \theta \rightarrow +0}} z(t, s, \varepsilon, \theta) &= \bar{z}(t, s) = \varphi(\bar{v}(t, s)), \quad 0 < t \leq T, 0 \leq s \leq \cdot, \\ \lim_{\substack{\varepsilon \rightarrow +0 \\ \theta \rightarrow +0}} v(t, s, \varepsilon, \theta) &= \bar{v}(t, s), \quad 0 < t \leq T, 0 \leq s \leq \cdot, \end{aligned} \quad (18)$$

Here  $T$  is an arbitrary number such that  $u = \psi\left(\int_0^{\cdot} g(s, \varphi(\bar{v}(t, s))) ds\right)$ ,  $z = \varphi(\bar{v}(t, s))$  are the isolated stable roots of the equations

$f\left(u, \int_0^{\cdot} g(s, \varphi(\bar{v}(t, s))) ds\right) = 0$ ,  $h(\varphi(\bar{v}(t, s)), \bar{v}(t, s)) = 0$  for  $0 \leq t \leq T$  accordingly.

The proof of this theorem is the same as one in [6].



## 7. Conclusion

In this paper the procedure that the original system of three integro-differential equations reduces to a single integro-differential equation is given for a viral evolution model with specific immune response. The theorem on the passage to the limit is also formulated. The limiting equalities for fast variables whose physical meaning is the concentration of populations of healthy cells and killer T-cells are valid only for some segment  $[\delta, T]$ ,  $\delta > 0$ , separated from zero. To construct an approximate solution in a neighborhood of the point  $t = 0$  Tikhonov-Vasil'eva boundary function method [11] can be applied. In the paper [12] a model of viral evolution without immune response (but this model is described by a system of the same type which this work deals with) was considered. By the method mentioned above the solutions in powers of small parameters were found.

It should be noted that the mathematical models of evolution biology are usually formulated as integro-differential equations and PDE. Thus the same concept and the same techniques can be used to a model of evolution based on any other model virus dynamics.

## Acknowledgements

The study was supported by the Russian Foundation for Basic Research and Samara region (grant 16-41-630529-p) and the Ministry of Education and Science of the Russian Federation as part of a program to increase the competitiveness of SSAU in the period 2013-2020.

## Appendix

Let us solve the initial value problem

$$\frac{\partial \hat{z}}{\partial \tau} = \hat{z}(1 - \hat{z}) + v^0(s), \quad \hat{z}(0, s) = z^0(s),$$

where  $\hat{z}(\tau, s)$  is unknown function, functions  $v^0(s)$ ,  $z^0(s)$  are given. This equation is the Riccati equation. Performing successively in the equation of change of variables  $\hat{z} = \hat{z}_1 + \sqrt{1 + v^0(s)}/2$ ,  $y = 1/\hat{z}_1$ , we first bring it to the Bernoulli equation, and then to a linear nonhomogeneous equation of the first order

$$\frac{\partial y}{\partial \tau} = y\sqrt{1 + 4v^0(s)} + 1,$$

with the initial condition  $y(0, s) = 1/(z^0(s) - \varphi^0(s))$ , where  $\varphi(v) = (1 + \sqrt{1 + 4v})/2$ . Solving this linear equation by the variation method, we find that  $y(\tau, s) = \frac{1}{2\varphi^0 - 1} \left( e^{(2\varphi^0 - 1)\tau} - 1 \right) + \frac{1}{z^0 - \varphi^0}$ ,  $\varphi^0 = \varphi(v^0(s))$ ,  $z^0 = z^0(s)$ , then  $\hat{z}(\tau, s) = \varphi^0 + \frac{1}{y(\tau, s)} \rightarrow \varphi^0$  as  $\tau \rightarrow +\infty$ .

## References

- [1] Shchepakina E, Sobolev V, Mortell MP. Reduction methods for chemical systems. Singular Perturbations. Introduction to System Order Reduction Methods with Applications. Lecture Notes in Mathematics 2014; 2114: 111–117.
- [2] Shchepakina E, Korotkova O. Condition for canard explosion in a semiconductor optical amplifier. Journal of the Optical Society of America B: Optical Physics 2011; 28(8): 1988–1993.
- [3] Shchepakina E, Korotkova O. Canard explosion in chemical and optical systems. Discrete and Continuous Dynamical Systems. Series B 2013; 18(2): 495–512.
- [4] Laarhoven N, Korobeinikov A. Within-Host Viral Evolution Model with Cross-Immunity. Extended Abstracts Spring 2015; 119–124. DOI: 10.1007/978-3-319-22129-8\_21.
- [5] Tikhonov AN. Systems of differential equations with small parameters multiplying the derivatives. Mat. Sb. 1952; 31: 575–586.
- [6] Archibasov AA, Korobeinikov A, Sobolev VA. Pasage to the limit in a singularly perturbed partial integro-differential system. Differential Equations 2016; 52(9): 115–1122. DOI: 10.1134/S0012266116090020.
- [7] Shchepakina E, Sobolev V, Mortell MP. Slow integral manifolds. Singular Perturbations. Introduction to System Order Reduction Methods with Applications. Lecture Notes in Mathematics 2014; 2114: 25–42.
- [8] Shchepakina E, Sobolev V. Invariant surfaces of variable stability. Journal of Physics: Conference Series 2016; 727(1).
- [9] Shchepakina E. Stable/unstable slow integral manifolds in critical cases. Journal of Physics: Conference Series 2017; 811.
- [10] Henry D. Geometric theory of semilinear parabolic equations. Berlin: Springer-Verlag, 1981; 358 p.
- [11] Vasil'eva AB, Butuzov VF, Kalachev LV. The boundary function method for singular perturbation problems. Philadelphia: SIAM, 1995; 236 p.
- [12] Archibasov AA, Korobeinikov A, Sobolev VA. Asimptotic expansions of solutins in a singularly perturbed model of virus evolution. Computational Mathematics and Mathematical Physics 2015; 55(2): 240–250. DOI:10.1134/S0965542515020037.

# Generalized model of pulse process for dynamic analysis of Sylov's fuzzy cognitive maps

R.A. Isaev<sup>1</sup>, A.G. Podvesovskii<sup>1</sup>

<sup>1</sup>*Bryansk State Technical University, 50 let Oktyabrya Blvd. 7, 241035, Bryansk, Russia*

---

## Abstract

The article deals with pulse process as a means of dynamic analysis of cognitive models of semi-structured systems. We introduce and substantiate a generalized model of pulse process for Sylov's fuzzy cognitive maps. We offer its implementations for various semantic interpretations of concept influence. The results of experimental validation of the proposed models are presented in the paper.

*Keywords:* cognitive modeling; fuzzy cognitive map; dynamic analysis; pulse process

---

## 1. Introduction

A cognitive approach is one of the approaches to the study of semi-structured systems, which is widely used at the present time. According to the definition given in [1], this approach focuses on the development of formal models and methods supporting the intelligent problem-solving process as they include human cognitive capabilities (perception, conception, cognition, understanding, explanation) in solving management problems. Structure and target modeling and simulation modeling methods based on cognitive approach are commonly subsumed under the umbrella term "cognitive modeling". In general terms, cognitive modeling refers to the study of structure, functioning and development of a system by analyzing its cognitive model. The cognitive model is based on a cognitive map, which reflects researcher's subjective notion (individual or collective) of the system as a number of semantic categories (known as factors or concepts) and a set of cause-and-effect relations between them.

A cognitive model is an effective tool for exploratory and estimative analysis of the situation. It does not give an opportunity to obtain accurate quantitative characteristics of the system under study, but it allows to assess trends related to its functioning and development, and to identify the key factors influencing these processes. Thus, we can search, generate and develop effective solutions for system management, as well as identify risks and develop strategies to reduce them.

Cognitive modeling starts with creating a cognitive map of the system under study on the basis of information received from experts. The next step includes direct simulation. Its main objectives are forming and testing hypotheses for the structure of the system under study, that can explain its behavior, also developing strategies for various situations in order to reach the specified target states.

Tasks solved by means of cognitive modeling can be divided into two groups:

1. Tasks of structure and target analysis:
  - finding the key factors influencing the targets;
  - identification of contradictions between the targets;
  - identification of feedback loops.
2. Tasks of dynamic analysis (scenario simulation):
  - self-development ("what if we do nothing");
  - managed development:
    - direct task ("what if");
    - inverse task ("how to").

Thus, the scenario simulation allows prediction of the simulated system states under different control actions, and search for alternative control solutions bringing the system to the target state.

Mathematical apparatus most commonly used to represent cognitive models and underlying the methods for their analysis is fuzzy logic. As a result, there appeared a whole class of cognitive models based on different types of fuzzy cognitive maps (FCM). A detailed overview of such models can be found, for instance, in monograph [3]. One of FCM varieties, well-proven in practical analyzing and modeling of semi-structured organizational, social and economic systems are Sylov's FCMs. They were firstly proposed in [7] and represent the development of signed cognitive maps [6]. For this type of FCMs there was developed quite a wide range of structure and target analysis methods based on the study of such FCM factors as consonance, dissonance and action. A detailed description of these methods can be found in the original monograph [7], and some examples of their application in the study of different organizational and social systems – in papers [2, 4]. The problem of developing and improving methods of Sylov's FCMs dynamic analysis was given far less attention. This article presents an approach to dynamic analysis of this FCM type using a generalized model of pulse process. The proposed approach is based on the notion of pulse process, originally introduced in [6] for the class of signed cognitive maps. We generalize this concept by extending it to the class of FCMs and develop the approach, first mentioned in [5] and described in more detail in monograph [4] (section 3.2).

## 2. Formal definition and structure of Sylov's fuzzy cognitive map

As previously mentioned, the cognitive model is based on formalization of cause-and-effect relations which occur between factors characterizing the system under study. The result of the formalization represents the system in the form of a cause-and-effect network, termed a cognitive map and having the following form:

$$G = \langle E, W \rangle,$$

where  $E = \{e_1, e_2, \dots, e_K\}$  is a set of factors (also called concepts),  $W$  is a binary relation on the set  $E$ , which specifies a set of cause-and-effect relations between its elements.

Concepts can specify both relative (qualitative) characteristics of the system under study, such as popularity, social tension, and absolute, measurable values – population size, cost, etc. Moreover, every concept  $e_i$  is connected with a state variable  $v_i$ , which specifies the value of the corresponding index at a particular instant. State variables can possess values expressed on a certain scale, within the established limits. Value  $v_i(t)$  of state variable at instant  $t$  is called the state of concept  $e_i$  at the given instant. Thus, the state of the simulated system at any given instant is described by the state of all concepts included in its cognitive map.

Concepts  $e_i$  and  $e_j$  are considered to be connected by relation  $W$  (designated as  $(e_i, e_j) \in W$  or  $e_i W e_j$ ) if changing the state of concept  $e_i$  (cause) results in changing the state of concept  $e_j$  (effect). In this case we say that concept  $e_i$  influences concept  $e_j$ . Besides, if the value increase of the concept-cause state variable leads to the value increase of the concept-effect state variable, then the influence is considered positive (“strengthening”); if to the decrease – then negative (“inhibition”). Therefore, the relation  $W$  can be represented as a union of two disjoint subsets  $W = W^+ \cup W^-$ , where  $W^+$  is a set of positive relations and  $W^-$  is a set of negative relations.

Fuzzy cognitive model is based on the assumption that the influence between concepts may vary in intensity, whereas, intensity may be constant or variable in time. Taking into account this assumption,  $W$  is set as a fuzzy relation, however, its setting depends on the adopted approach to formalization of cause-and-effect relations. A cognitive map with fuzzy relation  $W$  is termed a fuzzy cognitive map.

Sylov's fuzzy cognitive map represents FCM, characterized by the following features.

State variables of concepts can possess values on the interval  $[0, 1]$ .

Influence intensity is considered constant, so relation  $W$  is specified as a set of numbers  $w_{ij}$ , characterizing the direction and degree of influence intensity (weight) between concepts  $e_i$  and  $e_j$ :

$$w_{ij} = w(e_i, e_j),$$

where  $w$  is a normalized index of influence intensity (characteristic function of the relation  $W$ ) with the following properties:

- a)  $-1 \leq w_{ij} \leq 1$ ;
- b)  $w_{ij} = 0$ , if  $e_j$  does not depend on  $e_i$  (no influence);
- c)  $w_{ij} = 1$  if positive influence of  $e_i$  on  $e_j$  is maximum, i.e. when any changes in the system related to concept  $e_j$  are univocally determined by the actions associated with concept  $e_i$ ;
- d)  $w_{ij} = -1$  if negative influence is maximum, i.e. when any changes related to concept  $e_j$  are uniquely constrained by the actions associated with concept  $e_i$ ;
- e)  $w_{ij}$  possesses the value from the interval  $(-1, 1)$ , when there is an intermediate degree of positive or negative influence.

Clearly, FCM of this structure can be graphically represented as a weighted directed graph, which points correspond to elements of set  $E$  (concepts) and arcs correspond to nonzero elements of relation  $W$  (cause-and-effect relations). Each arc has a weight which is specified by the corresponding value  $w_{ij}$ . In this case, relation  $W$  can be represented as a matrix of dimension  $n \times n$  (where  $n$  is the number of concepts in the system), which can be considered as the graph adjacency matrix and is termed a cognitive matrix. In addition, each point of the graph also has a weight, which corresponds to the associated concept state and can change over time.

## 3. Pulse process as a means of dynamic analysis of cognitive maps

Dynamic analysis of cognitive maps is based on modeling of concept state dynamics over time. Besides, concept state may change, firstly, due to changes in the state of other concepts influencing this one, and, secondly, due to external actions. We understand external action as a change of the concept state as to the current one under the impact of external factors, i.e. irrespective of the concepts included in the cognitive map. At the same time external actions can be targeted, i.e. they come from the subject performing system control, and untargeted, i.e. due to uncontrollable factors, external to the system. Thus, in the first case we speak about control actions, and in the second case – about disturbing actions (or disturbance).

To describe the dynamics of concept states we use pulse processes. This approach is based on the assumption that changes in the states of all concepts occur at discrete moments of time. State change of concept  $e_i$  at instant  $t$  is called pulse and is denoted by  $p_i(t)$ . Thus,

$$p_i(t) = v_i(t) - v_i(t-1).$$

It is additionally assumed that influence transmission occurs in one step: changing the state of the concept-cause at instant  $t$  results in changing the state of the concept-effect at instant  $t + 1$ .

Let us first give the model of pulse process for signed cognitive maps, i.e. maps which take into account only the directions of influence but not their intensity. For such maps  $w_{ij}$  can only take values  $-1, 0$  or  $1$ , and the graph arcs are marked with signs “+” and “-”, respectively. The model of pulse process was proposed in [6]:

$$p_i(t+1) = \sum_{j=1}^K \text{sgn}(w_{ji}) p_j(t),$$

accordingly

$$v_i(t+1) = v_i(t) + \sum_{j=1}^K \text{sgn}(w_{ji}) p_j(t).$$

Thus, the state change (pulse) of each concept in the current step is determined by the pulses of all concepts influencing it and by the ratio of influence signs. Moreover, transmission of positive influence is neutralized by simultaneous transmission of negative influence, and vice versa.

In [4, 5], a modified model of pulse process for Sylov's FCM is proposed. The model takes into account both influence transmission between concepts and external actions:

$$v_i(t+1) = \min \left( v_i(t) + u_i(t+1) + q_i(t+1) + \sum_{j=1}^K w_{ji} p_j(t), 1 \right), \quad (1)$$

where  $u_i(t+1)$  is a control action on concept  $e_i$  at instant  $t+1$ ;  $q_i(t+1)$  is disturbance  $e_i$  at instant  $(t+1)$ .

#### 4. Generalized model of pulse process

In the framework of Model (1) it is assumed that the state change of concept  $e_j$  is equal to the difference between its states at the current step and the previous step:

$$p_j(t) = v_j(t) - v_j(t-1).$$

Thus, in dynamic simulation in order to determine the state of dependent concepts we take into account *absolute change* in states of influencing concepts. This approach is acceptable, but at the same time, it is not the only possible one. In this regard, it is advisable to consider other, alternative approaches to interpreting concept influence and propose alternative models of pulse process on their basis.

However, it is necessary to define a number of requirements to models of pulse process, which must be met by all proposed models in the future, regardless of the assumptions which they are based on.

Firstly, a model of pulse process should unambiguously determine the state of an arbitrary concept  $e_i$  at instant  $(t+1)$ , using for this purpose the following available information:

- the state of the same concept  $e_i$  at instant  $t$ ;
- the states of concepts  $e_j, \dots, e_k$ , influencing concept  $e_i$ , at instant  $t$ ;
- the states of these concepts influencing  $e_i$ , at instant  $(t-1)$ ;
- connection weights (influence intensity)  $w_{ji}, \dots, w_{ki}$  among all influencing concepts and  $e_i$ ;
- control and disturbance actions on  $e_i$  at instant  $(t+1)$ , if there are any.

Or, more formally:

$$v_i(t+1) = f(v_i(t), v_j(t), \dots, v_k(t), v_j(t-1), \dots, v_k(t-1), w_{ji}, \dots, w_{ki}, u_i(t+1), q_i(t+1)). \quad (2)$$

Secondly, the following conditions should be met:

- the values of state variables of concepts belong to the interval  $[0, 1]$ , that is  $v_i(t+1) \in [0, 1]$ ;
- if influence intensity between concepts  $e_j$  and  $e_i$  is equal to 0, then changing  $e_j$  state should not cause changing  $e_i$  state;
- if the states of influencing concepts at the previous step did not change ( $v_j(t) = v_j(t-1)$  for all  $j$ ), and there are no control and disturbance actions, then the state of the dependent concept at the current step should not change:  $v_i(t+1) = v_i(t)$ ;
- when the state of the influencing concept increases (decreases) and the relation is positive, the state of the dependent concept should *not decrease (not increase)*:  $v_i(t+1) \geq v_i(t)$  if  $w_{ji} > 0$  and  $v_j(t) > v_j(t-1)$ ;  $v_i(t+1) \leq v_i(t)$  if  $w_{ji} > 0$  and  $v_j(t) < v_j(t-1)$ ;
- when the state of the influencing concept increases (decreases) and the relation is negative, the state of the dependent concept should *not increase (not decrease)*:  $v_i(t+1) \leq v_i(t)$  if  $w_{ji} < 0$  and  $v_j(t) > v_j(t-1)$ ;  $v_i(t+1) \geq v_i(t)$  if  $w_{ji} < 0$  and  $v_j(t) < v_j(t-1)$ ;
- a more significant change of the influencing concept with other factors equal should result in a more significant change of the dependent concept:  $p_i^1(t+1) \geq p_i^2(t+1)$ , if  $p_j^1(t) \geq p_j^2(t)$ ;
- higher intensity of the influence with other factors equal should result in a more significant change of the dependent concept:  $p_i^1(t+1) \geq p_i^2(t+1)$ , if  $w_{ji}^1 \geq w_{ji}^2$ .

Let us call the Expression (2) together with the above mentioned conditions a generalized model of pulse process. This model, on the one hand, comprises Model (1) as a possible particular case, and on the other hand, it provides the basis for building other implementations of the pulse process model.

## 5. Implementation of the generalized model of pulse process

Let us consider alternative implementations of the described generalized model of pulse process, involving different interpretations of concept influence.

### 5.1. Pulse process model, based on relative changes of concept states

Let us assume that concept influence on the system is determined not by the change of its state in general, but by the significance of this change relative to the previous state of this concept. In other words, we consider a relative change of concept states, not an absolute one.

With this view, let us consider pulse  $p_i(t)$  as a relative state change of concept  $e_i$  at instant  $t$ :

$$p_i(t) = \frac{v_i(t) - v_i(t-1)}{v_i(t-1)}.$$

Thus, the value of pulse  $p_i(t)$  shows *by what fraction* of its state at instant  $(t-1)$  concept  $e_i$  has changed.

Now, let us define the way of influence transmission between directly related concepts. Suppose there is a relation between concepts  $e_j$  and  $e_i$ , whose strength is equal to  $w_{ji}$ . To begin with, knowing  $p_j(t)$  – the relative change of state  $e_j$  at instant  $t$ , let us define the relative change of state  $e_i$  at instant  $(t+1)$ .

It is necessary to consider the conditions of the generalized model, and the following additional conditions:

- if  $p_j(t) = 0$  or  $w_{ji} = 0$ , then  $p_i(t+1) = 0$ ;
- if  $w_{ji} = 1$ , then  $p_i(t+1) = p_j(t)$ .

The following operation satisfies these conditions:

$$p_i(t+1) = w_{ji} p_j(t).$$

Finally, let us define the state of concept  $e_i$  at instant  $(t+1)$ . Note that

$$p_i(t+1) = \frac{v_i(t+1) - v_i(t)}{v_i(t)}.$$

So,

$$v_i(t+1) = v_i(t) + v_i(t) w_{ji} p_j(t).$$

The resulting model can be easily generalized in the case of multiple influencing concepts:

$$v_i(t+1) = v_i(t) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t).$$

As one of the conditions of the generalized model is that the concept states range within the interval  $[0, 1]$ , then we should add the following constraints to the model:

$$v_i(t+1) = \max \left( \min \left( v_i(t) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t), 1 \right), 0 \right).$$

Moreover, control and disturbance actions on  $e_i$  should also be defined in terms of relative changes. For example, the control action  $u_i(t+1) = 0,1$  means “to increase the value of  $i$ -concept state variable by 10% of its current value”.

Thus, we obtain the final version of the model:

$$v_i(t+1) = \max \left( \min \left( v_i(t) + v_i(t) u_i(t+1) + v_i(t) q_i(t+1) + v_i(t) \sum_{j=1}^K w_{ji} p_j(t), 1 \right), 0 \right). \quad (3)$$

### 5.2. Multiplicative model of pulse process

Let us consider another model, which also takes into account relative changes of concept states but implies a slightly different interpretation of these changes. This model is not equivalent to that described above, but they both proceed from similar prerequisites.

In this case, relative change of concept  $e_j$  state shows *what fold* this concept changed at instant  $t$  compared with its state at instant  $(t-1)$ :

$$p_i(t) = \frac{v_i(t)}{v_i(t-1)}.$$

Let us define the way of influence transmission between directly related concepts. In this case, the following conditions should be taken into account in addition to those of the generalized model:

- if  $w_{ji} = 1$ , then  $p_i(t+1) = p_j(t)$ ;
- if  $w_{ji} = 0$  or  $p_j(t) = 1$ , then  $p_i(t+1) = 1$ ;

- if  $w_{ji} = -1$ , then  $p_i(t+1) = \frac{1}{p_j(t)}$ .

Exponential operation satisfies these conditions:

$$p_i(t+1) = (p_j(t))^{w_{ji}}$$

Now we can easily determine the state of concept  $e_i$  at instant  $(t+1)$ :

$$v_i(t+1) = v_i(t)(p_j(t))^{w_{ji}}$$

Generalization of the model in case of multiple influencing concepts is the following:

$$v_i(t+1) = v_i(t) \prod_{j=1}^K (p_j(t))^{w_{ji}}$$

This model does not operate on negative values (excluding connection weights used as exponents). This guarantees the fulfillment of the condition  $v_i(t+1) \geq 0$ . To fulfill the other condition of the generalized model, namely  $v_i(t+1) \leq 1$ , let us add the constraint:

$$v_i(t+1) = \min \left( v_i(t) \prod_{j=1}^K (p_j(t))^{w_{ji}}, 1 \right)$$

Control and disturbance actions within this model should be specified on the basis of the interpretation “concept state has changed  $n$ -fold”. For example, the control action  $u_i(t+1) = 2$  means “to double the concept state as compared to its current state”.

Thus, the final version of the model under study is:

$$v_i(t+1) = \min \left( v_i(t) u_i(t+1) q_i(t+1) \prod_{j=1}^K (p_j(t))^{w_{ji}}, 1 \right) \tag{4}$$

### 6. Experimental validation of the discussed pulse process models

For experimental validation and comparison of the examined models, let us perform dynamic analysis of a cognitive map using each of them, with the same initial data.

Fig. 1 shows a fragment of the cognitive map used for the experiment. Connection weights are assigned the following values:  $w_{12} = 0,9$ ;  $w_{23} = -0,8$ ;  $w_{31} = 0,7$ . The initial concept states are specified as:  $v_1(1) = 0,2$ ;  $v_2(1) = 0,3$ ;  $v_3(1) = 0,8$ .

Suppose there is a control action on concept 1, which results in its transition to a state  $v_1(2) = 0,6$ . Influenced by the initial pulse, concept states begin to change in accordance with the rules defined by each model of pulse process.

Fig. 2-4 give graphs of concept state changes during the operation of three models of pulse process. The horizontal axis measures simulation steps; the vertical axis measures the state of the corresponding concept. For the graphs we use the following notations:

- “Model 1” – the results obtained using the additive model (1);
- “Model 2” – the results obtained using the additive model (3) based on relative changes of concept states;
- “Model 3” – the results obtained using the multiplicative model (4) based on relative changes of concept states.

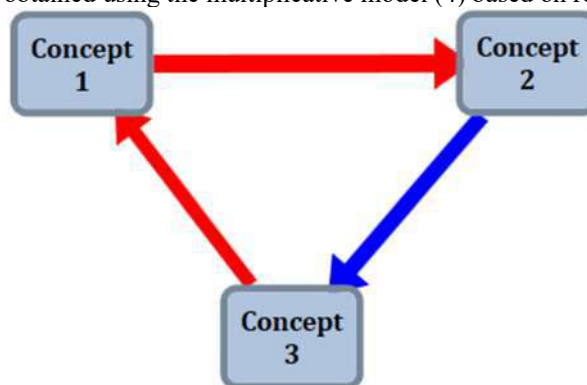


Fig. 1. Fragment of a fuzzy cognitive map used for the experiment.

Of principal interest for interpretation is influence transmission between directly related concepts, differently occurring within different models, eventually providing different results. Thus, in models 2 and 3, implying relative change of concept states, the state of the second concept at the 3rd simulation step increased more than in model 1. Similarly, account taken of relative changes results in more significant decrease in the state of the third concept at the 4th step. Similar regularities are typical for the subsequent steps.

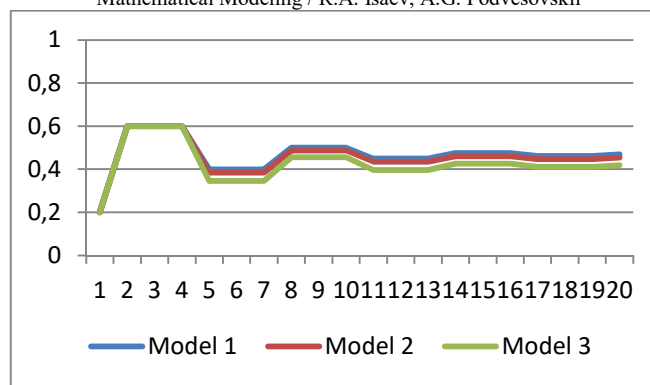


Fig. 2. Dynamics of state change of concept 1.

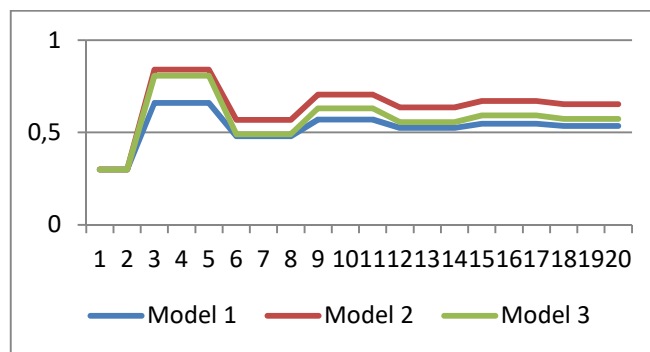


Fig. 3. Dynamics of state change of concept 2.

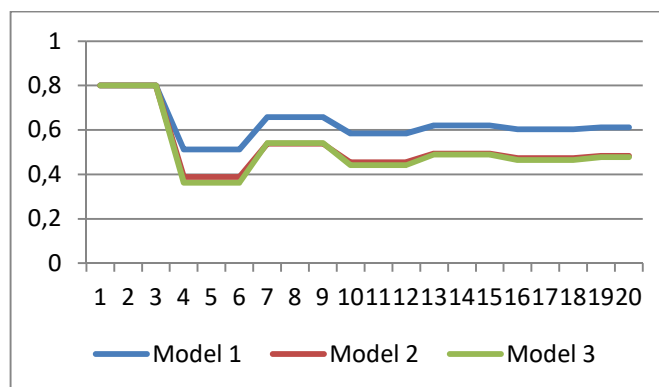


Fig. 4. Dynamics of state change of concept 3.

Describing the results in general, it should be noted that:

- all models operate correctly regarding the influence transmission: the directions of concept state changes correspond to the signs of influences;
- all models are stable: pulse decays with time, which results in the system transition to some stable state;
- as a result of simulation the state of each concept has changed in the same direction for all models (the states of the first and second concepts increased, the state of the third one decreased in comparison with the initial one). These results are generally consistent with the intuitive notion of the system changes pattern, which also validates the models;
- differences in predictions obtained by different models are quite well explained by the underlying prerequisites (concerning the nature of influences among concepts).

## 7. Conclusion

The paper introduces a generalized model of pulse process for Sylov's fuzzy cognitive maps. This model, on the one hand, represents a generalization of previously developed models, and on the other hand, can serve as a basis for building other variations of the pulse process model.

Also, the proposed alternative implementations of the described generalized pulse process model provide diverse interpretations of concept influence. Experimental validation of these implementations has been carried out confirming their correctness and operability.

Among the possible directions for further research, the following are of major interest:

- identifying characteristics and making requirements to the methods of expert identification of FCM parameters in different pulse process models;
- identifying characteristics and making requirements to the methods of identification of FCM parameters on the basis of statistical data in different pulse process models;
- developing methods for selecting an optimal pulse process model based on the analysis of available statistical and expert data.

## References

- [1] Avdeeva ZK, Kovriga SV, Makarenko DI. Cognitive Modeling Approach to Control of Semi-Structured Systems (Situations) in Managing Large Systems 2007; 16: 26–39. (in Russian)
- [2] Averchenkov VI, Kozhukhar VM, Podvesovskii AG, Sazonova AS. Monitoring and Prediction of Regional Demand for Highest Scientific Degree Specialists: monograph. Edited by Averchenkov VI, Kozhukhar VM. Bryansk: Bryansk State Technical University Press, 2010; 163 p. (in Russian)
- [3] Borisov VV, Kruglov VV, Fedulov AS. Fuzzy Models and Networks. Moscow: “Goryachaya Liniya – Telekom” Publisher, 2012; 284 p. (in Russian)
- [4] Erokhin DV, Lagerev DG, Laricheva EA, Podvesovskii AG. Strategic Enterprise Innovation Management: monograph. Bryansk: Bryansk State Technical University Press, 2010; 196 p. (in Russian)
- [5] Podvesovskii AG, Lagerev DG, Korostelyov DA. Application of Fuzzy Cognitive Models for Construction of Alternatives Set in Decision Problems. Bulletin of Bryansk State Technical University 2009; 4(24): 77–84. (in Russian)
- [6] Roberts FS. Discrete Mathematical Models with Application to Social, Biological and Environmental Problems. Prentice-Hall, Englewood Cliffs, 1976.
- [7] Sylov VB. Strategic Decision Making in Fuzzy Environment. Moscow: “INPRO-RES” Publisher, 1995; 228 p. (in Russian)



# The method of augmented regularized normal equations for systems with sparse matrices

S.Y. Gogoleva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

A new approach for solving ill-posed problems is proposed. The approach makes it possible to effectively calculate normal pseudosolutions for ill-conditioned systems of linear algebraic equations and to find an acceptable solution with a minimum filling of sparse matrices.

*Keywords:* regularization method; augmented system; sparse matrices; filling, pivoting

---

## 1. Introduction

Many practical problems of finding solutions based on available data are typical representatives of ill-posed problems. It should be noted that such problems have a number of unpleasant properties of manipulating, and for their solution standard methods are inapplicable. Thanks to the works of academician A.N. Tikhonov developed a general strategy for constructing stable methods for solving ill-posed (unstable problems) in operator form [1]. It is based on the notion of a regularizing operator or a regularizing algorithm. Realizing this algorithm, it is necessary to solve the normal regularized systems of linear algebraic equations. This system is often ill-conditioned. It is necessary to choose the regularization parameter correctly in order to reduce the condition number. It is also important to choose a solution method that is numerically stable. Often ill-posed problems lead to systems with large and sparse coefficient matrices, in which most of the elements are zero.

When storing and manipulating sparse matrices on a computer, it is beneficial and often necessary to use specialized algorithms and data structures that take advantage of the sparse structure of the matrix. Operations using standard dense-matrix structures and algorithms are slow and inefficient when applied to large sparse matrices as processing and memory are wasted on the zeroes. Sparse data is by nature more easily compressed and thus require significantly less storage.

A serious problem in the storage and processing of sparse matrices is the fill-in. The fill-in of a matrix are those entries which change from an initial zero to a non-zero value during the execution of an algorithm. To reduce the memory requirements and the number of arithmetic operations used during an algorithm it is useful to minimize the fill-in.

In this paper we propose an approach using a special form of augmented regularized normal equations. This approach allows solve the system of equations for substantially smaller values of the regularization parameter, as well as to reduce the error of the solution and reduce the fill-in.

## 2. Statement of the Problem

Consider the system linear algebraic equations

$$Ax = b, \quad (1)$$

where  $A \in R^{n \times m}$ ,  $b \in R^n$ .

The regularized solution of the system (1) is found as  $x = \text{Argmin}_{x \in R^m} \{\|Ax - b\|_2^2 + \alpha^2 \|x\|_2^2\}$ , which is equivalent to solving the regularized normal system

$$(A^T A + \alpha^2 E)x = A^T b, \quad (2)$$

where  $\alpha^2$  is a regularization parameter.

The condition number of the system (3) is found as

$$\text{cond}_2(A^T A + \alpha^2 E) = \frac{\sigma_1^2 + \alpha^2}{\sigma_m^2 + \alpha^2},$$

where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m$  are the singular values of  $A$ .

Since the matrix of the system is symmetric, then in the case of well conditionality, it is solved by the Cholesky method. System (2) is often ill-conditioned, then methods based on orthogonal transformations are applied, but they lead to a significant increase in the number of the arithmetic operations. Therefore, instead of system (2), we propose to consider an approach based on an augmented regularized system of equations.

## 3. The Method of Augmented Regularized Normal Equations with Pivoting

Instead of system (3), it is proposed to consider the equivalent system of algebraic equations [2]:

$$\begin{pmatrix} E & A \\ A^T & -\alpha^2 E \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} \quad (3)$$

where  $r = b - Ax$  is the residual vector.

The condition number of the system matrix (3) is slightly less than the condition number of the normal system equations matrix (2). Therefore, in order to reduce the condition number, the parameter  $\beta > 0$  is introduced into the system (3):

$$\begin{pmatrix} \beta E & A \\ A^T & -\frac{\alpha^2 E}{\beta} \end{pmatrix} \begin{pmatrix} r \\ \beta \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} \leftrightarrow C(\beta) = d. \tag{4}$$

A regularized normal system is equivalent to a regularized augmented system. The minimum of the condition number of the matrix (4) is attained for  $\beta_* = \sqrt{\frac{\sigma_m^2}{2} + \alpha^2}$ , where  $\sigma_m$  is the minimal singular number of the matrix  $A$ .

When choosing  $\beta_{**} = \sqrt{\alpha^2}$  the spectral condition number of the system matrix (4) will be  $\sqrt{\frac{\sigma_1^2 + \alpha^2}{\alpha^2}}$ . Thus, this approach make it possible to increase the numerical stability of the problem and to reduce errors in solving of the equations system (1).

The augmented system of equations modification leads to an increase in the dimension of the original problem. Using known methods to solve it leads to computational difficulties. Therefore, it is proposed to consider the modification of the direct projection method [4, 6] with the pivoting, which allows to reduce the number of arithmetic operations to obtain the augmented system of equations solution.

Due to the special structure of the linear algebraic equations augmented system matrix and direct projection method vectors in the augmented system, from  $p = n + m$  equations  $n$  are solved analytically. This means that it is possible to calculate in advance the values of the first  $n$  vectors and indicate the vectors structure in the next steps of the algorithm.

For sparse systems, in order to reduce the fill-in, it is proposed to apply the Markowitz strategy in the direct projection method.[3]

Let the  $k$ -th step of the direct projection method be performed. The number  $r(i, k)$  denotes the number of non-zero entries in the  $i$ -th row of the active submatrix  $C_k$  and  $s(j, k)$  is the number of non-zero elements in the  $j$ -th column of  $C_k$ . The Markowitz count of an entry  $c_{ij}^{(k)}$  is a value

$$M_{ijk} = (r(i, k) - 1)(s(j, k) - 1), (i, j = \overline{1 \dots k}).$$

The count  $M_{ijk}$  is equal to the number of elements that change the value at the transition to the next elimination step, if the entry  $c_{ij}^{(k)}$  is chosen as the pivot one, it is the upper border for the fill-in that occurs when  $c_{ij}^{(k)}$  is selected.

Let

$$M_k = \min\{M_{ijk} \mid i, j = \overline{k \dots n}\}.$$

The Markowitz strategy is that at each step  $k$ , the entry with the Markowitz count  $M_k$  is taken as the pivot.

This does not necessarily mean that the fill-in minimum at the  $k$ -th step will be obtained; however, finding Markowitz count is much easier than calculating the value of the fill for each entry  $C_k$ .

To ensure numerical stability, we will choose the elements of the active submatrix for the role of the pivot, satisfying the condition

$$|c_{ij}^{(k)}| u \geq \max_{k \leq x \leq m, k \leq y \leq n} |c_{xy}^{(k)}|,$$

where it is recommended to select the parameter  $u > 1$ .

Table 1 lists the matrices from the Harwell-Boeing Collection with their characteristics: the size, the number of non-zero elements, and the condition number. [5]

Table 1 . The tested matrices characteristics.

Matrix	Size	Non-zeros	Condition number
ash958	958 × 292	1916	2,1903E+6
flower 8 1	628 × 513	1538	7,0295E+15
ch7-8-b1	1176 × 56	2352	4,7861E+14
mk11-b1	990 × 55	1980	9,8787E+7
well1033	1033 × 320	4732	1,6613E+2
photogrammetry	1388 × 390	11816	4.3591E+08
ash608	608 × 188	1216	1,7661E+6

We give the system of equations solution (1) using the ill-conditioned matrix photogrammetry. The results of the numerical experiment are shown in Table 2.

Table 2. The results for photogrammetry matrix.

Method	Matrix	Pivoting	Relative error	Time, c
Cholesky factorisation	$A^T A$	-	2.8400E-8	20.1600
Direct projection method	C	row	3.7416E-11	42.3949
		Markowitz strategy	3.4203E-12	50.0140
QR factorization	A	-	4.6837E-12	78.6559

From Table 2 we see that the direct projection method for the augmented system with pivoting and the use of the Markowitz strategy yields exactly the same results as the QR method, but requires less execution time.

#### 4. Conclusion

A new approach to solving ill-posed problems is considered. This approach makes it possible to effectively calculate normal pseudosolutions of ill-conditioned linear equations systems and to find an acceptable solution in accuracy. Its modification for this problem, taking into account the sparseness of the augmented system, allows to significantly reduce the number of steps of the algorithm, as well as to reduce the amount of random-access memory and arithmetic operations. The Markowitz strategy in this modification allows to reduce the fill-in of a sparse matrix. This fact significantly simplifies the problem solving and reduces the time for calculation, which is a rather significant advantage.

#### References

- [1] Tikhonov AN, Goncharky AV, Stepanov VV, Yagola AG. Numerical methods for solving ill-posed problems . Moscow: Nauka, 1990; 229 p.
- [2] Zhdanov AI. The method of solving regularized normal equations. Journal of Computational Mathematics and Mathematical Physics 2012; 52(2): 205–208.
- [3] Zlatev Z, Esterby O. Direct methods for sparse matrices. M.: Mir, 1987; 120 p.
- [4] Zhdanov AI. A direct sequential method for solving systems of linear algebraic equations. Russian Academy of Science Report 1997; 356(4): 442–444.
- [5] Harwell-BoeingCollection. MatrixMarket. URL: <http://math.nist.gov/MatrixMarket/data/Harwell-Boeing/> (5.02.2016).
- [6] Gogoleva SY, Zoteeva OV. Solution of the least squares problem on the basis of the augmented equations system method with a sparse matrix. Vestnik SSAU 2008; 2: 175–178.

# On possibilities for studying the problem of human society's evolution using simple mathematical models

L.G. Teklina<sup>1</sup>

<sup>1</sup>Lobachevsky State University of Nizhny Novgorod, Gagarin Ave. 23, 603950, Nizhny Novgorod, Russia

---

## Abstract

The possibilities for transition from an unstable community with periodical crisis phenomena to a globally steady, stable and dynamically developing community in the situation of technical progress are discussed on the base of a simple mathematical model for an isolated “producers–managers–product” community. The essential tool for studying the mathematical model is pattern recognition methods.

*Keywords:* mathematical model; dynamic system; phase and parametric portraits; pattern recognition

---

## 1. Introduction

From our point of view, the willingness to study the evolution of the human society more thoroughly may be implemented by creating and investigating mathematical models capable of explaining the observed reality and suggest possible ways to improve it. Constructing an adequate model for such a sophisticated and diverse object as a human society is the task unlikely to be fulfilled. Instead, simple mathematical models are of interest that can open an opportunity to analyze quite complex objects. At the same time, the evolution of the society can hardly be described as a dynamic system. It is possible to introduce major features of the society and describe their interaction using a dynamic system, completing the model with a significant number of parameters characterizing the society under consideration. This way there will be a possibility to study the evolution of the society depending on the model parameters. A sample of this model is a simple mathematical model of the “producers–managers–product” community given in [1]. A simplified version of this model was analyzed analytically, partially confirmed, and partially supplemented by a small numerical study, whose possibilities turned out to be rather limited for a system with 15 parameters. However, even the incomplete results obtained in this process turned out to be quite interesting. They were analyzed by an historian, confirmed by the facts from the history of global community development and encouraged vivid feedback from the readers [2]. Owing to our attention to this topic and because of the new opportunities for the numerical studies of multidimensional dynamic systems with a large number of parameters we decided to return to this model, but in its full original version.

## 2. Brief presentation of the model

In the study of the model of an isolated “producers (those who actually make the product) – managers (do not make the product but assist in its production) – product (everything necessary for human life, what people consume and use)” community, the values of  $x$ ,  $y$ ,  $z$  are the numbers of producers, managers and products accumulated by the community. The interaction between them is described (roughly and approximately) in the following simple model as a system of three differential equations:

$$\begin{aligned}x &= (a - bx - ly + cz)x \\y &= (-d - mx - ey + fz)y \\z &= \begin{cases} F = g \frac{1 + \varepsilon_1 y}{1 + \varepsilon_2 y} \frac{\mu x}{1 + \delta z} - hx - ky & \text{if } z > 0 \quad \text{or } z = 0 \& F > 0 \\ 0 & \text{if } z = 0 \& F \leq 0 \end{cases}\end{aligned}$$

This model reflects the fact that competing people join in the community for more efficient production of the vital product. The model includes 15 parameters that reflect the level of technology development ( $g$ ), the level of production management ( $\varepsilon_1, \varepsilon_2$ ), the features that take into account an increase in production complexity along with its volume growth and depreciation ( $\mu, \delta$ ), redistribution of the product between producers and managers ( $c, f, h, k$ ), competition inside each group ( $a, b$  and  $d, e$ ), the impact of one group on the other ( $l, m$ ). The detailed description of the model can be found in [1].

## 3. Application of pattern recognition methods to numerical studies of dynamic systems

A new technique of numerically studying dynamic systems by pattern recognition methods with an active experiment is represented in [3,4]. This technique is based on forming selected data on the phenomenon in question using an appropriate mathematical model followed by its pattern recognition analysis. Standard procedures of the technique include the following:

- studying all possible kinds of steady motions in the system phase space (attractors);
- constructing **rough phase portraits** as the total of attractors and the domains of their attraction in the phase space under given parameter values;

- studying the dependence of the rough phase portrait on parameter values by constructing a rough parametric portrait of the dynamic system.

These problems are solved for any mathematical models described by the systems of ordinary differential equations regardless of their specific content. Their solution is formal and partially automated. All the obtained results are statistically reliable with the given probability  $p_0$ . In addition, they may provide the basis for solving non-standard problems that are specific for each mathematical model considered. These problems are normally related to studying the dependence of motions in the system phase space on the model parameters. The algorithm for solving them includes the following stages:

- The formulation of the problem as a task for the analysis and research into the dynamics of attractors or system phase portraits.
- The statement of the problem as a pattern recognition task.
- Forming a learning sample to solve the task in the space of the system parameters based on the data on attractors or system phase portraits.
- The selection of the informative features for solving the problem. Informative features for recognition are those system parameters whose change leads to the transition of an object from one recognizable class to another.
- The search of hidden regularities by using different data mining methods on the set of selected features (constructing decision rules of recognition, cluster and regression analysis etc.).

#### 4. Results of analytical and numerical studies of the mathematical model by pattern recognition methods

A qualitative study of the model as a dynamic system can be defined as the study of the system phase portraits and their dependence on parameters. Finding all possible types of attractors and system phase portraits in their rough version, constructing a rough parametric portrait of the system was fulfilled by numerical methods based on the use of the ideas and algorithms of pattern recognition. All the results are statistically reliable with the probability  $p > 0.99$ . Below are listed some already published and new data on the qualitative study of the model [1,4,5] that are required for a better understanding of the issue.

Attractors, or steady motions in the system phase space under the given parameter values, correspond to the possible steady communities. The system attractors include the equilibrium states (steady and stable communities) and periodic motions (steady but unstable communities). Only three kinds of possible equilibrium states  $(x^*, y^*, z^*)$  were found: a stable community “producers–managers–product”  $P(x^* \neq 0, y^* \neq 0, z^* \neq 0)$ , a stable community «producers–product”  $P_y(x^* \neq 0, y^* = 0, z^* \neq 0)$  and a stable community “producers”  $P_{yz}(x^* \neq 0, y^* = 0, z^* = 0)$ . A more extensive presence was noted of steady periodic motions, when all the three variables periodically change within the range of  $x_{\min} \leq x \leq x_{\max}$ ,  $y_{\min} \leq y \leq y_{\max}$ ,  $z_{\min} \leq z \leq z_{\max}$ , but more often the following three types of cycles are encountered:  $C(x_{\min} > 0, y_{\min} > 0, z_{\min} > 0)$ ,  $C_z(x_{\min} > 0, y_{\min} > 0, z_{\min} = 0)$  и  $C_{yz}(x_{\min} > 0, y_{\min} = 0, z_{\min} = 0)$ . Some cycles are characterized by the periods when some variables stay nearly unchanged. In particular, the moments of achieving and maintaining  $z_{\min} = 0$  can be considered as crisis phenomena.

Rough phase portraits of the dynamic system give us an insight into how stable communities may exist in the society characterized by a specific set of parameters. Our research has shown that the system may include phase portraits with one attractor, these include three types of equilibrium states  $P_{yz}, P_y, P$  and all possible cycles, as well as portraits with two attractors:  $P_{yz} \& P$ ,  $P_{yz} \& C$ ,  $P_y \& P$ ,  $P_y \& C$ . What are the conditions for these or those phase portraits?

The refined **statistically reliable** data on the correlation between parameters for different phase portraits are given in tables 1-

2, where  $g_0 = \frac{h}{\mu} \left( 1 + \delta \frac{am + bd}{bf - cm} \right)$ ,  $\varepsilon = \frac{\varepsilon_1}{\varepsilon_2}$ .

Table 1. Phase portraits under  $bf - cm \leq 0$ .

Parameters correlation	$bf - cm \leq 0$	
	$g \leq h$	$g > h$
$ce - lf < 0$		$P_y, P, C, C_z, C_{yz}, \dots$
$ce - lf \geq 0$	$P_{yz}$	$P_y$

Table 2. Phase portraits under  $bf - cm > 0$ .

Parameters correlation	$bf - cm > 0$		
	$g \leq h$	$h < g \leq g_0$	$g > g_0$
$ce - lf < 0$	$P_{yz}, P_{yz} \& P, P_{yz} \& C$	$P_y, P_y \& P, P_y \& C$	$P, C, C_z, C_{yz}, \dots$
$ce - lf \geq 0$	$P_{yz}, P_{yz} \& P$	$P_y, P_y \& P$	$P$

The obtained results indicate that the following conclusions can be considered **statistically reliable**:

**Conclusion 1.**  $bf - cm > 0$  &  $\varepsilon > 1$  are the necessary conditions for the emergence of steady communities  $P$  and  $C$  when  $g \leq h$ .

**Conclusion 2.**  $ce - lf < 0$  &  $\varepsilon > \varepsilon^*(g)$  are the necessary conditions for the emergence of unstable communities  $C, C_z, C_{yz}, \dots$

**Conclusion 3.**  $g > g^* > \frac{h}{\mu} \left( 1 + \delta \frac{am + bd}{bf - cm} \right)$  &  $bf - cm > 0$  &  $ce - lf \geq 0$  are the sufficient conditions for the existence of globally steady and stable community  $P$  “producers–managers–product”.

**5. Problem statement**

The research into the possible ways of the community evolution under certain conditions comes down to the study of the dependence of attractors and phase portraits on parameters. It is worth considering the possibility of moving from an unstable community with periodic crises to the globally steady, stable and dynamically developing community “producers–managers–product” in the conditions of technical progress. In terms of mathematics, this task can be formulated as follows: which parameters and how should be changed in order to go over from the steady cycle  $C (C_z, C_{yz}, \dots)$  to the globally steady and stable equilibrium  $P(x^* \neq 0, y^* \neq 0, z^* \neq 0)$  which is characterized by the fact that the value  $\frac{z^*}{x^* + y^*}$  increases along with the growth of  $g$ .

Based on the data of the analytical and numerical study shown in tables 1-2 and on the condition of the problem  $g \rightarrow \infty$  (technical progress), we can assume that there is a globally steady cycle of any kind  $C, C_z, C_{yz}, \dots$ , whereas the initial conditions in the phase space may vary. This cycle may exist only provided  $ce - lf < 0$ , but  $bf - cm$  may have any values.

**6. On the transition from an unstable community “producers–managers–product” to a stable one**

Is it possible to move to the stable community provided  $g \rightarrow \infty$  without changing any other system parameters? Regardless of the  $bf - cm$  sign, the specific feature of the globally steady periodic motion (an unstable community) is the fact that with the growth of  $g$  while preserving other parameters instability does not disappear, i.e. the cycle never goes into a stable state of equilibrium. The level of the production power development does not ensure stability in the society. Moreover, the growth of  $g$  is always accompanied by the growing fluctuations and may cause crisis phenomena: the transition of cycle  $C$  to cycles  $C_z, C_{yz}, \dots$ . An example of this process is given in figure 1 where the graph of dependence  $z(t)$  shows the emergence of the time interval when  $z = 0$  as the parameter  $g$  increases. In the presence of crisis phenomena (for example, cycle  $C_z$ ) the growth of  $g$  does not lead to their disappearance either. Furthermore, the crisis phenomena may worsen, e.g., the duration of  $z = 0$  period may be extended as it occurred in the case shown in figure 2. It is curious that similar effects can be also observed when changing the parameters characterizing the level of production management: when  $\varepsilon = \frac{\varepsilon_1}{\varepsilon_2}$  increases, crisis phenomena do not disappear (figure 2). They may even emerge (figure 1), but on the whole the picture is rather complicated: the period of fluctuations becomes longer, as well as the duration of an interval between crises. At times, crisis periods get shorter (but do not disappear), however at the same time crisis phenomena ( $z$  decline, changes in the population of community groups) are even more pronounced.

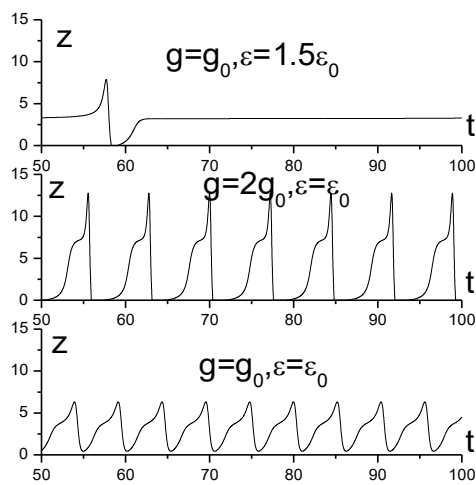


Fig. 1. Changes in  $C$  cycle along with growing  $g$  and  $\varepsilon$ .

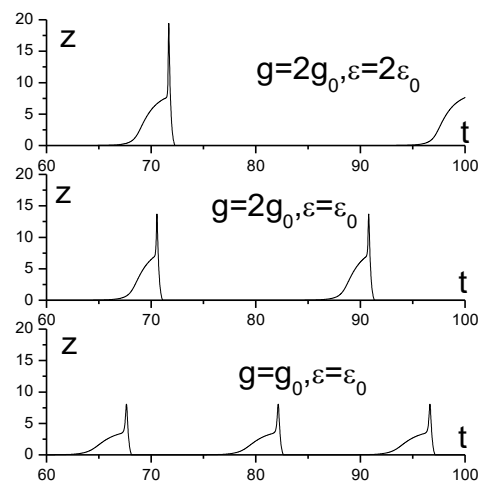


Fig. 2. Changes in  $C_z$  cycle with growing  $g$  and  $\varepsilon$ .

On retention of all the other parameters and the growth of  $g$  the cycle goes to the state of equilibrium only in the case of decreasing  $\varepsilon$ . At the same time, the higher  $g$  is, the lower  $\varepsilon$  must be:  $\varepsilon < \varepsilon^*(g)$ , and under  $g \rightarrow \infty$   $\varepsilon \rightarrow 0$ . Thus, it turns out

that with  $ce - lf < 0$  it is necessary to reduce the level of economic management for the transition to a stable community. But what is this stable community? For this equilibrium with the growing  $g$  there may be two kinds of changes:

- with growing  $g$  (often rather insignificant) the equilibrium transits to the cycle whose changes with the growing  $g$  are described above;

- for small  $\varepsilon$  the equilibrium  $(x^*, y^*, z^*)$  is preserved, but with the growth of  $g$  when  $bf - cm \leq 0$  all the coordinates of the equilibrium decrease (and  $x^*$  decreases faster than the other coordinates), while for  $bf - cm > 0$  similarly  $x^* \rightarrow 0$ , but  $y^*$  grows, which, most likely, is the effect of the idealization of the processes described by the model, since this phenomenon will necessarily lead to a change in the parameters of the model, and, consequently, to a change in the phase portrait.

Considering the aforesaid it is possible to conclude that the transition to the stable (non-crisis) community may occur only when changing the parameters characterizing relations between the groups and the redistribution of the products made. In particular, studies have shown (see section 4) that sufficient conditions for emerging and keeping the globally steady equilibrium (stable community of type  $(x^*, y^*, z^*)$ ) are expressed by the two inequalities  $bf - cm > 0$  and  $ce - lf > 0$ , that can

be written in the form  $\frac{l}{e} < \frac{c}{f} < \frac{b}{m}$  and must be fulfilled under all  $g > g^* > \frac{h}{\mu} \left( 1 + \delta \frac{am + bd}{bf - cm} \right)$ .

## 7. On the possibility for the existence of a dynamically developing stable community

$bf - cm > 0$  &  $ce - lf > 0$  are statistically reliable sufficient conditions for the existence of a rough phase portrait with the only steady equilibrium  $(x^*, y^*, z^*)$  whose coordinates increase alongside with the growth of  $g$ . However, in this case there are

two options of the change in the value of  $\chi = \frac{z^*}{x^* + y^*}$ : a gradual rise or a very slow fall. What are the conditions for the first version of  $\chi$  changes, which might serve as an estimate of the effectiveness of the existing community? To answer this question, we should look at the derivative  $\frac{\partial \chi}{\partial g} = \chi'_g$ .

The equilibrium coordinates satisfy the equations

$$\begin{cases} a - bx^* - ly^* + cz^* = 0 \\ -d - mx^* - ey^* + fz^* = 0 \end{cases} \text{ wherefrom } x^* = \frac{ae + dl}{be - lm} + \frac{ce - fl}{be - lm} z^* \text{ and } y^* = \frac{am + bd}{lm - be} + \frac{cm - bf}{lm - be} z^*.$$

Expression for  $\chi$  looks as follows:  $\chi = \frac{z^*}{x^* + y^*} = \frac{(be - lm)z^*}{a(e - m) + d(l - b) + (c(e - m) + f(b - l))z^*}$  and, consequently,

$$\chi'_g = \frac{(be - lm)(a(e - m) + d(l - b))(z^*)'_g}{(a(e - m) + d(l - b) + (c(e - m) + f(b - l))z^*)^2}.$$

Provided that  $z^*$  increases with the growing  $g$ , i.e.  $(z^*)'_g > 0$ ,  $\chi'_g > 0$  if  $(be - lm)(a(e - m) + d(l - b)) > 0$ .

From the conditions for the existence of the community under consideration, expressed by inequalities  $\frac{l}{e} < \frac{c}{f} < \frac{b}{m}$ , it follows that  $be - lm > 0$ , i.e.  $\chi'_g > 0$  if  $a(e - m) + d(l - b) > 0$ .

Thus, the conditions for the existence of a globally steady, stable and economically efficient community “producers-managers-product” may be described with the following inequalities for the parameters describing this community:

- 1)  $g > g^* > \frac{h}{\mu} \left( 1 + \delta \frac{am + bd}{bf - cm} \right)$
- 2)  $bf - cm > 0$  &  $ce - lf > 0$ , or  $\frac{l}{e} < \frac{c}{f} < \frac{b}{m}$ ;
- 3)  $g'_i > 0$  - technical progress;
- 4)  $a(e - m) + d(l - b) > 0$

Since the above conclusions are based not only on the analytical results, but also on numerical experiments, a control analysis of a large statistical sample (over 30000 cases) was carried out. When conditions (1-2) were met, the rough phase portrait consisted of the only steady equilibrium  $(x^*, y^*, z^*)$  whose coordinates increased when condition (3) was satisfied, and when condition (4) was fulfilled, the value of  $\chi = \frac{z^*}{x^* + y^*}$  also increased.

## 8. Qualitative description of the inequalities expressing the conditions for the existence of a globally steady, stable and economically efficient community “producers–managers– product”

The simple model in question reflects the essential connections and interactions of the three main elements of the human community: producers– managers– product, but it is very difficult to choose a quantitative equivalent for them. What is there behind the resulting formulas? What features and relations between the groups are reflected by inequalities (1-4)?

It is evident that inequality (1) requires a certain level of productive forces development, while inequality (3) means the development of the community under technological change. Inequalities (2) and (4) relate the values that reflect production distribution and competition between the groups to the values characterizing relations within each group. These relations should be considered in detail.

Thus, inequality (2) implies that *an increase in the share of the product produced for one of the groups (redistribution of products between producers and managers) must be either insignificant or must be accompanied by strong competition within this group and/or a decrease in the pressure on another group.*

The inequalities under consideration also impose constraints on the correlation between the values that characterize the competition within the group and the pressure on it, i.e. between  $b$  and  $l$ , between  $e$  and  $m$ .

With  $m \geq e$  and  $b \geq l$   $a(e-m)+d(l-b) \leq 0$ , inequality (4) fails, whereas with  $m \geq e$  and  $b < l$  inequality (2) fails, because it should be  $l < \frac{be}{m} = b \frac{e}{m} \leq b$ . Consequently, inequalities (2,4) require  $e > m$ , when *the competition between managers must be higher than producers' pressure on them.*

Suppose  $e > m$ . If  $b \geq l$ , inequality (2) is fulfilled by increasing  $b(e)$  or decreasing  $l(m)$ , while inequality (4), which may be recorded as  $\frac{b-l}{e-m} < \frac{a}{d}$ , by increasing  $e$  or reducing  $m$ , i.e. for the simultaneous achievement of the required correlations it is possible to increase competition between managers. If  $b < l$ , then inequality (4)  $a(e-m)+d(l-b) > 0$  is fulfilled, and to fulfil

inequality (2)  $\frac{l}{e} < \frac{c}{f} < \frac{b}{m}$  it is possible to increase  $e$  or decrease  $m$ . In addition, when  $b < l$ , inequality (2) imposes the

constraint on  $l$ , namely:  $l < b \frac{e}{m}$ , i.e.  $b < l < b \frac{e}{m}$ . Therefore, for producers the correlation between  $b$  and  $l$  may differ, but at the same time the *pressure on producers must be limited, and if the pressure increase is desirable, and sometimes mandatory, it is to be accompanied by an increased competition between managers.*

## 9. Conclusion

When discussing the possibilities of studying the objects described by mathematical models, we do not focus on the issue of the adequacy of the models considered (this is to be done by experts from respective fields). Though it should be noted that our research gives additional material for assessing model significance. Nonetheless, the above results, in our opinion, firstly, confirm the wide possibilities of applying new statistical methods in the study of mathematical models, especially those with a large number of parameters, which is a topical problem that can hardly be solved in the general case, and, secondly, the results of this research, without claiming to be complete and comprehensive, allow us to identify significant factors affecting the dynamics of the object under analysis. The data on the community obtained as a result of studying its mathematical model, provide a new knowledge of the possible ways of our society's development and give an insight into the processes taking place in society.

It should be emphasized once again that the results of the research presented in this paper refer to an isolated society and do not take into account any specific features of different communities' interaction. However, even these results are indeed thought-provoking.

## References

- [1] Neimark YuI. Mathematical Models in Natural Science and Engineering. Springer, 2003; 570 p.
- [2] Neimark YuI, Levin AYa. Does God Play Dice? Izvestiya VUZ. Applied nonlinear dynamics 2009; 17(3): 98–136. (in Russian)
- [3] Neimark YuI, Kotel'nikov IV, Teklina LG. Coarsened statistical study of applied dynamical systems using pattern recognition methods (part I). Vestnik of Lobachevsky University of Nizhni Novgorod 2012; 5(2); 159–171. (in Russian)
- [4] Teklina L, Kotel'nikov I. Analysis and synthesis of dynamic systems using methods of pattern recognition. LAP LAMBERT Academic Publishing, 2015; 129 p. (in Russian)
- [5] Neimark YuI, Kotel'nikov IV, Teklina LG. Coarsened statistical study of applied dynamical systems using pattern recognition methods (part II). Vestnik of Lobachevsky University of Nizhni Novgorod 2012; 6(1): 164–174. (in Russian)



# Mathematical model of power characteristics of the diagnostic fluorimeter

V.N. Grishanov<sup>1</sup>, V.S. Kulikov<sup>1</sup>, K.V. Cherepanov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

A mathematical model for evaluation of power characteristics of the optical link of the fluorimeter is developed. The main objective of fluorimeter is measurement of intensity of fluorescent radiation of human skin in vivo. The model is realized in a packet of computer mathematics Mathcad and consists of the units modeling energetic characteristics of passive optical elements, radiators and photodetectors by analytic functions – laws of photometry. For creation of models elements reference, literary and experimental data on them are used. Basic purpose of model – operational quantitative comparing of constructive solutions for instrument by energetic criterion – photodiode's output signal. The given obviously mathematical functions provide openness of model and accessibility for modification by the user.

*Keywords:* mathematical model; radiation stream; laser; light-emitting diode; photodiode; light filter; fluorimeter; photometry

## 1. Introduction

Human skin is the most available object of diagnostics in vivo. The skin integument is a peculiar accumulator of the products which reflecting the processes happening in an organism. Measurement of skin autofluorescence is demanded by physicians for an assessment of maintenance of the Advanced glycation end products (AGE). According to their contents complications in case of diabetes, coronary heart disease, operations on renal transplantation and a chronic hemodialysis are predicted [1 - 4]. Measurement of maintenance of AGE on the AFR level is used in dermatology for determination of a biological age of skin [5] and an assessment of activity of processes of biooxidation in fabrics [6]. Fluorescence of AGE is excited by radiation from the spectral range of 300 - 420 nanometers, and highlighting of radiation of fluorescence of AGE is watched in the spectral range of 420 - 600 nanometers.

In the West measurement of maintenance of AGE on the AFR level is accepted as one of required parameters for prediction risk of complications in cardiovascular system of the diabetic [7] and received the instrumental support in the form of family of the instruments AGE Reader of the DiagnOptics Technologies B.V company. [8], by means of which researches [1-3] are conducted. The instruments AGE Reader have the sizes of the netbook and are easy-to-work. The procedure of diagnostics doesn't exceed 5 minutes and consists that the patient puts forearm inside to an optical window of the instrument, and the operator clicks "Start-up". Results of diagnostics are displayed on a panel of the instrument and fixed in its memory. Fluorescence of AGE is excited by the mercury lamp, and is elastic dispersed by skin and fluorescent the radiations reaching an input end face of receiving optical fiber are transferred to them to a compact spectrometer. Diagnostic parameter in the instruments AGE Reader the integral criterion of the AFR [1,2] level appears

$$AU = \frac{\int_{420}^{600} I_f(\lambda) d\lambda}{600 - 420} \times \frac{420 - 300}{\int_{300}^{420} I_{bs}(\lambda) d\lambda}, \quad (1)$$

where  $I_f(\lambda)$  – a range of intensity of fluorescent radiation of skin in the range of lengths of waves (420 – 600) nanometers;  $I_{bs}(\lambda)$  – a range of intensity is elastic the radiation of excitation of fluorescence reflected by skin in the range of lengths of waves (300 – 420) nanometers. The experimental ranges of  $I_f(\lambda)$  and  $I_{bs}(\lambda)$  register the spectrometer which is an AGE Reader part.

Due to the lack of the AGE Reader equipment available to most medical institutions in Russia similar operations are carried out only at the research level [4-6, 9]. Researches are conducted on original universal spectrofluorometers [10-14] which operation assumes an involvement in it a highly qualified staff, there was no consensus also by diagnostic criterion [9].

In the Samara university with the assistance of authors of the real operation the diagnostic fluorimeter capable to solve the problem of measurement of AFR caused by AGE and the implementing integral diagnostic criterion (1) is created. From original, universal, research spectrofluorometers it shall differ in compactness, simplicity of construction and operation at the expense of optimized under the decision of the task set above by optical, electronic and algorithmic structures, and from instruments of the AGE Reader family - the budgetary element basis and easy replicability. The on-stage performance group managed to create two prototypes of the fluorimeter [15,16] meeting the advanced criteria, first of which, single-channel, allowed to validate experimentally the made circuitry decisions by convincing demonstration of its ability to register very feeble radiation of AFR in the presence of destabilizing operational factors, and the second, being dual-channel, - as in laboratory, and in case of approbation in Regional clinical hospital of V. D. Seredavin showed ability to register age features of AFR and pathological processes at the patients having coronary heart disease.

As development is in a stage of optimization of designer decisions now, creation of simple mathematical models of the principal structural components of the diagnostic fluorimeter is necessary for operational comparing of possible modifications of optical and optical-electronic elements and their relative positioning.

## 2. Object of simulation

The optical circuit of the diagnostic fluorometer [16] is provided in fig. 1. Excitation of fluorescence of AGE which are contained in skin – a research object 5, is carried out by the radiation of a short-range ultra-violet or violet LED or the junction laser 1 which passes through the collimating optics 2 and the clearing light filter 3. The range of radiation of excitation of AFR of 350 - 415 nanometers is caused by the AGE fluorescent properties [1-2, 17], and in the specified range noticeable spectral non-uniformity efficiency of AFR excitation isn't marked that causes a designer level of freedom in a radiator choice.

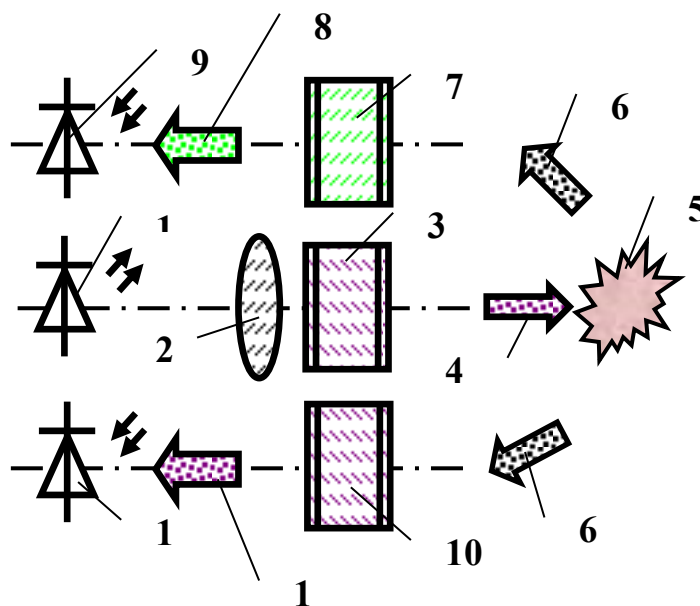


Fig. 1. Optical circuit of the diagnostic fluorometer: 1 – radiation source; 2 – the collimating optics; 3 – the clearing light filter; 4 – autofluorescence excitation radiant flux; 5 – the researched object; 6 – a compound it is elastic scattered and fluorescent radiations; 7 – the light filter which is cutting off fluorescence excitation radiation; 8 – flow of fluorescent radiation; 9 – photodiode of the channel of measurement of intensity of fluorescent radiation; 10 – the light filter which is cutting off fluorescent radiation; 11 – a flow it is elastic scattered radiation; 12 – the photodiode of the channel of measurement of intensity it is elastic scattered radiation.

Requirements of compactness and small energy consumption are narrowed by a choice to semiconductor sources of radiation. However it is impossible to recognize this restriction essential if for the solution of the main objective it is possible to pick up a LED or the junction laser of mass production, then on set of optical, operational and economic parameters it and will be, in most cases, the optimal designer solution [18].

Assignment of the clearing light filter 3 set in an exciting branch of the optical circuit consists in suppression of parasitic long-wave radiation which range is superimposed on a range of AFR [19]. Presence of additional long-wave radiation is characteristic of commercial UF and a blue range LEDs and its nature is described in operation [20]. It is caused by a radiant recombination in the upper layer of p-GaN heterostructure of a LED. Intensity peak of parasitic long-wave radiation is in the ultra-violet and violet ranges next LEDs: LEUVS33G10TZ00, FYL-5013UVC, T5F36, EOLD-365-525, by the experimental estimates [19] made from  $1 \cdot 10^{-3}$  to  $7,5 \cdot 10^{-3}$  from intensity peak of the main radiation with a maximum on wavelength, inhering to an interval (560; 580) nanometer. Elements 1 – 3 form optical link of excitation of AFR. A part of the radiation dispersed by skin 6 through the light filter which is cutting off exciting radiation 7 falls on the photodiode 9 of the channel of measurement of intensity of AFR. The signal of the photodiode 9 is proportional to numerator of expression (1). Other part of the radiation dispersed by skin through the light filter which is cutting off fluorescent radiation 10 falls on the photodiode 12 of the channel of measurement of intensity is elastic scattered radiation. Its signal is proportional to a denominator of expression (1). The sketch of an optical system of the fluorometer is provided in fig. 2.

Object of simulation is also the optical system provided in fig. 1 and 2. Owing to geometrical symmetry of construction to channels for measurement intensity of elastic and fluorescent radiations the mathematical model of energetic characteristics shall describe dependence of the radiant flux falling on the photosensitive platform on one of photodiodes, and, therefore, and value of its output signal from spatial and energetic radiation parameters of excitation AFR set of such parameters of the photodiode as the sizes of its photosensitive site and indicatrix of sensitivity taking into account the dispersing research object properties. Optimized by means of model of energetic characteristics will be two key design parameters of an optical system: distance between optical axes of a excitation fluorescence source and the photodiode  $b$  and distance between a surface of the researched object and an input window of the photodiode.

The mathematical model of spectral characteristics shall provide an assessment photodiodes suitability on their spectral sensitivity; selection of type, materials and thickness of light filters for spectral dependences of their absorption coefficients or synthesis of a spectral characteristic of passage an interference light filter. Thus, important project and designer problems are also solved with its help.

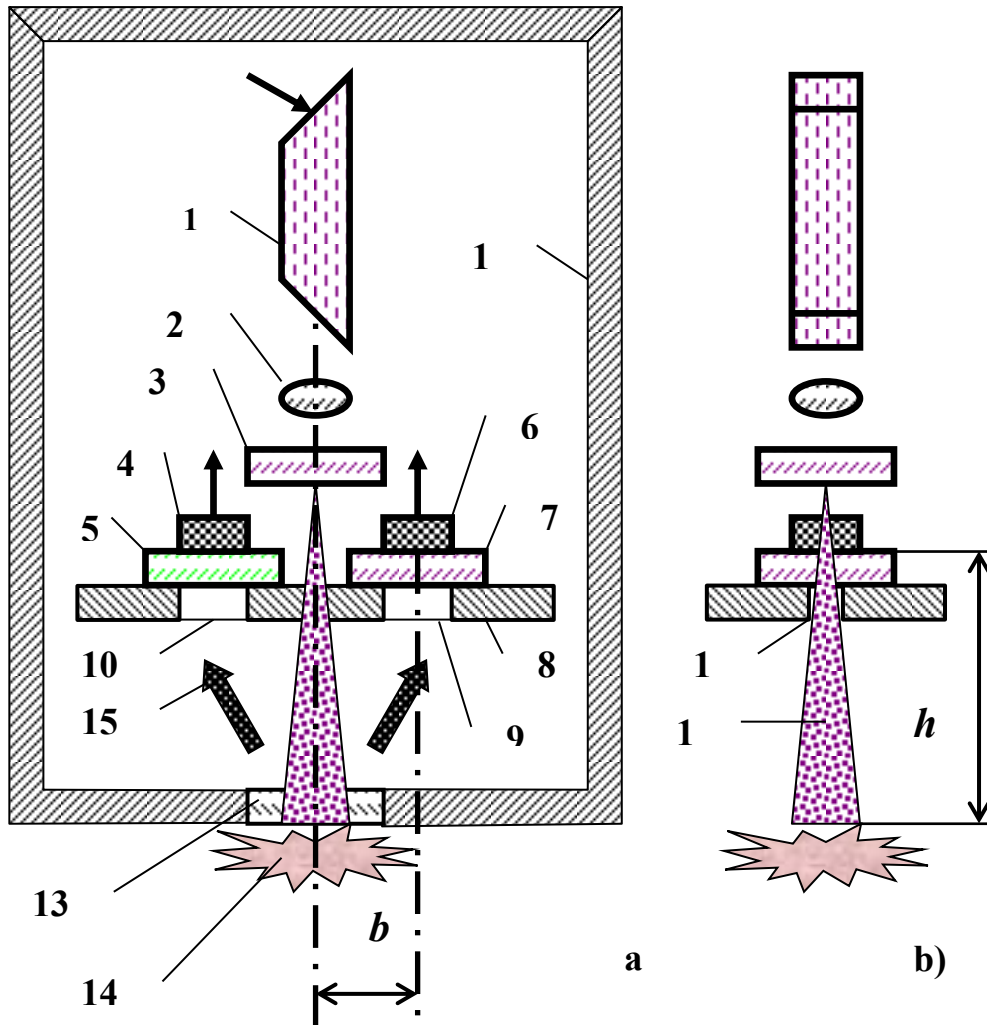


Fig. 2. Construction of an optical system of the diagnostic fluorometer: a) – frontal look; b) – side view; 1 – radiation source; 2 – the collimating optics; 3 – the clearing light filter; 4 – photodiode of the channel of measurement of intensity of fluorescent radiation; 5 – the light filter which is cutting off fluorescence excitation radiation; 6 – the photodiode of the channel of measurement of intensity it is elastic scattered radiation; 7 – the light filter which is cutting off fluorescent radiation; 8 – mounting plate; 9, 10, 11 – holes in a mounting plate; 12 – lightproof casing; 13 – optical window; 14 – research object; 15 – a compound it is elastic scattered and fluorescent radiations; 16 - autofluorescence excitation radiant flux; b – distance between optical axes of a source of excitation of fluorescence and the photodiode; h – distance between a surface of the researched object and an input window of the photodiode.

### 3. Mathematical model of the energy characteristics

Let us demonstrate the structure and operation of the energy characteristic model on the problem of optimizing the  $b$  and  $h$  design parameters. We propose using a laser module that includes the SLD3233VF semiconductor laser [21] and a built-in adjustable collimator as a radiation emitter in the design under development. Its peak emission wavelength is 405 nm, and the peak emission power in a continuous mode is equal to 65 mW. Due to the adjustable collimator, the emission power density in the examined skin area and the effective characteristic beam size can be varied within an order of magnitude. Therefore, the radiation emitter exciting SAF (skin autofluorescence) in the present paper is modeled by distributing the  $E(x, y)$  power density-irradiance-over the object of investigation.

We expect to use as the emission detector the BPW21R silicon photodiode with an integrated light filter that shifts the peak of its spectral sensitivity to a wavelength of 560 nm, which matches better with the AGE (advanced glycation endproducts) fluorescence spectrum. Since the photodiode manufacturers provide its  $S(\alpha)$  sensitivity indicatrix [22], it is logical to consider a model of a point-contact photodetector the photodiode model to analyze the fluorimeter energy. The point-contact photodetector model features a defined direction of the sensitivity indicatrix axis, along which the sensitivity reaches its peak  $S_{max} = 1$  value and the  $\alpha$  angle, which determines the direction to the radiation emitter, is measured from it as well. The analysis of the angular dependence of  $S(\alpha)$ , given in graphical form in [22], showed that it is not differentiated from cosine:

$$S(\alpha) = \cos \alpha, \quad (2)$$

The geometry of an optical system model is shown in Fig. 3. The object of investigation is considered flat and located in the  $xOy$  plane. The skin test area has the shape of a square with  $\alpha$  side. The origin of the reference system coincides with the geometric center of the examined area. The photodiode is located at a  $P$  point belonging to the  $yOz$  plane and is at a  $b$  distance from the  $z$  axis that is equal to the length of  $PH$  or  $OB$  segments and at a  $h$  height from the  $xOy$  plane equal to the length of  $PB$  or  $HO$  segments. The axis of the sensitivity indicatrix is directed vertically down along the  $PB$  segment.

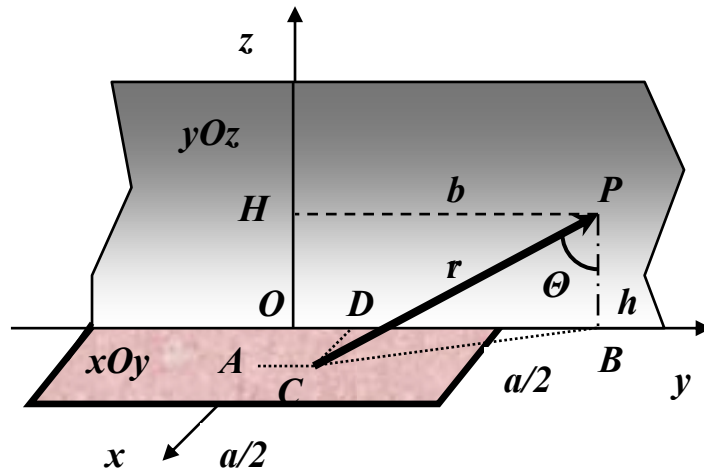


Fig. 3. Geometry of mathematical model of power characteristics.

The element of the  $dx \times dy$  scattering surface with the center at the  $C$  point, the boundary point of the  $CP$  segment that connects the surface element and the photodiode serves as an elementary radiation emitter for the photodiode. The  $A$  and  $D$  points are projections of the  $C$  point onto the corresponding axes of reference. The  $CB$  segment is a projection of the  $CP$  segment onto the  $xOy$  plane.

Since the surface of an examined object – skin – can be assumed to be Lambertian [23] in the first approximation with the  $p$  reflectance at the angles of radiation incidence and radiation scattering up to  $70^\circ$ , its  $L$  brightness will not depend on the scattering angle, and a brightness value of the surface element will be determined by its irradiance [24]:

$$L(x, y) = \rho E(x, y) / \pi, \quad (3)$$

Then the  $dx \times dy$  element will have normal radiation intensity:

$$I_0(x, y) = L(x, y) dx dy = (\rho / \pi) E(x, y) dx dy, \quad (4)$$

and the radiation intensity itself will vary according to the cosine law:

$$I(x, y, \theta) = I_0(x, y) \cos \theta = (\rho / \pi) E(x, y) \cos \theta \cdot dx dy, \quad (5)$$

where  $\theta$  is the angle between the  $CP$  segment and the normal to the  $xOy$  plane. Denoting the length of the  $CP$  segment by  $r$ , we obtain the irradiance created by the  $dx \times dy$  element at the location of the  $P$  photodiode:

$$E_p(x, y, \theta) = I(x, y, \theta) / [r(x, y)]^2, \quad (6)$$

Denoting the irradiance transfer ratio into an electrical signal at the  $k$  output of the photodiode, and taking into consideration its  $S(\alpha)$  sensitivity indicatrix, we will have the following dependence of the  $U$  output signal on the design parameters:

$$U = (\rho k / \pi) \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} \{E(x, y) / [r(x, y)]^2\} \cos \theta \cdot \cos \alpha \cdot dx dy, \quad (7)$$

Since the axis of the sensitivity indicatrix is perpendicular to the  $xOy$  plane, then we have  $\alpha = \theta$  and from  $\triangle CPB$ :

$$\cos \alpha = PB / PC = h / r. \quad (8)$$

From  $\triangle CPB$  and  $\triangle CDB$  we obtain the  $r(x, y)$  dependence in explicit form:

$$r(x, y) = \sqrt{CB^2 + PB^2} = \sqrt{DB^2 + CD^2 + PB^2} = \sqrt{x^2 + (b - y)^2 + h^2}, \quad (9)$$

which substituted in (8) and (7), results in the expression for the  $U$  output signal of the photodiode:

$$U = (\rho k / \pi) \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} \{E(x, y) \cdot h^2 / [x^2 + (b - y)^2 + h^2]^2\} dx dy \quad (10)$$

that is convenient for calculations using the Mathcad software package.

#### 4. Optimization of the design parameters of a fluorimeter using a mathematical model

Let us demonstrate the model optimization potential by solving the following problems: 1) using a wide or narrow beam to excite SAF, i.e. to select the window  $a$  size; 2) the influence of the form of a spatial distribution of the  $E(x, y)$  power density over an object per an output signal magnitude; 3) how sharply the output signal varies depending on a distance change between the  $2b$  photodiodes and 4) the  $h$  height of their location above an investigated object. Since the optimization refers to a particular design shown in Fig. 2 with the selected type of BPW21R photodiode whose case diameter [22] is 9 mm, we have  $b \geq 4$  mm. It is not feasible to reduce the  $h$  height to values less than 10 mm without complicating the optical system by using beam splitters, mirrors, etc. as well, i.e. we have  $h \geq 10$  mm. Clinically tested devices [8, 16] diagnose a skin area with a characteristic size of 6 - 10 mm. A diagnosed area of less than 1 mm in size can hardly be representative. Therefore, it is admissible to restrict the range of the  $a$  parameter variation by the segment [1; 10] mm.

The optimization goal is to obtain a set of  $a$ ,  $b$  and  $h$  geometric parameters with the design constraints discussed above that do not significantly reduce the  $U$  output signal, all other things being equal. Then the value of the constant factor before the  $(\rho k/\pi)$  integral (10) is taken equal to 1000 in order to obtain single-valued integers along the ordinate axis, the value of the  $P_u$  emission power of the SAF exciting source is assumed equal to 1, the  $U$  output signal is measured in nominal units and the normalization condition is used:

$$\int_{-a/2}^{a/2} \int_{-a/2}^{a/2} E(x, y) \cdot dx dy = 1. \quad (11)$$

The uniform distribution of power density is the simplest one with simulation results easily verified physically:  $E(x, y) = E_0 = P_u / a^2 = Const$  (12)

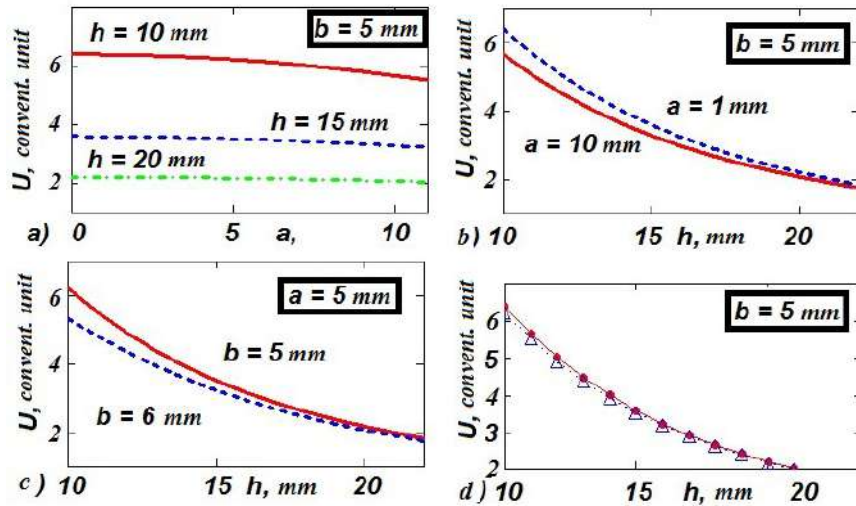


Fig. 4. Results of modeling (parameter in a frame has identical value for all curves of the schedule): a) - influence of the cross size of a bunch on the size of an output signal; b) - dependence of an output signal from photo diode arrangement height over a research object; c) - influence of distance of the photo diode from an optical axis of the probing bunch; d) - influence of the law of distribution of density of power of radiation of excitement of AFK on an object::  $\Delta\Delta\Delta$  - uniform distribution;  $\bullet\bullet\bullet$  - Gaussian distribution.

The simulation results are shown in Fig. 4. It follows from Fig. 4a that, in terms of the device energy, the formation of a small aperture beam does not provide any tangible advantage. The dependence on the height of the photodiode above the object of investigation appears to be more significant (Fig. 4b). It can be seen that this distance should be minimized; on the other hand, small  $\sim 1$  mm height variations caused, for example, by the need to replace a light filter with a light filter of a different thickness or errors in manufacturing optical element holders should not significantly affect the magnitude of an output signal. The last remark is related to the variation of the  $b$  design parameter (Fig. 4c).

Fig. 4d illustrates the effect of the law of the  $E(x, y)$  power density distribution over the object. In addition to the uniform distribution, a Gaussian distribution is introduced into the model, as it is characteristic of laser radiation emitters:

$$E(x, y) = M_0 \exp[-2(x^2 + y^2)/w^2], \quad (13)$$

where  $M_0$  is the power density on the Gaussian beam axis ;  $w$  is the radius (the distribution parameter) of the beam. The normalization to the full power of the SAF exciting source is carried out with the help of the expression:

$$M_0 = 2P_u / \pi \cdot w^2. \quad (14)$$

The comparison in Fig. 4d was carried out for the following values of the parameters:  $a = 5$  mm and  $w = 1.25$  mm. The choice of the value of a Gaussian beam radius was determined by the condition that the total radiation power of the emitter should almost completely fall on the object without masking it with an output window with a characteristic size of 5 mm. The condition should be taken into consideration since it is known [25] that only 86.5% of the total power passes through the cross section of a Gaussian beam of  $2w$  diameter and to increase the total power to 99.99%, the  $4w$  cross section diameter is required. As could be expected from physical considerations, according to the results of running the model with a uniform distribution of the power density of different cross sections (Figure 4a) the output signal is insensitive to the law of power density distribution, at least in the category of radially symmetric distributions. The  $(U_{\text{gaussian}} - U_{\text{uniform}})/U_{\text{gaussian}}$  relative difference of values does not exceed 2.5%, which confirms the correctness of the mathematical apparatus used.

## 5. Conclusion

A mathematical model has been developed that enables to predict the energy characteristics of a device by automating a calculating component of designing according to the manufacturer's specifications, literature or experimental data on the parameters of optoelectronic system components of the designed diagnostic fluorimeter and the optical properties of a diagnosed object.

Simulation showed that the most significant contribution to the device energy is made by the distance between the surface of a diagnosed object and the photosensitive pad of the photodetector. To obtain a maximum output signal of the photodetector, it is required to minimize the distance within the range of permissible design constraints. A small effect of the size of a diagnosed area on the output signal provides an additional degree of freedom for medical applications.

The model is implemented in the Mathcad software package to which mathematical models of the optical system components described above, the number of components and their parameters and expressions that the spectrum transfer over propagating radiation through an optical component obeys are introduced. The explicitly defined mathematical functions ensure the openness of the model and the feasibility of its modifying by a user.

## Acknowledgments

This research was supported by the Ministry of Education and Science of the Russian Federation and results have been received within performance of the state task of the Ministry of Education and Science of the Russian Federation on the project 15.6567.2017/BCh (hands. V.P. Zakharov) – 15.6567.2017/8.9, and RFBR r\_a (project 17-42-630907).

## References

- [1] Meerwaldt R, Graaff R, Oomen PHN et al. Simple non-invasive assessment of advanced glycation endproduct accumulation. *Diabetologia* 2004; 47: 1324–1330.
- [2] Mulder D J, van Haelst PL, Graaff R et al. Skin autofluorescence is elevated in acute myocardial infarction and is associated with the one-year incidence of major adverse cardiac events. *Netherlands Heart Journal* 2009; 17(4): 162–168.
- [3] Meerwaldt R, Hartog JWL, Graaff R et al. Skin Autofluorescence, a Measure of Cumulative Metabolic Stress and Advanced Glycation End Products, Predicts Mortality in Hemodialysis Patients. *Journal of the American Society of Nephrology* 2005; 16: 3687–3693.
- [4] Golubev RV, Papayan GV, Glazunova AA, Korosteleva NYu, Petrishchev NN, Smirnov AV. Examination of skin autofluorescence for the determination of glycation end-products in patients on chronic hemodialysis. *Therapeutic Archive* 2016; 88(6): 65–77.
- [5] Papayan GV, Petrishchev NN, Krylova EV et al. Method of estimating the biological age of skin by means of a fluorescence multispectral video dermatoscope. *Journal of Optical Technology* 2010; 77(2): 60–67.
- [6] Blyumin RB, Naumova EM, Khadartsev AA. The Technologies of Non-Contact Diagnostics. *Journal of New Medical Technologies* 2008. 15(4): 146–149.
- [7] Lutgers HL, Gerrits EG, Graaff R et al. Skin autofluorescence provides additional information to the UK Prospective Diabetes Study (UKPDS) risk score for the estimation of cardiovascular prognosis in type 2 diabetes mellitus. *Diabetologia* 2009; 52: 789–797.
- [8] Page reader brochure: [www.diagnoptics.com](http://www.diagnoptics.com) | [www.age-reader.com](http://www.age-reader.com).
- [9] Dunaev AV, Dremin VV, Zherebtsov EA et al. Analysis individual variability of parameters of laser fluorescence diagnostics. *Biotechnosphere* 2013; 2(26): 39–47.
- [10] Kang Uk, Papajan GV, Berezin VB et al. Spectrometer for fluorescence- reflective Biomedical Research. *Journal of Optical Technology* 2013; 80(1): 56–67. (in Russian)
- [11] Papajan GV, Gurba VM, Kishalov AA et al. Fiber - reflective fluorescent spectrometer with multiwavelength excitation. *Journal of Optical Technology* 2014; 81(1): 38–43. (in Russian)
- [12] Bulgakova NN, Smirnov VV, Fabelinsky VI et al. Laser spectral fluorescence colposcope: preclinical testing on experimental mice tumor. *Biomedical* 2013; 2: 108–122. (in Russian)
- [13] Novikov IA, Grusha YO, Kiryushchenkova NP. Improving Efficacy of Fluorescent Diagnostics of Skin and Mucosal Tumors in Ocular Oncology. *Annals of the Russian Academy of Medical Sciences* 2012; 10: 62–69. (in Russian)
- [14] Rogatkin DA, Sokolovski SG, Fedorova KA, Stewart NA, Sidorov VV, Rafailov EU. Basic principles of design and functioning of multifunctional laser diagnostic system for noninvasive medical spectrophotometry. *SPIE Proc* 2011; 7890: 1–7. DOI: 10.1117/12.874258.
- [15] Kornilin DV, Grishanov VN. Portable fluorescence meter for medical applications. *Proc. of SPIE* 2016; 9887: 1–7. DOI: 10.1117/12.2227392.
- [16] Kornilin DV, Grishanov VN, Zakharov VP, Burkov DS. Portable fluorescence meter with reference backscattering channel. *Proc. SPIE* 2016; 9961: 1–8. DOI:10.1117/12.2237135.
- [17] Koetsier M, Lutgers HL, Smit AJ, Links TP, de Vries R, Gans ROB, Rakhorst G, Graaff R. Skin autofluorescence for the risk assessment of chronic complications in diabetes: a broad excitation range is sufficient. *Opt. Express* 2009; 17: 509–519.
- [18] Egorova OV, Schtejn GI. Comparison of fluorescence microscope lighting systems based on LEDs (LED) and a mercury lamp (HBO). *Journal of Optical Technology* 2011; 78(1): 99–101. (in Russian)
- [19] Grishanov VN, Kornilin DV, Kulikov VS. Adjustment of the emission spectra of the ultraviolet light-emitting diodes to excite the fluorescence of biological objects. *Proc. of Actual problems of electronics and telecommunications: Russian Scientific and Technical Conference, Samara, 2015*; 150–152.
- [20] Jmerik VN, Mizerov AM, Shubina TV et al. Deep UV AlGaIn quantum wells heterostructures grown by sub-monolayer digital molecular beam epitaxy with plasma-activated nitrogen. *Physics and Technology Semiconductors* 2008; 42(12): 1452–1457.
- [21] Sld3233vf: [www.alldatasheet.com/datasheet-pdf/pdf/228445/ETC2/SLD3233VF.html](http://www.alldatasheet.com/datasheet-pdf/pdf/228445/ETC2/SLD3233VF.html).
- [22] Bpw21r: [www.vishay.com](http://www.vishay.com).
- [23] Barun VV et al. Light scattering by a rough surface of human skin. 1. The luminance factor of reflected light. *Quantum Electronics* 2013; 43(8): 768–776.
- [24] Yakushenkov YuG. Theory and calculation of optoelectronic devices. Moscow: Logos, 1999; 480 p.
- [25] Klimkov YuM. Applied laser optics. Moscow: Mashinostroenie, 1985; 128 p.

# Modeling and investigating the stability of a solution to the inverse problem of signal separation

V.A. Zasov<sup>1</sup>, Ye.N. Nikonorov<sup>1</sup>

<sup>1</sup>Samara State Transport University, 2B Svobody Street, 443066, Samara, Russia

---

## Abstract

This paper proposes a method for modeling and analyzing the stability of a solution to the inverse problem of extracting individual signals from an additive mixture of several signals that come to measurement points from various signal sources inaccessible for direct measurement. Stability analysis is accomplished by determining those intervals (singular intervals) for parameters of a signal formation model in which steady signal separation is achievable. We have developed algorithms to calculate singular intervals for different parameter variations of a signal formation model—absolute, relative, critical, and their combinations—that simulate various practically significant types of model parameter perturbations that affect the stability of the solution to this inverse problem. The paper also presents the results of computer modeling for the proposed algorithms.

*Keywords:* signal separation; inverse problem; solution stability; signal models; singular intervals; algorithm; modeling

---

## 1. Introduction

Signal separation is a solution to the problem of extracting individual signals from an additive mixture of several signals that come to measurement points from various signal sources inaccessible for direct measurement. That solution is needed in many practical fields such as monitoring and diagnosis of technical facilities (e.g., vibroacoustic diagnosis), communications, medical diagnosis, and speech processing. This is because in complicated facilities, measured signals present an additive mixture of signals received from many components, and in most practical applications the extraction of parameters that describe the state of specific components is impossible without signal separation.

In addition, signal separation enables further parallel processing of each extracted signal, thereby improving the efficiency of data-processing systems.

Problem of signal separation relates to the class of inverse problems, which may be ill-posed in the general case. From that, it follows that the solution may be unsteady because slight changes in the parameters of the mixing matrix  $\mathbf{H}$  of the signal formation model or in characteristics of source signals lead to impermissibly large changes in the solution: an unstable computation of source signals [1, 2]. For a stable solution, the parameters of the object described by the signal formation model must satisfy several prior restrictions [3]. For instance, the mixing matrix must be invertible; the polynomials describing the transfer functions of channels must not have common roots; the number of receivers must equal that of sources.

In practice, prior restrictions may be violated since object parameters may change because of the object's evolution in time, measurement error, fabrication inaccuracies, and other causes that often are unpredictable. Thus, changes in parameters mixing matrix and in characteristics of source signals may cause a steady solution to migrate toward an unstable one, unsuitable for practical use.

That is why it is important to investigate how deviations of the above source properties from those presumed a priori affect the solution's stability and to investigate deviations from requirements for channel characteristics. It is relevant that the stability of the solution to the problem of signal separation should be determined a priori by calculating whether the parameters of the signal formation model fall within the model's parameter intervals in which stability is achievable. Currently the methods described below are used to analyze the stability of the solution to the inverse problem of signal separation.

Given that the signal formation model in the frequency domain is described, for each frequency, by a system of linear algebraic equations [ ], the stability of signal separation is determined by the stability of the solution to equation systems, which, as is known (see, e.g., ref. [4]), is determined by the condition number  $cond(\mathbf{H})$ . Thus, by calculating the condition number, one can evaluate the stability of signal separation: The higher the condition number, the worse the stability.

As reference [5] shows, for the matrix of parameter intervals, the minimal norm reducing the initial matrix  $\mathbf{H}$  to the degenerated matrix  $det(\mathbf{H} + \Delta\mathbf{H}) = 0$  equals  $\|\Delta\mathbf{H}\|_2 = 1 / \|\mathbf{H}^{-1}\|_2 = \sigma_M$ . With the value (real number)  $\|\Delta\mathbf{H}\|_2$ , one can also evaluate the stability of the solution to the problem of signal separation.

Stability analysis methods based on using condition numbers  $cond(\mathbf{H})$  and matrix norm  $\|\Delta\mathbf{H}\|_2$  exhibit limited functionality. Indeed, values  $cond(\mathbf{H})$  and  $\|\Delta\mathbf{H}\|_2$  are integral estimates of stability, and they do not allow singular intervals  $\Delta H$  to be determined in singular-interval matrices  $\Delta\mathbf{H}$ , which is important for practical applications. It is evident that the knowledge, in matrices, of elements (singular intervals) close to zero makes it possible to determine elements of mixing matrices  $\mathbf{H}$  on which the stability of the problem of signal separation mainly depends. In other words, values  $cond(\mathbf{H})$  and  $\|\Delta\mathbf{H}\|_2$  do not take into account the structure of a perturbation (for the same condition numbers and matrix norms, there can be an infinite number of perturbation realizations). But in practice, perturbations can have a structure: Each matrix element can have its own perturbation that is unlike the others, and that perturbation can, in turn, be absolute, relative, or critical—or a combination of the three.

Reference [6] proposes a method for analyzing the stability of the solution to a system of linear algebraic equations, offering



broader functionality compared with the methods described above. The method can be used for analyzing and verifying the stability of the solution to the problem of signal separation. The algorithm that uses that method determines the direction of the worst parameter variations that cause instability (singularity). If, in a set parameter interval, the condition number for absolute or relative variations has increased significantly (e.g., exceeded a threshold), a solution within this interval is assumed unstable. Thus, the algorithm makes it possible to analyze the stability of a solution for signal formation models at a set variation value of mixing-matrix elements.

The method proposed in [10] does not solve the problem of determining the matrix  $\Delta\mathbf{H}$  of singular parameter intervals; nor does it enable use of complex mixing matrices  $\mathbf{H}(\omega)$ , and that limits the method's functionality.

Thus, our analysis of existing methods' functionality indicates the relevance of developing a method for modeling, analyzing, and verifying the stability of the solution to the problem of signal separation.

This paper proposes analyzing stability by determining singular intervals with the model's parameter variations directed toward the maximal deterioration of the solution's stability.

## 2. The object of the study

To state the problem formally, we will consider a signal formation model presented as a linear multivariable system that has  $N$  inputs and  $M$  outputs. The model's input signals are  $s_n(k)$ ,  $n=1,2,\dots,N$ ; output signals,  $x_m(k)$ ,  $m=1,2,\dots,M$ . The input signals are generated by various signal sources, and the output signals may be signals of various receivers such as sensors, measurement transducers, and antennas. Let us assume that each of the  $M$  outputs of the multivariable system is connected with all the  $N$  inputs through linear signal transmission channels.

At any discrete instant of time  $k$ , the  $M$ -dimensional vector of sensor-measured discrete signals  $\mathbf{x}(k)=[x_1(k),x_2(k),\dots,x_M(k)]^T$  results from the  $N$ -dimensional vector of source signals  $\mathbf{s}(k)=[s_1(k),s_2(k),\dots,s_N(k)]^T$ . The mathematical model of signal formation is described by an equation system of discrete convolution type (1), where the  $m$ th observed signal is an additive mixture of channel-distorted source signals and noise [7]; that is,

$$x_m(k) = \sum_{n=1}^N \sum_{g=0}^{G-1} h_{mn}(g, \mathbf{l}) s_n(k-g) + y_m(k), \quad (1)$$

where  $h_{mn}(g, \mathbf{l})$  is the element  $M \times N$  of the  $\mathbf{h}(g, \mathbf{l})$  matrix for the impulse characteristics of channels, and

$\mathbf{y}(k)=[y_1(k),y_2(k),\dots,y_M(k)]^T$  is the noise vector. For purposes of further discussion, we will assume that the  $h_{mn}(g, \mathbf{l})$  impulse characteristics are finite and are represented by the counting number  $G$ . The dynamic characteristics of channels  $h_{mn}(g, \mathbf{l})$  are quasistationary in that they change depending on parameter vector  $\mathbf{l}$  (time, temperature, location, etc.).

In the frequency domain, model (1) is described as

$$\mathbf{X}(\omega) = \mathbf{H}(\omega, \mathbf{l}) \cdot \mathbf{S}(\omega) + \mathbf{Y}(\omega), \quad (2)$$

where  $\mathbf{H}(\omega, \mathbf{l}) = \begin{pmatrix} H_{11}(\omega, \mathbf{l}) & H_{1N}(\omega, \mathbf{l}) \\ H_{M1}(\omega, \mathbf{l}) & H_{MN}(\omega, \mathbf{l}) \end{pmatrix}$  - is the mixing matrix  $M \times N$ , comprising Fourier transforms of the channels;

$\mathbf{X}(\omega)=[X_1(\omega), \dots, X_M(\omega)]^T$  is the vector of observed signals, and it comprises Fourier transforms of receiver signals;  $\mathbf{S}(\omega)=[S_1(\omega), \dots, S_N(\omega)]^T$  is the vector of source signals, and it comprises Fourier transforms of source signals;  $\mathbf{Y}(\omega)=[Y_1(\omega), \dots, Y_M(\omega)]^T$  is the noise vector, comprising Fourier transforms of noise signals. Signals of sources  $\mathbf{S}(\omega)$  and of noise  $\mathbf{Y}(\omega)$  are considered independent, and channels can be modeled by spectral converters such as various filters.

Generally, the solution to the problem of separating signal sources reduces to calculating the separating matrix  $\mathbf{w}(g)$ , which is, in terms of specific criteria, equal or close to the matrix inverse to matrix  $\mathbf{h}(g, \mathbf{l})$ . Thus, generally, the solution to the problem of separating source signals is the solution to system (1), and it can be expressed as

$$s_n(k) = \sum_{m=1}^M \sum_{g=0}^{G-1} w_{nm}(g, \mathbf{l}) x_m(k-g), \quad (3)$$

where  $\mathbf{w}(g, \mathbf{l})$  is the matrix of impulse characteristics of tunable filters with  $w_{nm}(g, \mathbf{l})$  elements. In the frequency domain, equation (3) can be written as

$$\mathbf{S}(\omega) = \mathbf{W}(\omega, \mathbf{l}) \mathbf{X}(\omega), \quad (4)$$

where  $\mathbf{W}(\omega, \mathbf{l}) = \mathbf{H}^{-1}(\omega, \mathbf{l})$ . It is evident that calculating the separating matrix  $\mathbf{w}(g, \mathbf{l})$  requires prior information about parameters of the signal formation model (object parameters). For separating signal sources, a variety of approaches are used that are based on various prior knowledge of the item under study. Signal separation methods can be classified into two groups—deterministic and statistical [1].



The deterministic group is based on prior information about characteristics of signal transmission channels—that is, on the knowledge of the matrix of impulse characteristics  $\mathbf{h}(g, \mathbf{l})$ , which are either measured or determined from theoretical premises.

A feature of the statistical group is that  $\mathbf{h}(g, \mathbf{l})$  matrix elements are unknown explicitly, and the information used to determine input signals  $\mathbf{s}(k)$  is provided by the realization of the vector of measured signals  $\mathbf{x}(k)$  and the knowledge of source properties of signals  $\mathbf{s}(k)$ .

The deterministic group is based on principal information about signal transmission channels (statistical, frequency, amplitude, and other channel characteristics); that is, transmission channels and signals are known.

The statistical group is based on principal information about signal sources such as lacking source correlation and the knowledge of signal distribution laws. In this case, explicit information about transmission channels is unavailable, and only observed signals are known. For that reason, the methods within this group are often called “blind” [8].

Thus, the solution to the problem of separating of signal sources reduces to using a deterministic or statistical method to calculate the separating matrix  $\mathbf{w}(g, \mathbf{l})$  equal or close, in terms of specific criteria, to the matrix inverse to matrix  $\mathbf{h}(g, \mathbf{l})$ .

There are separation methods that fall within neither group because they use information both about channels and the properties of signal sources (e.g., adaptive noise cancellation [9]).

From general solution (3) it follows that the problem of signal separation relates to the class of inverse problems, which may be ill-posed in the general case.

Our paper aims to:

—Develop an algorithm for modeling the problem of signal separation with a solution whose stability is variable by setting variations for the parameters of the signal formation model.

—Develop algorithms for analyzing and verifying stability by determining parameter intervals in which stable signal separation is achievable for various practically significant variations of model parameters.

This paper investigates the stability of the solution to the problem of signal separation with varied parameters of channels  $H_{mn}(\omega, \mathbf{l})$ , constituting the mixing matrix  $\mathbf{H}(\omega, \mathbf{l})$ .

### 3. Methods. Algorithms for Modeling, Analyzing, and Verifying the Stability of the Solution to the Problem of Signal Separation

#### 3.1. Mathematical Signal Formation Model with the Capability to Set Variations for Channel Parameters

To investigate how prior indefinite perturbations affect stability, we propose introducing singular-variation blocks for parameters of channels  $\delta h_{mn}$  into the signal formation model. Then the signal formation model with parameter variations shown in figure 1 will take the form of [7]

$$x_m(k) = \sum_{n=1}^N \sum_{g=0}^{G-1} (h_{mn}(g, \mathbf{l}) + \delta h_{mn}(g, \mathbf{l})) s_m(k-g) + y_m(k) \quad (5)$$

In the frequency domain, expression (5) can be written as

$$\mathbf{X}(\omega) = (\mathbf{H}(\omega, \mathbf{l}) + \delta \mathbf{H}(\omega, \mathbf{l})) \cdot \mathbf{S}(\omega) + \mathbf{Y}(\omega), \quad (6)$$

where  $\delta \mathbf{H}(\omega, \mathbf{l})$  is the matrix of singular parameter variations.

Unlike objectively existing perturbations, parameter variations in the model for stability studies can be modeled by introducing a block for setting types of variation. Thus, mathematical model (5) can be used to investigate how perturbations from various types of variation affect the stability of the solution to the problem of signal separation.

For purposes of further discussion, we will refer to matrices of parameter intervals varying from the initial state  $\mathbf{H}(\omega, \mathbf{l})$  to the degenerated (singular) state  $\mathbf{H}(\omega, \mathbf{l})$  as singular-interval matrices and designate them  $\Delta \mathbf{H}(\omega, \mathbf{l})$ .

Among the different types of variations, we will consider those most often encountered in engineering practice—absolute, relative, and critical variations, which simulate related real perturbations [7]. Critical variations are variations that cause the initial model (5, 6) to become degenerated at a minimal spectral variation norm of  $\|\delta \mathbf{H}(\omega, \mathbf{l})\|_2$ . Relative variations, as the name suggests, have values proportional to those of matrix elements. Absolute variations can be of any value unconnected with the value of the current matrix element. Thus, this paper aims to determine singular intervals for parameters  $\Delta \mathbf{H}_{abs}(\omega, \mathbf{l})$ ,  $\Delta \mathbf{h}_{abs}(g, \mathbf{l})$ ,  $\Delta \mathbf{H}_{rel}(\omega, \mathbf{l})$ ,  $\Delta \mathbf{h}_{rel}(g, \mathbf{l})$ ,  $\Delta \mathbf{H}_{crit}(\omega, \mathbf{l})$ , and  $\Delta \mathbf{h}_{crit}(g, \mathbf{l})$  for the variations above. In the introduced matrices of singular parameter intervals,  $mn$ th elements  $\Delta H_{mn}$  indicate changes in parameters of  $mn$ th elements of the initial matrix  $\mathbf{H}(\omega, \mathbf{l})$ .

In particular, if model channels are frequency-independent and parameter vector  $\mathbf{l}$  independent, the designations of singular-interval matrices omit arguments  $(\omega)$ ,  $(g)$  and  $(\mathbf{l})$ ; for instance,  $\Delta \mathbf{H}_{abs}$ ,  $\Delta \mathbf{h}_{abs}$ .

Thus, the singular intervals obtained from calculations reflect those absolute, relative, and critical perturbations of the signal formation model’s parameters that cause the model to be unstable.

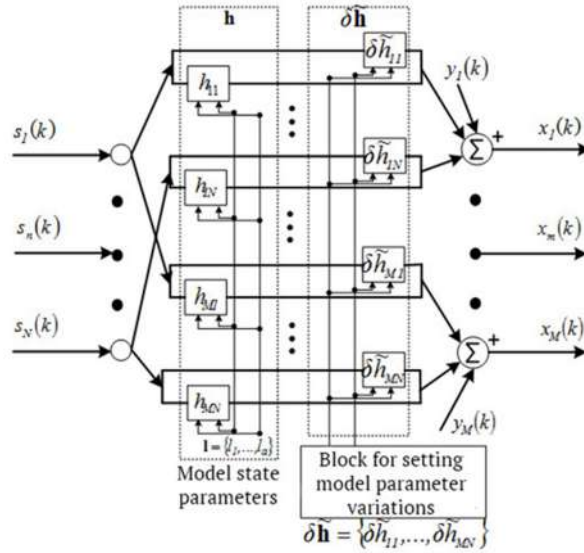


Fig. 1. Schematic of the signal formation model with the capability to set variations for channel parameters to analyze the stability of signal separation.

### 3.2. Analyzing and Verifying the Stability of Signal Separation by Determining Singular Parameter Intervals for the Signal Formation Model

For purposes of determining singular parameter intervals for different variations, a generalized algorithm has been developed [7, 10], in which three stages can be distinguished: determining the singular direction of parameter variation; determining a singular matrix; and determining a singular-interval matrix.

Singular directions for absolute, relative, and critical variations are determined by direction matrices whose analytic expressions, obtained in references [7, 10], are given in table 1.

Direction matrices can be calculated both on the basis of singular value decomposition (SVD) and on the basis of the inverse matrix. Determining singular directions with proposed matrices  $\mathbf{Z}$  has lower computational complexity compared with known algorithms.

Table 1. Analytic expressions for calculating direction matrices  $\mathbf{Z}_{crit}$ ,  $\mathbf{Z}_{abs}$ , and  $\mathbf{Z}_{rel}$ .

Calculation method	Type of variation		
	Critical	Absolute	Relative
Based on inverse matrix	$\frac{\mathbf{H}^{-*}}{\ \mathbf{H}^{-1}\ _2}$	$\frac{ \mathbf{A}  \otimes \mathbf{sign}(\mathbf{H}^{-*})}{\ \mathbf{A}  \otimes \mathbf{sign}(\mathbf{H}^{-*})\ _1}$	$\frac{ \mathbf{H}  \otimes \mathbf{sign}(\mathbf{H}^{-*})}{\ \mathbf{H}  \otimes \mathbf{sign}(\mathbf{H}^{-*})\ _1}$
Based on SVD	$-u_N v_N^*$	$\frac{ \mathbf{A}  \otimes \mathbf{sign}(u_N v_N^*)}{\ \mathbf{A}  \otimes \mathbf{sign}(u_N v_N^*)\ _2}$	$\frac{ \mathbf{H}  \otimes \mathbf{sign}(u_N v_N^*)}{\ \mathbf{H}  \otimes \mathbf{sign}(u_N v_N^*)\ _2}$

The proposed matrices  $\mathbf{Z}$  can be used to determine singular parameter intervals not only for real elements (as in known algorithms) but also for complex (frequency-dependent) elements of mixing matrix  $\mathbf{H}(\omega)$ .

Table 1 refers to:  $|\mathbf{A}|$ , a matrix composed of element modules of matrix  $\mathbf{A}$  and determining absolute parameter variations;  $\mathbf{sign}(\mathbf{A})$ , matrix operation whose elements are calculated as  $\mathbf{sign}(A_{mn}) = A_{mn} / |A_{mn}|$ ;  $\otimes$ , element-by-element multiplication of matrices  $\mathbf{C} = \mathbf{A} \otimes \mathbf{B}$ , where  $C_{mn} = A_{mn} \cdot B_{mn}$ ;  $v_n$  and  $u_n$ , right and left singular vectors of singular value decomposition  $\mathbf{H} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* = \sum_{n=1}^N \sigma_n u_n v_n^*$ ;  $\|\cdot\|_1$ , the maximum column sum matrix norm.

It is proposed that singular matrix  $\mathbf{H}$  be calculated by finding the roots of the equation  $f(\mathbf{H}_j + \delta h \cdot \mathbf{Z}) = \det(\mathbf{H}_j + \delta h \cdot \mathbf{Z}) = 0$  under restrictions caused by parameter variations.

The numerical algorithm (table 2) for determining singular matrices  $\mathbf{H}$  for absolute, relative, and critical variations is based on the Newton method, in which, unlike in the classical method, derivative  $f'_{z_j}(\mathbf{H}_j)$  is calculated on the basis of the matrix of directions  $\mathbf{Z}$  and refined in each step. This improves accuracy and simplifies computation compared with known algorithms [10].

At the third stage, singular-interval matrix  $\Delta \mathbf{H}$  is calculated as follows:  $\Delta \mathbf{H} = \mathbf{H} - \mathbf{H}$ . In the proposed algorithm, function  $f(\mathbf{H}_j + \delta h \cdot \mathbf{Z})$  must satisfy the conditions of convergence theorems from the Newton method, including the Lipschitz condition.

Table 2. Algorithm for determining singular matrix  $\mathbf{H}$  on the basis of the Newton method.

Step	Action	Note
1	Parameter $\gamma > 0$ is set, and it determines error, increment value $\delta h$ , and initial iteration value $j = 1$	Initialization takes place
2	$\mathbf{Z}_j$ is determined according to type of parameter variation (table 1) for matrix $\mathbf{H}_j$	Direction matrix is calculated
3	$f'_{\mathbf{Z}_j}(\mathbf{H}_j) = \frac{df(\mathbf{H}_j)}{d\mathbf{Z}_j} = \frac{f(\mathbf{H}_j + \delta h \cdot \mathbf{Z}_j) - f(\mathbf{H}_j)}{\delta h}$	Derivative is calculated for direction $\mathbf{Z}_j$
4	$\delta \mathbf{H}_j = -\frac{\mathbf{Z}_j \cdot f(\mathbf{H}_j)}{f'_{\mathbf{Z}_j}(\mathbf{H}_j)}$	Matrix for singular variations of parameters $\delta \mathbf{H}_j$ is calculated
5	$\mathbf{H}_{j+1} = \mathbf{H}_j + \delta \mathbf{H}_j$	New approximation is calculated
6	If $ f(\mathbf{H}_{j+1}) - f(\mathbf{H}_j)  < \gamma$ , then algorithm ends ( $\mathbf{H} = \mathbf{H}_j$ ); else, $j = j + 1$ , go to step 2	Algorithm completion is verified

The method for calculating singular intervals provided the basis for algorithms for verifying the stability of signal separation. Table 3 outlines an example algorithm [10].

Table 3. Algorithm for verifying the stability of the solution to the inverse problem of signal separation.

Step	Action	Note
1	Frequency index $g = 0$ is set for spectral matrix $\mathbf{H}(\omega_g, \mathbf{l})$	Initialization
2	Singular-interval matrix $\Delta \mathbf{H}(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) - \mathbf{H}(\omega_g, \mathbf{l})$ is calculated	For frequency $\omega_g$
3	Threshold matrix $\mathbf{H}_t(\omega_g, \mathbf{l})$ is determined	For frequency $\omega_g$
4	Matrices are determined for intervals of parameters of stable ( $\Delta \mathbf{H}_R(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) - \mathbf{H}_t(\omega_g, \mathbf{l})$ ) and unstable ( $\Delta \mathbf{H}_S(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) - \mathbf{H}_t(\omega_g, \mathbf{l})$ ) separation	For frequency $\omega_g$
5	If $ \Delta \mathbf{H}_{\max}(\omega_g, \mathbf{l})  \leq  \Delta \mathbf{H}_R(\omega_g, \mathbf{l}) $ , solution is stable; otherwise message appears stating that stable signal separation for frequency of $\omega_g$ is impossible	Conditions of separation stability at frequency of $\omega_g$ are verified
6	If $ \Delta \mathbf{H}_{\text{pot}}(\omega_g, \mathbf{l})  \leq  \Delta \mathbf{H}_R(\omega_g, \mathbf{l}) $ , solution is stable; otherwise message appears stating that stable signal separation for frequency of $\omega_g$ is not guaranteed	Conditions of stable separation at frequency of $\omega_g$ are verified
7	$g = g + 1$ . If $g > G - 1$ , algorithm ends, and final message on stability verification appears; else, step 2	Transition to next spectral matrix

The matrix  $\Delta \mathbf{H}_{\max}(\omega_g, \mathbf{l})$  added in step 5 for maximal allowable variation intervals is defined from theoretical and practical information about the object being modeled. Matrix inequalities of  $|\mathbf{A}| \leq |\mathbf{B}|$  type should be understood as systems of componentwise inequalities  $|A_{mn}| \leq |B_{mn}|$ .

Threshold matrix  $\mathbf{H}_t(\omega_g, \mathbf{l})$ , determined in step 3 of the algorithm, is mixing matrix  $\mathbf{H}_j(\omega_g, \mathbf{l})$ , for which  $\text{cond} \mathbf{H}_j(\omega_g, \mathbf{l})$  exceeds a given threshold value of  $\text{cond}_t$ . To determine matrix  $\mathbf{H}_t(\omega_g, \mathbf{l})$ , mixing matrix  $\mathbf{H}_j(\omega_g, \mathbf{l})$  is changed in accordance with the expression  $\mathbf{H}_j(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) + j \times \delta \mathbf{H}(\omega_g, \mathbf{l})$ , and in each step  $j = 1, \dots, J$  its  $\text{cond} \mathbf{H}_j(\omega_g, \mathbf{l})$  is compared with threshold value  $\text{cond}_t$ .

The calculated matrix  $\mathbf{H}_t(\omega_g, \mathbf{l})$  of threshold values serves as the basis for determining the matrix of parameter intervals for stable separation  $\Delta \mathbf{H}_R(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) - \mathbf{H}_t(\omega_g, \mathbf{l})$  and the matrix of parameter intervals for unstable separation  $\Delta \mathbf{H}_S(\omega_g, \mathbf{l}) = \mathbf{H}(\omega_g, \mathbf{l}) - \mathbf{H}_t(\omega_g, \mathbf{l})$ .

#### 4. Results and Discussion

Figure 2 shows the relationship between relative error  $\xi_{\Delta H}$  in determining singular intervals  $\Delta \mathbf{H}$  and the reduced error of the parameters of mixing matrix  $\mathbf{H}$  (determined by the number of binary digits of the ADC) and its values  $\text{cond}(\mathbf{H})$  for critical parameter variations. Figure 3 shows results testing algorithms for verifying the stability separation of signals.

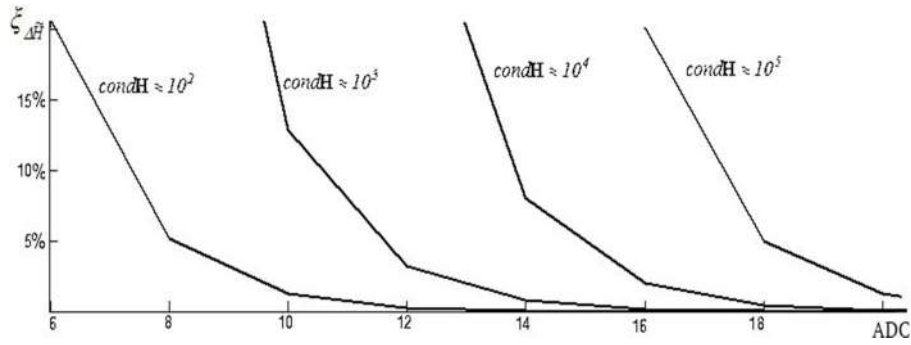


Fig. 2. The relationship between relative error  $\xi_{\Delta H}$  in determining singular intervals and the reduced error of matrix  $\mathbf{H}$  parameters (determined by the number of bits of the ADC) under critical parameter variations.

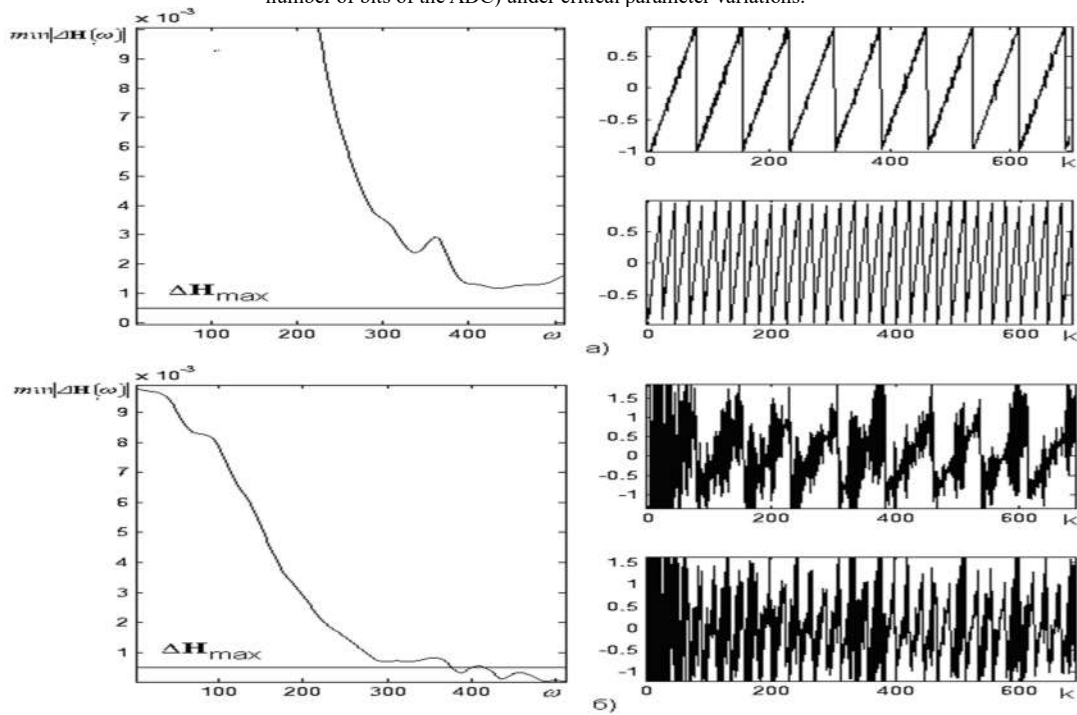


Fig. 3. Results testing algorithms for verifying the stability: illustrate stable (a) and unstable (b) separation of signals.

The proposed algorithms have been incorporated in a software system for modeling signal separation and restoration. Figure 4 shows examples of modeling and investigating the stability of signal separation for test signals and signals in an automatic cab signaling system, which transmits, via the track, traffic-light coding signals to train cab [11].

The automatic cab signaling system, which is used for train safety, operates under the influence of various interference sources. Suppressing that interference is important for the reliable operation of the system. Sometimes for interference sources to be suppressed, interference signals need to be extracted in order to determine their physical nature; that is, to identify interference sources [11]. This needs to be done as part of monitoring the condition of track circuits and automatic cab signaling systems.

Figures 4-a(1) and 4-a(2) show initial triangular test signals and their mixtures, and figures 4-a(3) и 4-a(4) are examples of an unstable and a stable solution to the problem of separating test signals. The Information window indicates that the separation is unstable: singular intervals at a frequency of 400 Hz are close to zero, and the condition number is on a sharp increase.

Figures 4-b(1) and 4-b(2) show examples of the automatic cab signaling system’s amplitude-modulated signals under the influence of interference: fluctuation noise from traction current, a 50 Hz harmonic interference from the power line, and a low-frequency interference of 4 Hz due to intake coils’ wobbling in relation to the track. Figures 4-b(3) and 4-b(4) illustrate unstable and stable separation of signals in the system and the above interference. Whether the solution is stable or unstable is displayed in the Information window.

This results confirms the possibility of using the developed algorithms in engineering applications.

### 5. Conclusion

The key results of our investigation are as follows:

We proposed an algorithm for modeling signal separation that makes it possible to study the stability of solutions to the problem of signal separation under stability-crucial parameter variations (perturbations) controlled in a signal formation model.

We also developed algorithms for analyzing and verifying stability by determining singular intervals for parameters of the signal formation model for various parameter variations.

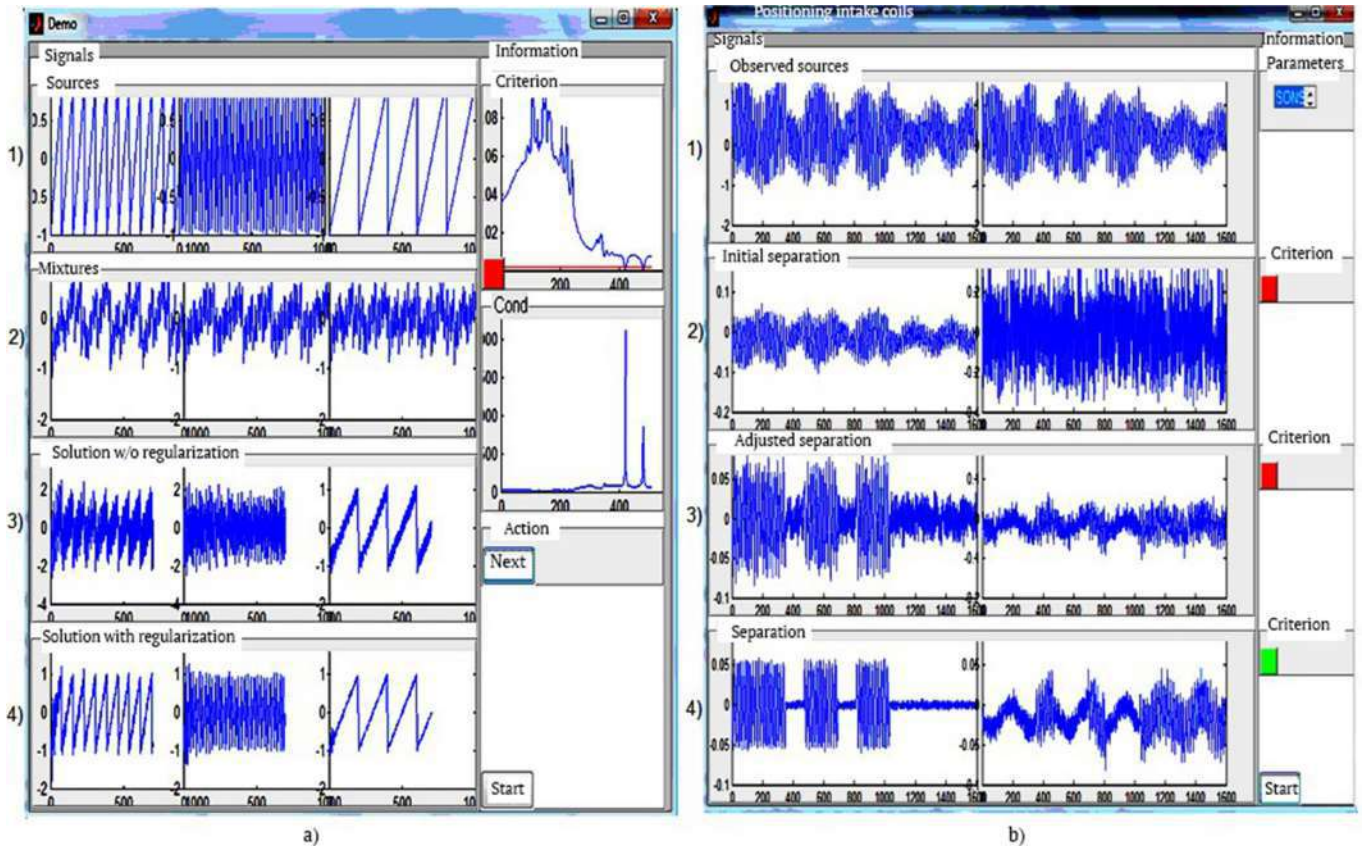


Fig.4. Results of analyzing the stability of the solution to the problem of signal separation: a) test signals; b) measured signals in the automatic cab signaling system.

## References

- [1] Bakushinskiy AB, Goncharov AV. Ill-Posed Problems: Numerical Methods and Applications. Moscow: Moscow State University Press, 1989; 199 p. (in Russian)
- [2] Petrov Yu.P, Sizikov VS. Well-Posed, Ill-Posed, and Intermediate Problems with Applications: A Textbook for Institutes of Higher Education. Saint Petersburg: "Politekhnik" Publisher, 2003; 261 p. (in Russian)
- [3] Digital Signal and Image Processing in Radiophysical Applications. Ed. Kravchenko VF. Moscow: "Fizmatlit" Publisher, 2007; 544 p. (in Russian)
- [4] Tyrtshnikov YeYe. Matrix Analysis and Linear Algebra. Moscow: "Fizmatlit" Publisher, 2007; 480 p. (in Russian)
- [5] Demmel J. Applied Numerical Linear Algebra. Theory and applications. Moscow: "Mir" Publisher, 2001; 430 p. (in Russian)
- [6] Petrov Yu P. Obtaining Reliable Solutions to Equation Systems. Saint Petersburg: Izdatelstvo "BKHV-Peterburg" Publisher, 2009; 176 p. (in Russian)
- [7] Zasov VA. Algorithms and Computational Devices for Separating and Restoring Signals in Multivariable Dynamic Systems: A Monograph. Samara: Samara State Transport University Press, 2013; 233 p. (in Russian)
- [8] Cichocki A, Amari Sh. Adaptive Blind Signal and Image Processing: Learning algorithms and applications. John Wiley & Sons Ltd, 2002; 555 p.
- [9] Windrow B, Stearns S. Adaptive Signal Processing. Moscow: "Radio i svyaz" Publisher, 1989; 440 p. (in Russian)
- [10] Zasov VA, Nikonorov YeN. Algorithms for Verifying the Stability of a Solution to the Problem of Separating Signal Sources under Conditions of Prior Uncertainty. Published in Hardware and Software Means for Management, Control, and Measurement Systems: Proceedings of a Conference with Russian and International Attendance. Moscow: Russian Academy of Sciences, Trapeznikov Institute of Control Sciences Press, 2010; 482–491. (in Russian)
- [11] Zasov VA, Nikonorov YeN, Tarabardin MA. Identifying Input Signals in Problems of Controlling and Diagnosing Dynamic Objects. Proceedings of the IV International Conference on Control Problems. Moscow: Russian Academy of Sciences, Trapeznikov Institute of Control Sciences Press, 2009; 1478–1486. (in Russian)



# Modeling and analysis of motion of a spacecraft with a tether aerodynamic stabilizer

D. Elenev<sup>1</sup>, Y. Zabolotnov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

Space tether system consists of two solid bodies connected by a tether. The deployment of this system is produced mainly by aerodynamic forces which act on the bodies. One of these bodies has higher ballistic coefficient thus acting as an aerodynamic stabilizer on low orbit. Such tether systems allow to lower the requirements for characteristics of a spacecraft and can be used for purposes of stabilization in higher layers of the atmosphere, for utilization of space debris. For mathematical modeling purposes the tether is represented as a set of mathematical points with elastic connections.

*Keywords:* tether system; stabilization; spacecraft; deployment; multipoint model of the tether

## 1. Introduction

The modeling of the motion of a spacecraft with a tether aerodynamic stabilizer is made on the atmosphere part of the motion during the deployment of the tether system. The tether system consists of a spacecraft, a stabilizer, and a tether. The deployment and orientation of the tether system are produced by aerodynamic forces. The stabilizer compared to spacecraft has a high ballistic coefficient  $\sigma = C_{xv}S/m$ , where  $C_{xv}$  is the drag coefficient,  $S$  is the characteristic square and  $m$  is mass.

This method of passive aerodynamic stabilization is described in [1], where some design solutions are describes. Aerodynamic stabilizer is a light inflatable or metal part and can be used for different purposes in the layers of the atmosphere: for aerodynamic stabilization of motion and to obtain the stability of motion of the spacecraft in lower layers of the atmosphere; for preliminary stabilization in the higher layers before the descent; for utilization of space debris by their descent in the high-density layers of the atmosphere. Aerodynamic stabilizer allows to lower the requirements for characteristics of a spacecraft which might have high mass-inertia and geometrical asymmetry. For the system, the stable motion can be obtained by the proper choice of parameters of the stabilizer.

The conditions for stability of the motion of the system consisting of two solid bodies connected by weightless non-stretchable tether are analyzed in [2], where it is assumed that the system is already deployed.

This research is focused on the deployment process. The mathematical model takes into consideration the extensibility of the tether and allows to evaluate the influence of a method of deployment.

The forms of a spacecraft and a stabilizer are close to spheres, but both of them can have high mass asymmetry. At the initial moment bodies are not separated and move on the circular orbit. The separation takes place with a relatively low relative velocity, and further motion of the bodies depends on the control method which is based on regulation of the release of the tether. The release mechanism works on deceleration only.

## 2. Mathematical model of the deployment

For mathematical modeling the geocentric coordinate system  $Oxyz$  is used. This coordinate system is connected to the plane of the orbit of the center of mass of the system and is defined at the moment of separation of the stabilizer from the spacecraft.  $Ox$  axis is directed to the ascending node of the orbit,  $Oy$  axis is parallel to the vector of velocity of the center of mass at the moment of separation. The spacecraft and the stabilizer are connected by the tether solid bodies (Fig. 1).

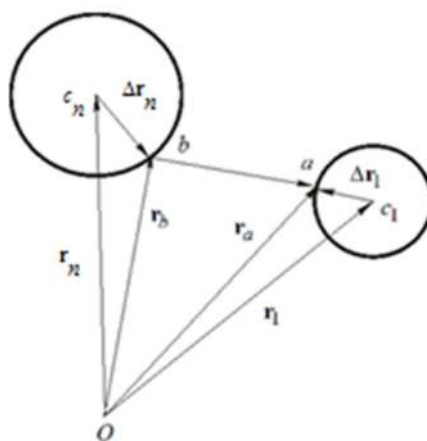


Fig. 1. The scheme of the system.

The equations of motions of this system are

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{G}_i + \mathbf{T}_i - \mathbf{T}_{i+1} + \mathbf{R}_i, \quad i = 1, 2, \dots, n \quad (1)$$

$$J_{x_i} \frac{d\omega_{x_i}}{dt} + \omega_{y_i} \omega_{z_i} (J_{z_i} - J_{y_i}) = M_{x_i}, \quad J_{y_i} \frac{d\omega_{y_i}}{dt} + \omega_{x_i} \omega_{z_i} (J_{x_i} - J_{z_i}) = M_{y_i}, \quad J_{z_i} \frac{d\omega_{z_i}}{dt} + \omega_{x_i} \omega_{y_i} (J_{y_i} - J_{x_i}) = M_{z_i}, \quad (2)$$

where indexes  $i=1$  and  $i=n$  are for center of masses of a spacecraft and stabilizer respectively,  $m_i$  are masses and  $\mathbf{r}_i$  are radius-vectors for bodies and material points on which the tether is divided;  $\mathbf{G}_i$  and  $\mathbf{R}_i$  are gravitational and aerodynamic forces,  $t$  is time,  $\mathbf{T}_i$  and  $\mathbf{T}_{i+1}$  are acting on adjacent areas of the tether tension forces;  $J_{x_i}, J_{y_i}, J_{z_i}$  are moments of inertia of bodies in coordinate systems  $c_i x_i y_i z_i$ ;  $\omega_{x_i}, \omega_{y_i}, \omega_{z_i}$  are projections of angular velocities;  $M_{x_i}, M_{y_i}, M_{z_i}$  are projections of acting on each body moments.

During the modeling, gravitational moments are neglected while aerodynamic moments and moments from tension force are taken into consideration. Tension forces are defined by Hooke's law

$$\mathbf{T}_i = T_i \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}$$

where  $T_i$  is the value of the tension force number  $i$ . If  $|\mathbf{r}_{i+1} - \mathbf{r}_i|$  is less or equal to non-deformed length of the tether on the area number  $i$ , than the tension force is equal to zero. The spacecraft and the stabilizer are influenced by the tension force from one area of the tether only. For calculation of these forces vectors  $\mathbf{r}_a$  и  $\mathbf{r}_b$  are used. If the tether is not strained, than the free motion of bodies and material point of the tether takes place.

To define gravitational forces, the central gravitational model under Newton's law is used. For the tether, aerodynamic forces are calculated as forces acting on the cylinder [3]. These forces are distributed proportionally between material points of the tether. It is assumed that the motion of the systems takes place in low density gas and the hypothesis of diffuse reflection of gas molecules can be used. [3].

The equations for dynamics of the tether release mechanism are

$$m_u \frac{d^2 L}{dt^2} = T_1 - F_u, \quad (3)$$

where the constant coefficient  $m_u$  depict the inertia of the tether release mechanism,  $F_u$  is the control force,  $T_1$  is the tension force on the first area of the tether, starting from the spacecraft. The tether release mechanism works on deceleration only,  $F_u > 0$ , and cannot pull the tether in.

The stabilizer is being separated from the spacecraft with relative velocity  $\mathbf{V}_r$ , and it is necessary to re-calculate velocities basing on the law of impulse saving

$$\mathbf{V}_1^{(a)} = \mathbf{V}_c^{(a)} - \frac{m_2}{m_1 + m_2} \mathbf{V}_r, \quad \mathbf{V}_2^{(a)} = \mathbf{V}_1^{(a)} + \mathbf{V}_r,$$

where  $\mathbf{V}_c^{(a)}$  is the absolute velocity of the center of mass of the system before the separation,  $\mathbf{V}_1^{(a)}$  and  $\mathbf{V}_2^{(a)}$  are absolute velocities after the separation.

### 3. The deployment of a tether system and the regulation of the tether release

During the deployment process dynamic or kinematic control laws can be used. For example, the following dynamic law is used with areas of acceleration and deceleration

$$F_p = \begin{cases} F_{\min}, & t < t_1 \\ F_{\min} + (F_{\max} - F_{\min}) \sin^2 [k_p (t - t_1)], & t_1 \leq t \leq t_2 \\ F_{\max}, & t > t_2 \end{cases} \quad (4)$$

where  $t_{1,2} = t_p \pm \pi / 4 k_p$ ,  $t_p, k_p, F_{\min}, F_{\max}$  are parameters of control law. The switching of control force is made on time bases, here  $t_p$  - is the time then the force switches. Parameter  $k_p > 0$  defines the smoothness of switching, the lower is it, the smoother is the switching. The parameters of law (4) are defined basing on edge conditions for the end of deployment:  $L_p(t_k)$ ,  $L_p(t_k) = L_p(t_k) = 0$ , where  $t_k$  is the time of the finishing the deployment.

The dynamic law (4) can be realized using the feedback principle:

$$F_u = F_p(t) + p_L [L - L_p(t)] + p_V [L - L_p(t)], \quad (5)$$

where  $L_p(t)$  and  $L_p(t)$  are program, or nominal, dependencies of length and rate of change of the length of the tether.;  $p_L, p_V$  are feedback coefficients;  $L, L$  are perturbed length and rate of change of the length which meet the conditions (3);  $F_p(t)$  is the program decelerating force.

The principle of regulating on bases of changing of the length and the rate of change of the length of the tether (5) was used in the real orbital tether experiment YES2 [4] and other researches [5,6].

To calculate the control force (5) it is necessary to define the dependencies  $F_p(t)$ ,  $L_p(t)$ ,  $L_p(t)$ , which can be found numerical by solving the system [5]. According to this, it is required to make prior calculations for these values and to use interpolation during the regulation process. But it is also possible to use more simple principle based on kinematic control law

$$L_p(\tau) = V_{\max} \cos^2(\tau + \nu), \quad (6)$$

where  $V_{\max}$ ,  $\omega$  and  $\nu$  are parameters. These parameters are defined from the system of non-linear equations

$$L_p(\tau_k) = 0, \quad \frac{dL_p}{d\tau}(\tau_k) = 0, \quad L_p(0) = V_r, \quad \int_0^{\tau_k} L_p(\tau) d\tau = L_k, \quad (7)$$

where  $\tau_k = \omega t_k$  is dimensionless duration of the deployment.

#### 4. Numerical results

The modeling of the deployment of the tether system was made using equations (1-3) and tether release laws (4) and (6). During the release of the tether, the algorithm for inserting the material point was used. This algorithm is described in [5]. On adding new point of the tether, the velocity of a spacecraft and a point of the tether are recalculated basing on the law of conservation of impulse of the system. The relative velocity of a new point is being calculated basing on the relative velocity of the previous point. Relative velocities for the points are defined relatively to the spacecraft. The next results are made for the following initial data: the masses of the spacecraft and the stabilizer are 200 kg and 20 kg, the final length of the tether is 0.5 km, initial altitude of a circular orbit is 250 km, the linear density of the tether is 0.2 kg/km, rigidity of the tether is 7000 N, initial relative velocity of separation is 2 m/s, feedback coefficients  $p_L = 0.2$ ,  $p_V = 7.8$ . The number of material points for modeling the tether is eight. The task of finding the optimal feedback coefficients was not taken into consideration, and the values of these coefficients were chosen on the assumption of obtaining the stability of regulation processes under the initial perturbations on initial velocity of separation (25%) and the direction of separation ( $\pm 1 \text{ rad}$ ).

The analysis of numerical results shows that kinematic deployment law (6) has great advantages compared to dynamic law (4). The advantages are based on smoother deceleration of the tether while conditions (7) are used. Usage of dynamic law (4) enforces to solve complex boundary-value problem using numerical calculation for the system of differential equations.

Figure 2 depicts the nominal dependence of rate of change of the tether length  $L_p(t)$  for this example. Figure 3 shows how the system reacts on the error in velocity of separation equal to 0.5 m/s. Figure 4 illustrates the dependence of the angle between longitudinal axis of the spacecraft and the direction of the tether  $\psi(t)$  from time. This dependence has high importance because it is a condition for exception of sagging and entanglement of the tether.

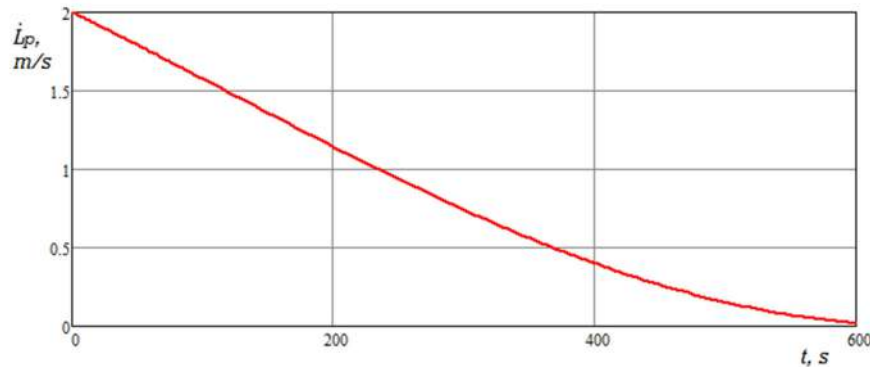


Fig. 2. Nominal dependence of rate of change of the tether length.

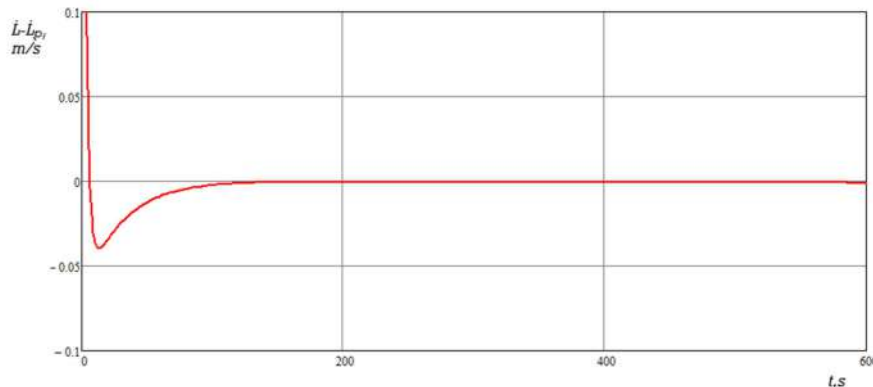


Fig. 3. Reaction on the error in velocity of separation.



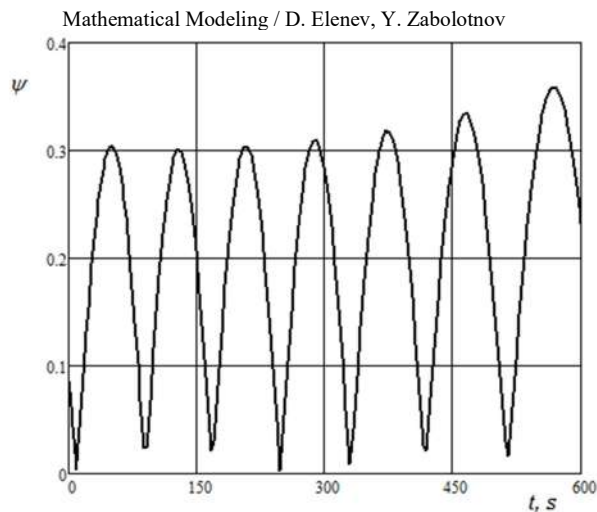


Fig. 4. The dependence of the angle between longitudinal axis of a spacecraft and the tether.

## 5. Conclusion

The analysis of different methods of deployment depicts that dynamics of motion of the system is mainly affected by the moment from tether tension force and almost is not influenced by the static stability of bodies, which is calculated as length between the center of mass and the center of pressure from aerodynamic forces. It is necessary to pay attention to the fact that this research was made for a system with a light and relatively short 0.5 km tether. The usage of a longer tether leads to increase of an aerodynamic pressure and to the necessity of dividing the tether into larger number of parts during calculations. This means that more computer resources are needed for calculations. Because of this the methods of high performance parallel calculations for analysis of deployment and optimization of parameters of the tether systems are being researched now.

This research is supported by the grant of Russian Foundation for Basic Research (RFBR) 16-41-630637.

## References

- [1] Alekseev KB, Bebenin GG. Spacecraft control. Moscow: Mashinostroenie, 1974; 343 p.
- [2] Zabolotnov YuM, Elenev DV. Stability of motion of two rigid bodies connected by a cable in the atmosphere. *Mechanics of solids* 2013; 48(2): 156–164. DOI 10.3103/S0025654413020064.
- [3] Arzhanikov NS, Sadekova. GS. Aircraft aerodynamics. Moscow: High School, 1983; 360 p.
- [4] Kruijff M. Tethers in Space. Netherlands: Delta-Utec Space Research, 2011; 423 p.
- [5] Zabolotnov YuM. Control of the deployment of a tethered orbital system with a small load into a vertical position. *J. of Applied Mathematics and Mechanics* 2015; 79(1): 28–34.
- [6] Williams P, Hyslop A, Kruijff M. Deployment control for the YES2 Tether-assisted Re-entry Mission. *Advance in the Astronautical Sciences* 2006; 123(2): 1101–1120.

# Development of algorithms for diagnosing forms of lichen planus and predicting of the disease's course

O.V. Serikova<sup>1</sup>, V.N. Kalaev<sup>2</sup>, N.A. Soboleva<sup>1</sup>

<sup>1</sup>Voronezh State Medical University, Studencheskaya, 10, 394000, Voronezh, Russia

<sup>2</sup>Voronezh State University, Universitetskaya PL. 1, 394000, Voronezh, Russia

---

## Abstract

The work is devoted to application of mathematical methods and development of algorithms for diagnostics of the lichen planus forms that located at mucous membrane of the mouth and lips, the differential diagnosis of severe forms of other diseases and its prediction in difficult cases through the use of expert knowledge base, the results of the selection of the most informative features and micronucleus test's interpretation in buccal epithelium.

*Keywords:* lichen planus; differential diagnosis; Kohonen neural network

---

## 1. Introduction

Diagnosis of lichen planus that located at the oral mucosa and the vermilion border has a significant interest among dentists, dermatologists, oncologists and doctors of other specialties. This is due to the lack of clear mechanisms of the disease development, severe, often permanent period, the existing trend towards malignancy elements of destruction, as well as frequent interaction with the General condition of the patient.

Treatment of patients applying for dental care with pathology of the oral mucosa and lips, is one of the most challenging problems in dentistry because of the difficulties that occur by diagnosis of diseases in this body's region. For example, the survey of 214 dentists, students of the improvement series in the Stomatology Department of the supplementary professional education Institute, Voronezh state medical University named after N. N. Burdenko, showed that only 30% of them are trying to make a diagnosis and prescribe treatment in cases of pathology of the oral mucosa and lips, and the remaining 70% of physicians send patients to the relevant medical Department of the University.

Analysis 568 advisory directions to the Department from dentists of the city's hospitals and region shows that the most often reveals of the discrepancy in the diagnosis of diseases such as lichen planus, various forms of cheilitis, erosive-ulcerative lesions of the oral mucosa, glossalgia. During the complex examination of patients by the Department staff revealed that the differences in the diagnoses at referral and the final diagnosis was 28%. In this regard, to improve the treatment of lichen planus of the oral mucosa and the vermilion border relevant has been the development and implementation of the educational and clinical practice of the Department by the algorithm computer system which allows on the basis of the most essential signs to diagnose forms of the disease, differential diagnosis with other diseases, to give the treatment regimen.

## 2. Methods

Due to the fact that the experience of our clinic covers almost a quarter of the observations' century, we can state the fact that the incidence of misdiagnosis has no tendency to decrease for years. This is despite the fact that in the literature, including and intended for dentists issues of clinical laboratory diagnostics of the diseases is paid more attention from year to year.

The authors analyzed clinical and laboratory characteristics of lichen planus, established expert knowledge base, the selection of the most informative parameters (erosive and ulcerative elements, the presence of local irritating factors, the presence of inflammatory infiltrate at the base of erosions, ulcers, nature of complaints, the course of the disease, age and gender of the patients, favourite localization, lesions of the skin, the reaction of the lymph nodes, the possibility of malignancy, specific characteristics, General condition), on the basis of which the basic principles of differential diagnosis of lichen planus with other diseases of the oral mucosa and lips was put forward.

The most common form was in the sample of 212 patients with the typical one (45,7%). Exudative-hyperemic form of lichen planus was diagnosed in 19,3%, erosive – 24,5%, bullous form – at 3.8%, erosive-ulcerative – 6.6% of patients. Significant difficulties in the diagnosis of lichen planus associated with the definition of severe forms of the disease. In determining of the criteria for identifying severe forms of lichen planus took into account the following: prevalence of the process on the mucous membrane of the mouth and lips, duration, and frequency of exacerbations, length of remissions, the severity of subjective sensations (pain), the effect of previous standard therapies, changes in the quality of life of patients.

The algorithm provides two basic modes of operation:

- 1) using the results of the micronucleus test in buccal epithelium, which informs the patient of lichen planus, and further work on determining the shape of disease and prediction of flow depending on the shape;
- 2) differential diagnosis of lichen planus with diseases such as ulcer decubitalis ulcers with signs of hyperkeratosis, trophic ulcers with signs of hyperkeratosis, leukoplakia and erosive forms of chronic lupus erythematosus, pemphigus vulgaris,

ulcerative-necrotic stomatitis, exudative erythema multiforme, recurrent aphthous stomatitis, and then in identifying of lichen planus implementation of the transition to the first mode.

In addition, in the reference part of the algorithm provides a practical introduction to the dentist with a full description of the knowledge base in the field of diagnostics of lichen planus forms and differential diagnosis with appropriate illustrations from the author's own clinical observations.

Micronucleus test in buccal the epithelium of the oral cavity is one of the most widely used methods for the evaluation of genetic homeostasis, because of its quickness, easiness, non-traumatic, cost-effective, allowing you to survey an unlimited number of times, requires no special equipment for cultivation of cells [1]. Practical introduction dentists can significantly improve the diagnostic level of lichen planus.

It should be noted that currently, often for the differential diagnosis of severe diseases and classification of their forms use neural networks, which are attractive from an intuitive point of view, because they are based on the primitive biological model of nervous systems.

The user of a neural network collates representative data and then runs the learning algorithm that automatically adapts to the data structure. However, the user requires a set of heuristic knowledge about how to select and prepare data, the desired network architecture and to interpret the results.

Currently, there are many successful examples of the application of neural network approach to building intelligent information systems [2, 4].

The ultimate goal of our research is to create neural network systems that allow for the diagnosis and differential diagnosis of lichen planus.

As it generally known, the creation of a neural network system includes the following stages: studying of the problem; problem statement; setting of training data and testing examples; training the neural network; optimal scheme; more experiments; the development and creation of the interface; the connection to the trained neural networks; system test examples not included in training data; a finish system in these examples [3].

As the neural structure was chosen as the Kohonen network, as it is, everyone carries out the classification. The Kohonen network can recognize clusters in data, and to establish intimacy and classes. Thus, the user can improve their understanding of data structures, and then to Refine the neural network model. If these recognized classes, they can be described, after which the network can solve classification problems. The Kohonen network can be used in the classification tasks where the classes are already set, then the advantage is that the network can identify similarities between different classes. Another possible area of application is the detection of new phenomena. The Kohonen network can recognize clusters in the training data and classifies all the data to the different clusters. If the network will meet with the dataset, unlike any of the known samples, it will not be able to classify such a set, and thus reveal its novelty. The network is trained by the Kohonen method of successive approximations. Starting from randomly chosen starting location of the centers, the algorithm progressively improves it in order to capture the clustering of the training data.

The principle of construction of system for differential diagnosis of lichen planus is as follows. Based on the table of differential diagnostics developed by the authors, was composed of simple questions, the answers to which are binary, i.e. "Yes" or "No". In drawing up a "vector of the survey", if the answer should be "Yes", then the vector component is assigned 1 if "No", then 0. This input vectors. Similar is the vector of output values, its components have a binary form.

The most informative characteristics that allow for differential diagnosis of exudative-hyperemic forms of lichen planus are presented in table 1.

Table 1. A list of the most informative signs for differential diagnosis of lichen planus.

Signs of disease	Code sign of the disease
Spontaneous pain	P1
The presence of papules of the oral mucosa	P2
The presence of local irritating factors	P3
The presence of inflammatory infiltration	P4
Localization on the surface of the tongue	P5
Localization on the red border of the lips	P6
Localization in the retromolar region	P7
The patient is a woman over 40	P8
Skin lesions	P9
Reaction of the lymph nodes	P10
The possibility of malignancy	P11
Specific features	P12
A burning sensation in the mouth	P13
The disease is a chronic	P14

The list of diseases for the differential diagnosis of exudative-hyperemic forms of lichen planus are presented in table 2.

The input vectors are eight diseases for the differential diagnosis of exudative-hyperemic forms of lichen planus are given in table 3.

Mathematical Modeling / O.V. Serikova, V.N. Kalaev, N.A. Soboleva  
 Table 2. The list of diseases for the differential diagnosis of exudative-hyperemic forms of lichen planus.

Disease	Code of disease
Lichen planus, exudative-hyperemic form	X1
Chronic mechanical trauma	X2
Papular syphilis	X3
Leukoplakia, flat shape,	X4
Lupus erythematosus chronic	X5
Allergic stomatitis	X6
Acute hyperplastic candidiasis	X7
Early signs of recurrent aphthous stomatitis	X8

Table 3. A list of the most informative signs for differential diagnosis of lichen planus.

Code sign of the disease	Disease							
	X1	X2	X3	X4	X5	X6	X7	X8
P1	0	0	0	0	0	0	0	1
P2	1	0	1	0	0	0	0	0
P3	0	1	0	1	0	0	0	0
P4	0	1	0	0	1	1	0	1
P5	1	1	1	1	1	1	1	1
P6	1	1	1	1	1	1	1	0
P7	1	1	1	0	1	1	1	0
P8	1	1	1	0	1	1	1	1
P9	1	0	1	0	1	0	0	0
P10	0	0	1	0	0	0	0	0
P11	0	1	0	0	1	0	0	0
P12	1	0	1	0	1	0	1	0
P13	1	0	0	0	1	1	1	0
P14	1	0	1	1	1	0	0	1

*The learning algorithm of Kohonen network*

The Kohonen network consists of one neurons' layer. The number of inputs of each neuron is  $n$  –and it is the total number of possible symptoms. The number of neurons  $m$  is the desired number of classes that need to divide (number of diseases). The significance of each of the inputs into a neuron is characterized by a numeric value called weight.

*Training*

**Step 1:** Initialize the network.

The weighted coefficients of the network  $w_{ij}, i = \overline{1, n}, j = \overline{1, m}$  are assigned small random values.

Values are defined  $\alpha_0$  - initial rate of training and

$D_0$  - the maximum distance between the weight vectors (columns of the matrix  $W$ ).

**Step 2.** Presentation the network a new input signal  $X$ .

**Step 3.** Calculate the distance from input  $X$  to all neurons of the network:

$$d_j = \sum_{i=1}^n (x_i - w_{ij}^N)^2, j = \overline{1, m}$$

**Step 4.** The choice of a neuron  $k, 1 \leq k \leq m$  with the shortest distance  $d_k$ .

**Step 5.** Adjusting the weights of a neuron  $k$  and all neurons that are at a distance of no more than

$$w_{ij}^{N+1} = w_{ij}^N + \alpha_N (x_i - w_{ij}^N).$$

**Step 6.** The decrease in the values of  $\alpha_N, D_N$ .

**Step 7.** Steps 2-6 are repeated until the weight stops changing (or until the total change of all weights will be very small).

After training, the classification is performed by supplying to the input network of the test vector, compute the distance from each neuron and then the neuron with the smallest distance as the indicator of correct classification.

### 3. Results and Discussion

For training the neural network were taken 180 cases, whose data were taken from the medical records of patients with already confirmed diagnosis. The data of 185 patients treated in the clinic were left to test the system. Table 4 shows examples of correct recognition diagnosis of a number of diseases as a result of the program.

Table 4. The distribution of patients in accordance with the forms of diseases of the oral mucosa and the results of testing

Nosological form of the disease	Number of cases	Number of correctly recognized cases
Ulcerative necrotic stomatitis	78	78 (100%)
Recurrent aphthous stomatitis	101	101 (100%)
Multiforme exudative erythema	33	31 (93,9%)
Lichen planus, erosive-ulcerative form	75	73 (97,3%)
Lichen planus exudative-hyperemic forms	82	78 (96,3)
Leukoplakia erosive forms	37	37 (100%)

For nanosistemy presented certain difficulties, for example, the differential diagnosis between lichen planus and erythema multiforme exudative (2 errors) that occurs frequently in clinical practice when doctors make mistakes that occurs in the 35% of cases.

Comparing the encountered incorrect ("guides diagnoses") of medical institutions in patients with diseases of the oral mucosa and the vermilion border, we can say that adequate "smart diagnosis" is recorded only in 72% of cases. In the remaining patients the diagnosis was incorrect. At the same time, the application developed by authors' algorithm of the neural network allows to obtain a correct diagnosis in 94-97%, which certainly contributes to improve early diagnosis of severe dental diseases.

### 4. Conclusion

Thus, the developed and implemented algorithm allows effective diagnostics of the forms of lichen planus and its differential diagnosis with other diseases. The system provides the possibility of reducing the amount of input data identifying the most significant indicators. The system is versatile and can be practical used by doctors to diagnose any other diseases by creating appropriate tests.

Program is performing differential diagnosis of lichen planus with the help of Kohonen network that is implemented in the programming system Delphi.

### References

- [1] Kalaev VN, Nechaev MS, Kalaeva EA. Micronucleus test of buccal epithelium of the oral cavity of the person. Voronezh: Publishing house Voronezh State University, 2016; 136 p.
- [2] Lvovich YaE, Kashirina IL, Shostak AA. Neural network approach to the selection of the most informative signs for the functional diagnostics of liquid rocket engines. Bulletin of the Voronezh State Technical University 2012; 8(8): 21–23.
- [3] Haykin S. Neural Networks: A Comprehensive Foundation Second Edition. M.: Publishing house "Williams", 2006; 1104 p.
- [4] Yelkova NL, Dubskaya EN, Kashirina IL, Soboleva NA. Using the Kohonen network for differential diagnosis of syndrome lesions of the mucous membrane of the mouth and skin. System analysis and management in biomedical systems 2006; 3(1): 52–60.

# Calculation of the electrostatic field distribution formed by the generator of the off-electrode plasma

M.A. Markushin<sup>1</sup>, V.A. Kolpakov<sup>1</sup>, S.V. Krichevskiy<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

A calculation of the electrostatic field distribution in the electrode system of a high-voltage gas-discharge device is made. The application of the conformal mapping method in order to obtain an analytical description of the of equipotentials and field lines distribution is described. The figures of the electrostatic field distribution are calculated, which made it possible to determine their relationship with the cathode-anode distance, the voltage at the electrodes and the hole diameter in the anode of the gas discharge device. The electrostatic field distribution of the device forming the off-electrode plasma is analyzed.

Keywords: plasma; high-voltage gas discharge; equipotentials; field lines; conformal mapping; the Schwarz-Christoffel integral

## 1. Introduction

The off-electrode gas-discharge plasma formed by a high-voltage gas discharge is used for plasmachemical etching of quartz, for preparing Ohmic contacts of semiconductor elements, cleaning the surfaces of semiconductor and dielectric substrates, contacts of small-size relays [1-4]. The high uniformity of the charged particles flow in the region of the gas discharge cross section and the independence of the discharge parameters from the dimensions of the treated area [5, 6] causes the wide propagation of this discharge. These advantages are a consequence of the singularity of the lines of the force and equipotentials distribution of the off-electrode plasma generator [7-9]. In turn, of the electrostatic field distribution depending on the design parameters determines the physics of the charged particles interaction processes with atoms and molecules of the residual gas. Existing publications contain no information on the relationship between such parameters and the electrostatic field distribution. Due with the laboriousness of the experimental study of this problem the computational model is proposed for the field distribution in the electrode system of a gas-discharge device.

## 2. Calculation of electrostatic field distribution of the off-electrode plasma generator

A high-voltage gas discharge is formed only in the area of the anode hole [8]. Outside this area, the electrode system design is a flat capacitor with uniform field distribution. Therefore, the design of a gas discharge device can be modeled by an electrode system in which the cathode and anode area outside of the electric field inhomogeneity are removed to infinity (fig. 1).

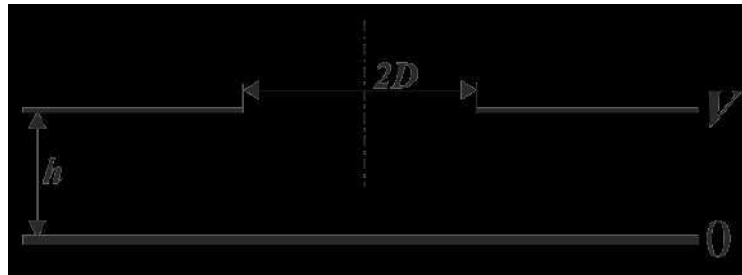


Fig.1. Schematic diagram of a device forming a high-voltage gas discharge:  $h$  is the anode–cathode distance,  $D$  is the radius of the anode orifice,  $V$  is the anode potential,  $0$  is the cathode potential.

Obtaining an analytical description of the distribution of the electrostatic field in the hole area in the anode is hampered because of field unevenness. To simplify such a problem, it is necessary to reduce it to the solution of the two-dimensional task, which will make it possible to simplify considerably the calculation of field lines and equipotentials by finding the complex potential for the canonical domain with a simple form of boundaries [10, 11]. Since the thickness of the anode has a value in the range of up to 0.5 mm, its influence on the formation of the electrostatic field is insignificant, because this thickness is smaller than the cathode-anode distance ( $h$  to 10 mm). Therefore, this quantity can be neglected.

The symmetry principle of the conformal mapping method allows us to consider only the right part of the obtained electrode system for the solution of the posed task by realizing the projection of the electrodes on the complex plane  $Z$ . This projection is shown in Fig. 2 (polygon)  $A_1A_2A_3A_4$ .

Simulation of the electrostatic field begins with finding the conformal mapping  $Z = f(\omega)$  of the upper half-plane  $\text{Im}\omega > 0$  to the region of the  $Z$  field with the electrodes  $A_1A_2$  (cathode),  $A_3A_4$  (anode) (fig. 2) with internal angles  $\alpha_k\pi$  at the vertices. Then an additional mapping  $\xi = f(\omega)$  of the half-plane  $\omega$  onto a strip  $0 < \text{Im}\xi < V$  with internal angles  $\beta_k\pi$  at the vertices (fig. 3).

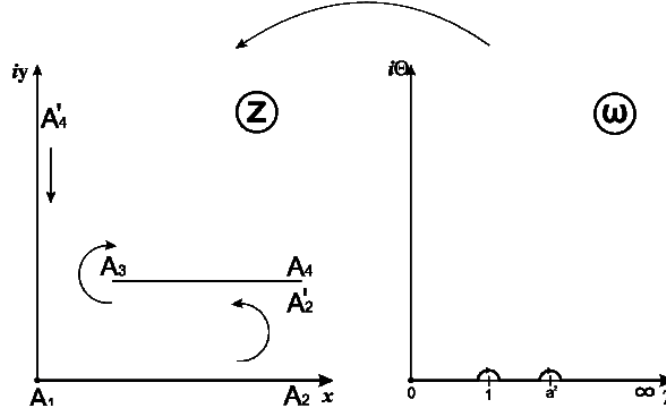


Fig.2. Diagram of the half-plane mapping onto plane (the electrode system).

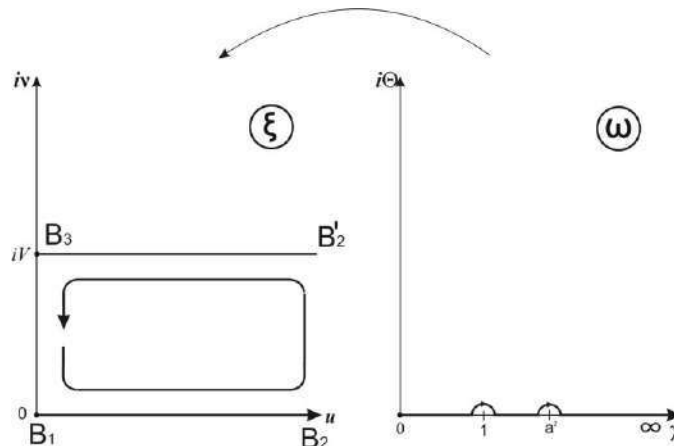


Fig.3. Diagram of the additional mapping of the half-plane  $\text{Im } w > 0$  onto the strip  $0 < \text{Im } \xi < V$ .

At the first stage, the vertices  $A_1 A_2 A_3 A_4$  of the  $Z$ -plane are associated with certain points of the real axis of the plane  $\omega$ . According to the theorem of uniqueness of a conformal mapping for a present correspondence of three arbitrarily chosen boundary points, for example,  $0, 1, \infty$ , we can obtain the correspondence [11]:

$$\begin{matrix} A_1 & A_2 & A_3 & A_4 \\ 0 & 1 & a^2 & \infty \end{matrix}$$

According to the technique developed in [10-12], the angles  $\mu_k$  are determined, which complement the internal angles  $\alpha_k$  at the vertices of the quadrangle  $A_1 A_2 A_3 A_4$  to  $\pi$ . Considering the inner region of a quadrilateral and moving in the positive direction of traversing its boundary, i.e. counterclockwise, we find the angles:  $\mu_1 = 1/2$  ( $\alpha_1 = 1 - \mu_1 = 1/2$ );  $\mu_2 = 1$  ( $\alpha_2 = 1 - \mu_2 = 0$ );  $\mu_3 = -1$  ( $\alpha_3 = 1 - \mu_3 = 2$ );  $\mu_4 = 3/2$  ( $\alpha_4 = 1 - \mu_4 = -1/2$ ).

To find the mapping function of a domain bounded by a polygon  $A_1 A_2 A_3 A_4$  the Schwarz-Christoffel integral [11] is used:

$$Z = C \int_{\omega_0}^{\omega_1} (\omega - a_1)^{\alpha_1 - 1} (\omega - a_2)^{\alpha_2 - 1} \dots (\omega - a_n)^{\alpha_n - 1} d\omega + C_1, \quad (1)$$

In the expression (1) instead of  $a_1 - a_n$  we substitute the corresponding points  $0, 1, a^2, \infty$ . According to [10], the factor related to the vertex  $a_4$  in the Schwarz-Christoffel integral is omitted, since  $a_4 = \infty$ .

In this case, the expression (1) has the form:

$$Z = C \int_0^{\omega} \omega^{-1/2} (\omega - 1)^{-1} (\omega - a^2) d\omega + C_1 = C \int_0^{\omega} \frac{(\omega - a^2)}{(\omega - 1)\sqrt{\omega}} d\omega + C_1$$

Let  $\omega = x^2$ , then:

$$Z = C \int_0^{\sqrt{\omega}} \frac{(x^2 - a^2)}{(x^2 - 1)x} dx^2 + C_1 = 2C\sqrt{\omega} + C(a^2 - 1) \ln \frac{1 + \sqrt{\omega}}{1 - \sqrt{\omega}} + C_1. \quad (2)$$

The value of the constant coefficient  $C_1$  is determined from the correspondence of the points  $A_1 \leftrightarrow 0$ , which allows us to write the equation:

$$Z = 2C \cdot 0 + C(a^2 - 1) \ln \frac{1 + \omega}{1 - \omega} + C_1 = C_1 = 0.$$

The transition from the lower electrode to the upper one, corresponding to the transition of the ray  $A_1A_2$  to the ray  $A_2A_3$  (fig. 2), allows us to determine the constants  $a^2$  and  $C$ . As a result, the function receives an increment:

$$\Delta Z = ih. \quad (3)$$

In addition, with such a small increment  $\Delta\omega$ , the increment of the first term in (2) is also small because of the continuity of this term at  $\omega = 1$ . Taking into account that the argument varies from  $\pi$  to 0 as we go around the point  $\omega = 1$ , the increment of the second term has the form:

$$\ln \frac{1 - \sqrt{\omega}}{1 + \sqrt{\omega}} = \ln(r) - \ln(re^{i\pi}) = -i\pi.$$

This allows us to write the expression:

$$\Delta Z = \lim_{r \rightarrow 0} [2C\sqrt{\omega} - C(1 - a^2) \ln \frac{1 - \sqrt{\omega}}{1 + \sqrt{\omega}}]_{\omega=r}^{\omega=re^{i\pi}} = C(1 - a^2)(-i\pi). \quad (4)$$

Equating (3) and (4), we obtain:

$$ih = C(a^2 - 1)i\pi.$$

Thus, the change of the coefficient value  $a^2$  can be described by the equation:

$$a^2 = \frac{h}{C \cdot \pi} + 1. \quad (5)$$

The correspondence of the points  $a^2$  and  $A_3$  makes it possible to transform expression (2) to the form:

$$D + ih = \frac{2 \times ha}{(a^2 - 1) \times \pi} + \frac{h}{\pi} \ln \left( -\frac{a+1}{a-1} \right) = \frac{2 \times ha}{(a^2 - 1) \times \pi} + \frac{h}{\pi} \ln \left( \frac{a+1}{a-1} \right) + \frac{h}{\pi} i\pi,$$

$$D = \frac{2 \times ha}{(a^2 - 1) \times \pi} + \frac{h}{\pi} \ln \left( \frac{a+1}{a-1} \right).$$

Whence we obtain the following equality:

$$\exp \left( D \frac{\pi}{h} - \frac{2 \times a}{a^2 - 1} \right) = \frac{a+1}{a-1}. \quad (6)$$

Given specific values of  $D = 0.9$  mm,  $h = 1.2$  mm, we can obtain a solution of the transcendental equation (6) and find the value of the constant  $a^2 = 4.179$ . Substituting it in (5), we obtain  $C = 0.24$  mm.

As a result, the function realizing the conformal mapping of the half-plane  $\omega$  onto the plane  $Z$  has the form:

$$Z = 2C\sqrt{\omega} + \frac{h}{\pi} \ln \left( \frac{1 + \sqrt{\omega}}{1 - \sqrt{\omega}} \right). \quad (7)$$

Thus, expressions (5), (6) allow us to find a constant  $C$  whose value depends on the design parameters  $D$  and  $h$ .

At the second stage, an additional mapping of the half-plane  $\text{Im}\omega > 0$  is applied to the strip  $0 < \text{Im}\xi < V$  with cuts along the corresponding rays (fig. 3). In this case we have a capacitor with infinite plates in the plane  $\xi$ .

Considering only the right triangle with vertices  $B_1B_2B_3$  because of the electrode design symmetry, we put the points 0, 1,  $\infty$  lying on the real axis  $\omega$  in correspondence to these vertices [11]:

$$\begin{array}{ccc} B_1 & B_2 & B_3 \\ 0 & 1 & \infty \end{array}$$

The inner angles  $\beta_k$  at the vertices of the triangle  $B_1B_2B_3$  and the angles  $\mu'_k$ , that complement the angles  $\beta_k$  to  $\pi$ , are defined similarly to  $\alpha_k, \mu_k$ :  $\mu'_2 = 1$  ( $\beta_2 = 1 - \mu'_2 = 0$ );  $\mu'_3 = 1/2$  ( $\beta_3 = 1 - \mu'_3 = 1/2$ );  $\mu'_1 = 1/2$  ( $\beta_1 = 1 - \mu'_1 = 1/2$ ).

The obtained values  $\beta_1 = 1/2, \beta_2 = 0, \beta_3 = 1/2$ , ensure equality  $\sum_{i=1}^3 \beta_i = 1$ , which confirms the correctness of the sought angles values according to [11].

An additional conformal mapping is also determined by the Schwarz-Christoffel integral [10]:

$$\xi = C_2 \int_0^\omega \omega^{-1/2} (\omega - 1)^{-1} d\omega = C_2 \int_0^\omega \frac{d\omega}{(\omega - 1)\sqrt{\omega}} + C_3.$$

The introduction of the new variable  $\omega = u^2$  allows us to obtain the solution of the given integral:



$$\xi = 2C_2 \int_0^{\sqrt{\omega}} \frac{udu}{(u^2-1)u} + C_3 = -C_2 \ln \frac{1+\sqrt{\omega}}{1-\sqrt{\omega}} + C_3. \quad (8)$$

From the correspondence of the points  $B_1 \leftrightarrow 0$  according to the technique outlined above, there is a constant  $C_3$ .

$$\begin{aligned} \xi &= -C_2 \ln \frac{1+\sqrt{0}}{1-\sqrt{0}} + C_3 = 0 + C_3, \\ C_3 &= 0. \end{aligned}$$

The constant  $C_2$  is defined similarly to the constant  $C$  in the first stage, namely, by traversal of the point  $\omega = 1$ , we get the increment

$$\Delta\xi = iV.$$

Since the increment of the argument changes from  $\pi$  to 0 upon traversal of the above point, the increment of function  $\xi$  corresponds to the expression:

$$\Delta\xi = \lim_{r \rightarrow 0} [-C_2 \ln \frac{1+\sqrt{\omega}}{1-\sqrt{\omega}}]_{\omega=r}^{\omega=re^{i\pi}} = -C_2(-i\pi) = C_2 i\pi,$$

which allows us to obtain equality:

$$iV = C_2 i\pi.$$

Solving this equality, we define  $C_2$  :

$$C_2 = \frac{V}{\pi}.$$

The final function conformally mapping the half-plane  $\omega$  to the strip  $0 < \text{Im}\xi < V$  has the form:

$$\xi = \frac{V}{\pi} \ln \frac{1+\sqrt{\omega}}{1-\sqrt{\omega}} = \frac{2V}{\pi} \text{arcth}\sqrt{\omega}. \quad (9)$$

Using (7) and (9), we obtain a system of equations:

$$\begin{cases} Z = 2C\sqrt{\omega} + \frac{h}{\pi} \text{Ln} \left( \frac{1+\sqrt{\omega}}{1-\sqrt{\omega}} \right) \\ \xi = \frac{2V}{\pi} \text{arcth}\sqrt{\omega} \end{cases} \quad (10)$$

From (10) we find

$$Z = 2C \cdot \text{th} \frac{\xi\pi}{2V} + \frac{h}{\pi} \ln \left( \frac{1 + \text{th} \frac{\xi\pi}{2V}}{1 - \text{th} \frac{\xi\pi}{2V}} \right) = 2C \cdot \text{th} \frac{\xi\pi}{2V} + \frac{h}{V} \xi. \quad (11)$$

By separating the real and imaginary parts of equation (11), the parametric equations of lines of equal potential and field lines of force are found. After separation, we obtain a system of equations describing the coordinates of the electric field distribution in the electrode system of a gas-discharge device:

$$\begin{cases} x = \frac{hu}{V} + 2C \frac{\text{sh} \frac{u\pi}{V}}{\text{ch} \frac{u\pi}{V} + \cos \frac{v\pi}{V}} \\ y = \frac{hv}{V} + 2C \frac{\sin \frac{v\pi}{V}}{\text{ch} \frac{u\pi}{V} + \cos \frac{v\pi}{V}} \end{cases}. \quad (12)$$

Substituting the parameters  $h, V, D$  into the expressions (5), (6) and system (12) and changing the values of the variables  $v$  and  $u$  with the necessary step, we can determine the number of the field lines and equipotential distribution (fig.4 and fig.5).

Changing the voltage at the electrodes does not lead to a change in the field configuration, but it affects the energy of charged particles. Thus, the equations system (12) allows to obtain the electrode system configuration to form the required electric field by varying the parameters  $h, V, D$ .

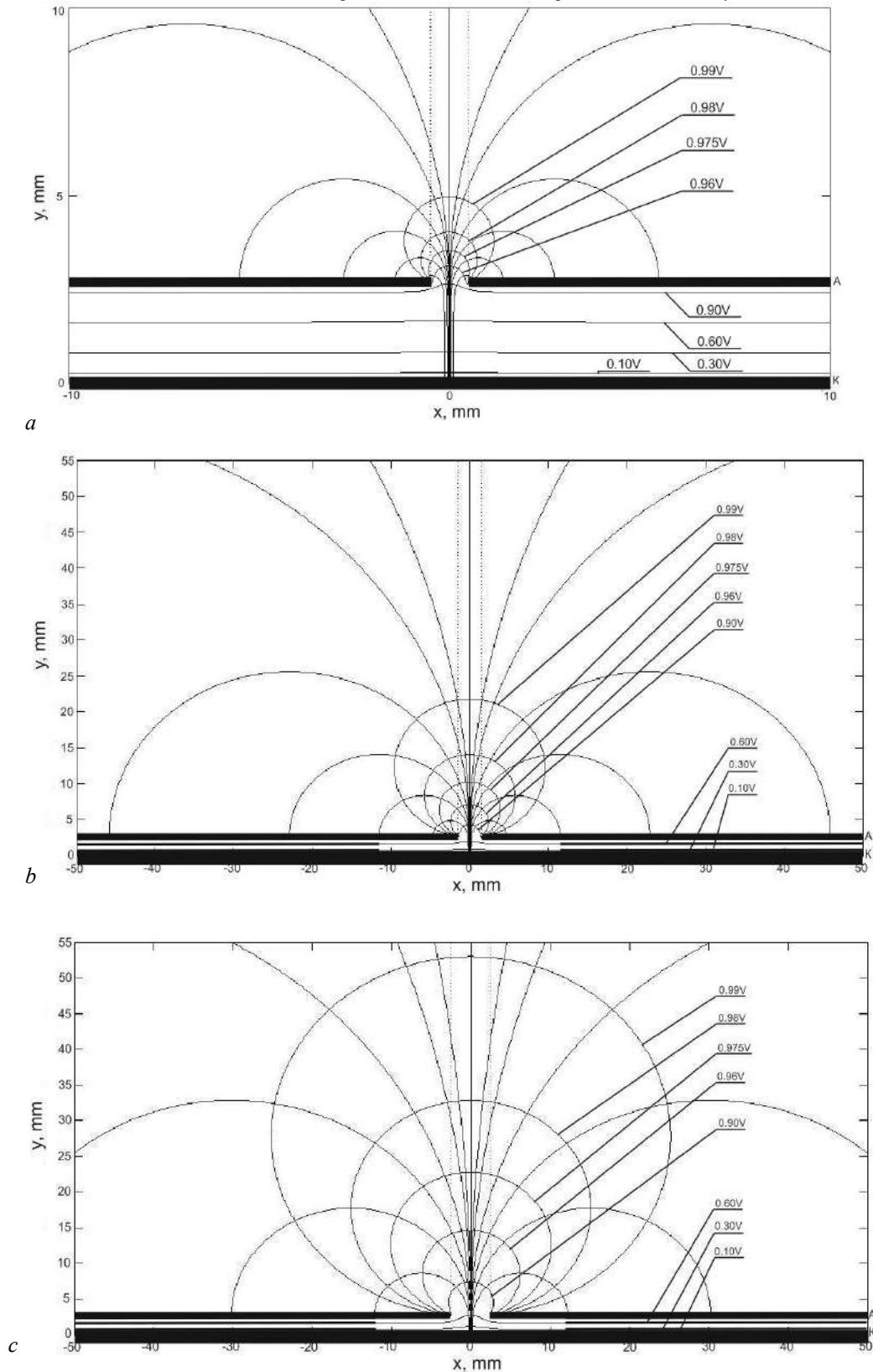


Fig.4. Field lines and equipotentials distribution in the electrode system of the gas-discharge device obtained by the equations system (12):  $a - h = 2.7$  mm,  $D = 1$  mm,  $V = 1200$  V;  $b - h = 2.7$  mm,  $D = 3$  mm,  $V = 1200$  V;  $c - h = 2.7$  mm,  $D = 5$  mm,  $V = 1200$  V.

### 3. Analysis of the field lines and equipotentials distribution

The initial coordinate ( $x = x_0, y = y_0 = 0$ ) of the rectilinear segment of the field line can be determined with the aid of the system (12), giving the values  $u = u_0$  and  $v_0 = 0$ . Then, searching further all the values of  $v = v_1 - v_n$  for which the coordinate  $x = x_0$  is constant, and the  $y$  varies in the limits  $y_1 - y_n$ . Further on, comparing the obtained maximum value of  $y_n$  with the mean free path of the electron  $k\lambda_e$  ( $k = 1, 2, 3$ ) and the potential at the given point with the ionization energy of the working gas atom (molecule)  $E_i$ , we verify the fulfillment of the condition for the emergence of a high-voltage discharge  $\gamma Q \geq 1$  [ 8 ] is, where  $\gamma$  is the number of electrons knocked by one ion from the cathode ( $\gamma$ -process),  $Q$  is the number of positive ions formed by the electron on the trajectory of its motion due to inelastic collisions with atoms and molecules of the working gas ( $\alpha$ -process). The energy

accumulated by the electron on the mean free path must be higher than the ionization energy of the working gas atom, and the energy of the positive ion bombarding the cathode must be sufficient for the emission of electrons necessary for sustaining the self-dependent discharge. Analogously, changing the values  $u = u_l - u_n$  for  $v_0 = 0$ , the corresponding  $x = x_l - x_n$  are determined. Further on, searching further values of  $v = v_l - v_n$  for each  $x$ , we find  $y = y_l - y_n = 0 - k\lambda_e$ . In other words, by repeating the comparison process, we can find all field lines with initial coordinates  $x_0, \dots, x_{\lambda_e}$ , on the rectilinear segments of which the ionization process takes place ( $\alpha$ -process), and, accordingly, the length of the cathode region  $\Delta x = 2x_{\lambda_e}$  where the electron emission from the cathode ( $\gamma$ -process) takes place [13].

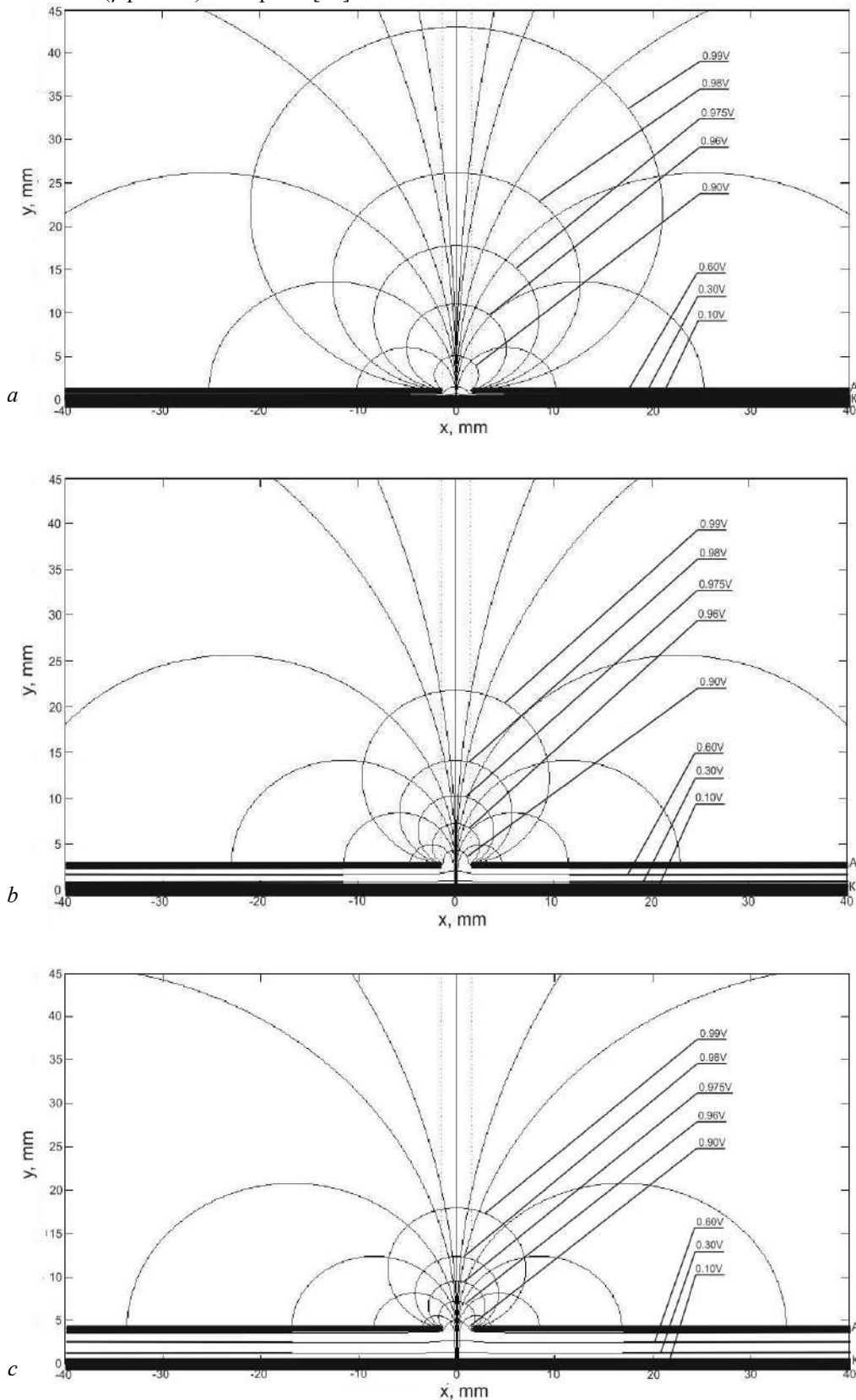


Fig.5. Field lines and equipotentials distribution in the electrode system of the gas-discharge device obtained by the equations system (12):  $a - h = 1$  mm,  $D = 3$  mm,  $V = 1200$  V;  $b - h = 2.7$  mm,  $D = 3$  mm,  $V = 1200$  V;  $c - h = 4$  mm,  $D = 3$  mm,  $V = 1200$  V.

In order to compare the maximum values of  $y_n$  with  $k\lambda_e$ , it is necessary to find the mean free path of an electron. Using the expression  $\lambda_e = l/(N\sigma_e)$  [14], we obtain the value 0.203 cm, which makes it possible to determine  $\Delta x = 318 \mu\text{m}$ . The calculated value of  $\Delta x$  correlated well with the experimental data of [9], namely, the size of the region on the cathode surface with intense sputtering by positive ions is 300  $\mu\text{m}$  (fig. 6).

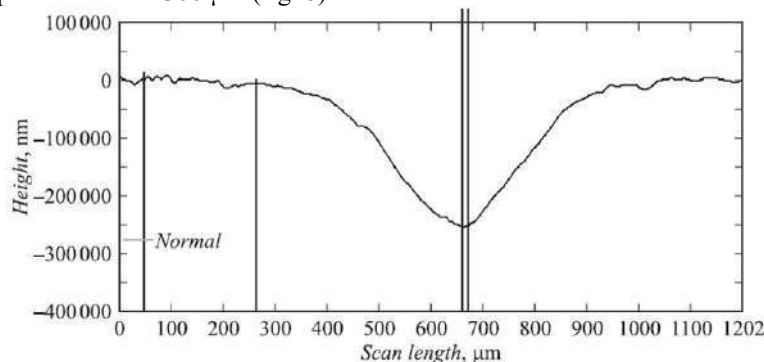


Fig 6. The profile of the etching pit on the surface of the cathode formed by positive ions.

This value is comparable with the size of the region  $\Delta x$  on which the rectilinear segments of field lines correspond to the value  $k\lambda_e$  and the condition for the emergence of an high-voltage discharge is observed.

#### 4. Conclusions

The parametric equations system presented in this paper makes it possible to simulate the of the field lines and equipotentials distribution in the electrode system of the off-electrode plasma generator and to monitor the dependence of this distribution on the design parameters of the system: the anode-cathode distance, the hole diameter in the anode, and also on the applied voltage at the electrodes. In addition, estimates are made in this paper: the length of the rectilinear segments of the field lines on which the condition is satisfied, the size of the cathode spot  $\Delta x$  within which the  $\gamma$  process is realized. The discrepancy between the calculated value and the experimental value does not exceed 6%, which indicates that the model corresponds to the actual physical processes occurring in the electrode system of a high-voltage gas discharge. Therefore, it becomes possible to optimize the devices design forming the off-electrode plasma without costly experimental investigations.

#### Acknowledgments

This work was supported by grants from the President of the Russian Foundation for State Support of Young Russian Ph.D. Scientists (MD-5205.2016.9) and the Russian Foundation for Basic Research (project no. 16-07-00494 A).

#### References

- [1] Komov AN, Kolpakov AI, Bondareva NI, Zakharenko VV. Electron-beam unit for soldering semiconductor devices. *Instruments and Experimental Techniques* 1984; 5: 218-220.
- [2] Kazanskiy NL, Kolpakov VA, Kolpakov AI. Investigation of the features of the anisotropic etching of silicon dioxide in plasma high-voltage gas discharge. *Microelectronics* 2004; 33(3): 209-224.
- [3] Kolpakov VA, Kolpakov AI, Krichevskiy SV. Ion-plasma cleaning of the low-power relay contacts surface. *Elektronnaya promyshlennost* 1996; 2: 41–44.
- [4] Kazanskiy NL, Kolpakov VA, Krichevskiy SV. Simulation of the cleaning process the surface of dielectric substrates in plasma formed by high-voltage gas discharge. *Computer Optics* 2005; 28: 80–86.
- [5] Kazanskiy NL, Kolpakov VA. Investigation of the mechanisms of low-temperature plasma formation by a high-voltage gas discharge. *Computer Optics* 2003; 25: 112–116.
- [6] Kazanskiy NL, Kolpakov VA, Kolpakov AI, Krichevskiy SV. Gas discharge devices forming directed flows of the off-electrode plasma. Part I. *Nauchnoe priborostroenie* 2012; 22(1): 13–18.
- [7] Soifer VA, Kazanskiy NL, Kolpakov VA, Kolpakov AI. Patent 2333619 Russian Federation, MPK H 05 H 1/24. Multi-beam gas-discharge plasma generator. Applicant and patent holder IPSI RAS N 2006121061; declared 13.06.06; published 10.09.08, bulletin. N 25; 5 p.
- [8] Kazanskiy NL, Kolpakov VA. Formation of an optical microrelief in off-electrode high-voltage gas discharge plasma. *M. : Radio and Communications*, 2009; 220 p.
- [9] Kolpakov VA, Kolpakov AI, Podlipnov VV. Formation of the off-electrode plasma in a high-voltage gas discharge. *Technical Physics* 2013; 83(4): 41–46.
- [10] Miroljubov NN, Kostenko MV, Levinshtejn ML, Tihodeev NN. Calculation methods of electrostatic fields. M.: Vysshaja Shkola, 1963; 415 p.
- [11] Lavrent'ev MA, Shabat BV. Methods of the theory of functions of a complex variable / M.A. Lavrent'ev. M.: Nauka, 1973; 736 p.
- [12] Novgorodtsev AB, Fethiev AR, Fethieva IS. Application of complex variable function to calculation of electrostatic fields of irregular shape electrodes: a tutorial. Ufa: Ufimskij ordena Lenina aviacionnyj institut im. Sergo Ordzhonikidze, 1986; 82 p.
- [13] Raizer YuP. Gas discharge physics. M.: Nauka, 1987; 592 p.
- [14] Kudryavtsev AA, Smirnov AC, Tsendin LD. Physics of glow discharge: a tutorial. SPb.: Lan', 2010; 512 p.

# Approaches to the optimization of the placement of service-oriented cloud applications in the software-defined infrastructure of the virtual data center

I. Bolodurina<sup>1</sup>, D. Parfenov<sup>1</sup>, K. Haenssger<sup>2</sup>

<sup>1</sup>Orenburg State University, 13 Pobedy ave., 460018, Orenburg, Russia

<sup>2</sup>Leipzig University of Applied Sciences, 132 Karl-Liebknecht-Straße, 04251, Leipzig, Germany

---

## Abstract

Nowadays, we see a steady growth in the use of service-oriented cloud applications in modern business. However, there are some issues related to the placement of service-oriented cloud applications in the software-defined infrastructure of the virtual data center. The goal of optimization is to control the service-oriented cloud applications within data centers. The advantage of modern infrastructure virtualization is the possibility to use software-defined networks and software-defined data storages. However, the existing optimization of algorithmic solutions does not take into account the specifics of working with multiple class service-oriented cloud applications types.

The paper describes the models which describe the basic structures of service-oriented cloud applications including: a level distribution model of the software-defined infrastructure with the technology of cloud applications containerization, a generalized model of a service-oriented cloud application, a model of virtualization of service-oriented cloud applications based on containers. Besides, we developed the efficient algorithm for optimizing the technology of containerization of cloud applications and services in the virtual data center (VDC) infrastructure. We propose an efficient algorithm for placing applications in the infrastructure of a VDC. The optimization of the placement of service-oriented cloud applications based on the VM template and containers with VDC disabilities infrastructure is reduced to packing in containers. Besides, we generalize the well-known heuristic and deterministic Karmakar-Karp's algorithms.

During the experimental studies, we have found that the use of our algorithm enables to decrease the response time of cloud applications and services and, therefore, to increase the productivity of user requests processing and to reduce the number of refusals.

*Keywords:* software-defined network; virtual data center; cloud applications and services; IT infrastructure; virtualization

---

## 1. Introduction

The technology of cloud computing is based on the virtualization of the individual components that make up the data center infrastructure. The approaches to the organization of the virtualization layer are divided according to the levels of application of this technology. Typically, the following types of virtualization are distinguished: operating system, software, memory, data storage, databases and network [1, 6]. The most active development level is the virtualization of the operating system. It allows you to create a virtual environment for running multiple instances of user space within the same operating system used to run service-oriented cloud applications within the network environment [24].

Nowadays, cloud applications are a fairly complex multi-level mechanism that interacts with various objects of the network infrastructure of the virtual data center in the course of its work [3, 5]. To deploy them, you need to use an integrated approach that can provide the performance with the least resource consumption. One of the approaches to virtualization, which allows implementing this approach, is containerization technology. A container is an object that provides the user with access to necessary libraries and contains the required set of software for launching the development environment, a ready application or service. Besides, the advantage of using virtualization based on containers for placing applications and services in the network environment of the virtual data center is an easy solution, which enables to reduce costs and increase the productivity of cloud-based service-oriented applications. The control system based on Docker the most effective technology of containerization [4]. The advantage of this technology is its reliability, the availability of open source code and a convenient API for sharing in a networked virtual data center environment. A significant drawback of containers is the inability to provide the proper level of data flows isolation, as well as the lack of support for migration between compute nodes in real time [7-9].

A key difference between virtual machine virtualization and virtualization is the use of a hypervisor that emulates the hardware of a physical computing node. Within each virtual machine, full-fledged operating system functions based on one or several cloud applications or services can be deployed. All this leads to significant overhead in terms of resource consumption in a virtual data center [2].

In this study, we propose a solution that enables to solve the problem of container migration and to optimize the overhead costs for computing resources for the virtual data center infrastructure. The developed solution is based on hybrid virtualization, which is a combination of two approaches to the deployment of cloud applications and services in the software-defined infrastructure of a virtual data center: a container and based on virtual machines.

With the resources of virtualization technology development, the number of layers that form infrastructure decisions and are used in the cloud computing technology is steadily increasing. Nowadays, we can find up to six levels of the software-defined infrastructure of the virtual data center used for the deployment of modern cloud platforms.

The paper is organized as follows. Section 2 is devoted to a level distribution model of the software-defined infrastructure with the technology of containerization of service-oriented cloud applications. Section 3 describes a generalized model of a service-oriented cloud application. Section 4 deals with the model of the virtualization of service-oriented cloud applications based on containers. Section 5 describes an optimization algorithm for the launching and deployment of applications and services in the virtual data center infrastructure using different methods of placement. Section 6 provides the experimental results of our investigation. Section 7 is devoted to the traditional approaches to route traffic based on load-balancing and the

solutions of this problem proposed by world scientists. Conclusion section includes a summary of our investigation and the overview of our future work.

Let us describe a model of the virtual data center structure based on the technology of containerized service-oriented cloud applications.

## 2. The level distribution model of the software-defined infrastructure with the technology of containerization of service-oriented cloud applications

Let us introduce the level model of the software-defined infrastructure of the virtual data center, which supports the containerization method of applications and services placement in the cloud system. The first level is the hardware component of any data center, which includes computing nodes (Nodes), filing systems (Storages) and physical network units (NetObj). Let us introduce it as a set of solutions:

$$PhysLayer = \{Nodes, Storages, NetObj\}. \quad (1)$$

The next level represents the software-defined layer. This layer consists of the same number of objects as the first level but the main difference is that all the infrastructure elements are dynamic, easily transformed and adjusted within the limits of the physical database network environment. The second level can be presented as the following set of connections:

$$SDLayer = \{SDNodes, SDStorages, SDNetwork\}, \quad (2)$$

where *SDNodes* are software-defined computing units; *SDStorages* are software-defined storages; *SDNetwork* is a software-defined network.

Above the layer of the software-defined infrastructure, there is a level of the specific objects virtualization. The main objects are computing nodes (VirtNodes), virtual data storages and the elements of the software-defined network used in the work of the cloud platform and consolidated in VirtNetwork multitude.

$$VirtLayer = \{VirtNodes, VirtStorages, VirtNetwork\} \quad (3)$$

In the software-defined infrastructure, computing nodes and data storages are more often presented as virtual machines that discharge the set of given functions.

To control such multi-layer infrastructure, a separate orchestration layer is needed (the forth level). It contains a number of functions as well as computing nodes and program systems to execute them. The main functions are to orchestrate the virtualization objects (virtual machines and data storages) (ONodes, OStorages) and the software-defined network (ONetwork). Lately, experimental Network function virtualization is also added to them.

$$OrchLayer = \{ONodes, OStorages, ONetwork\}. \quad (4)$$

The next level (service level) represents the services used in the working process (either the very cloud platform, or applications distributed there, for example, DBMS, Hadoop, Nginx and others).

$$ServiceLayer = \{Service_1, \dots, Service_n\}. \quad (5)$$

All the multitude of ServiceLayer cloud services that work in the virtual data center infrastructure can be divided into two disjoint subsets  $ServVM \cup ServDocker = ServiceLayer$ . The first set (Serv VM) involves services that use virtualization based on other machines. In the second set (ServDocker), there are services based on containers under Docker control.

The top level includes cloud applications that are exploited by users for flexible scalability providing (AppLayer).

$$AppLayer = \{App_1, \dots, App_m\}. \quad (6)$$

Like at the previous level, cloud applications *App<sub>i</sub>* can be placed in containers and form the AppVM set. Or they can form the AppDocker set using containerization. At the same time,  $AppVM \cup AppDocker = AppLayer$ .

Thus, the set of objects of the software-defined infrastructure can be divided into two groups by the methods of placing. Virtual objects that use a container placing method can be referred to the first group. Let us describe them in this way:

$$Docker = \{ServDocker, AppDocker\} \quad (7)$$

In the second group, there are services and applications that use virtual machines as a placing platform:

$$VM = \{ServVM, AppVM\} \quad (8)$$

Before we talk about the ways of placing service-oriented cloud applications in the virtual data center, we need to determine their structure, basic parameters, and key characteristics of their operation that affect the efficiency of their use. For this purpose, we have developed a generalized model of a service-oriented cloud application.

## 3. A generalized model of a service-oriented cloud application

The specific feature of the service-oriented cloud applications is the approach, where users have access to them and to their services; however, they do not know anything about their actual location. In most cases, users only know the address of the aggregation node and the application name. The cloud system automatically selects the optimal virtual machine for the request, on which it is to be processed.

The generalized model of the service-oriented cloud application is a multilayer structure, described in a form of graphs to characterize the connections of individual elements. The model can be represented in the form of three basic layers, detailing the connections of the specific objects of virtual cloud infrastructure: applications, related services and provided resources.

The cloud application is a weighted directed acyclic graph of data dependencies:

$$CloudAppl = (G, V), \quad (9)$$

Its vertices  $G$  are tasks that get information from the sources and process it in accordance with the user requests; its directed edges  $V$  between corresponding vertices are a link between tasks in a schedule plan. The schedule plan is defined as a procedure which is prepared to follow the user's request (*SchemeTask*).

Each vertex  $g \in G$  is characterized by the following tuple:

$$g = (\text{Re } s, \text{NAppl}, \text{Utime}, \text{SchemeTask}), \quad (10)$$

where *Res* are the resource requirements; *NAppl* is the number of application instances; *Utime* is the estimated time for of the users' request execution; *SchemeTask* is a communication scheme of data transmission between sources and computing nodes.

Each directed edge  $v \in V$  connects the application with the required data source. It is characterized by the following tuple:

$$v = (u, v, \text{Tdata}, \text{Mdata}, \text{Fdata}, \text{Vdata}, \text{Qdata}), \quad (11)$$

where  $u$  and  $v$  are linked vertices; *Tdata* is the type of transmitted data; *Mdata* is the access method to the data source; *Fdata* is the physical type of the accessed object (a file in the storage system, a local file, distributed database, data services and so on); *Vdata* is the traffic volume estimated by the accessed data (in Mb), *Qdata* is the requirements for the QoS (quality of services).

The model is original because it enables to calculate the consolidated assessment of its work with data sources for each application. It allows predicting the performance of the whole cloud system.

As mentioned earlier, a cloud service is one of the key slices in the generalized model of a cloud application. The cloud service is an autonomous data source for the application, for which it acts as a consolidated data handler. Generally, the cloud service is highly specialized and designed to perform a limited set of functions. The advantage of connecting a cloud application to the service is the isolated data processing, in contrast to direct access to the raw data, when a cloud application does not use a service. The usage of services reduces the execution time for user requests. The cloud service is described as a directed graph of data dependencies. The difference lies in the fact that from the user's viewpoint, the cloud service is a closed system.

The cloud service can be formalized as a tuple:

$$\text{CloudServ} = (\text{AgrIP}, \text{NameServ}, \text{FormatIN}, \text{FormatOUT}), \quad (12)$$

where *AgrIP* is the address of aggregation computing node; *NameServ* is the service name; *FormatIN* is the format of input data; *FormatOUT* is the format of output data.

The aggregator of a service selects the optimal virtual machine; it is executed on this machine. In addition, all its applications are distributed between predefined virtual machines or physical servers. Their new instances are scaled dynamically depending on the number of incoming requests from cloud applications, users or other services.

To describe the placement of cloud applications and services in the data center infrastructure, we have also implemented the model of a cloud resource. A cloud resource is an object of a data center, which describes the behavior and characteristics of the individual infrastructure elements, depending on its current state and parameters. The objects of data center are disk arrays including detached storage devices, virtual machines, software-defined storages, various databases (SQL/NoSQL) and others. In addition, each cloud service or application imposes requirements on the number of computing cores, RAM and disk sizes, and the presence of special libraries on physical or virtual nodes used to launch their executing environments.

Each cloud resource can be formalized as follows:

$$\text{CloudRes} = (\text{TRes}, \text{Param}, \text{State}, \text{Core}, \text{Rmem}, \text{Hmem}, \text{Lib}), \quad (13)$$

where *TRes* is the type of resource; *Param* is the set of parameters; *State* is the state of resource; *Core* is the number of computing cores; *Rmem* is the size of RAM; *Hmem* is the size of a disk; *Lib* are for the libraries requirements.

The distinctive feature of the model suggested implies analyzing cloud resources from the user viewpoint and from the viewpoint of a software-defined infrastructure of the virtual data center. The model is innovative, since it simultaneously describes the application data placements and the state of the virtual environment, taking into account the network topology.

We have developed the model of the software-defined storage, which details the resource model of the virtual data center. It is represented in the form of a directed multigraph; its vertices are the virtual data center elements, which are responsible for application data placement (e.g. virtual disk arrays, DBMS and so on):

$$\text{Stg}_{ki} = (\text{Max}V_{ki}, P_{ki}^{stg}, \text{Vol}_{ki}(t), \bar{R}_{ki}(t), \bar{W}_{ki}(t), s_{ki}^{stg}(t)), \quad (14)$$

where  $\text{Max}V_{ki} \in N$  is the maximum storage capacity in Mb;  $P_{ki}^{stg} = \{P_{kij}^{stg}\}_j$  is the set of network ports;  $\text{Vol}_{ki}(t) \in N \cup \{0\}$  is the available storage capacity in Mb;  $\bar{R}_{ki}(t)$  and  $\bar{W}_{ki}(t)$  are read and write speeds;  $s_{ki}^{stg}(t) \in \{\text{"online"}, \text{"offline"}\}$  is state of software-defined storage.

The data storage system for applications is like a layer cake. It uses the principles of self-organization of resources. The basis of self-organization of data storages is an adaptive model of dynamic reconfiguration when resources are changing. The model allows optimizing the organizational structure of the cloud platform based on algorithms for searching optimal control nodes and allocating control groups. Our control model assumes two control levels for nodes and resources.

When a software-defined storage is created on each virtual computing node, the software module for exchanging state data between devices is executed. This exchange is carried out within a group of nodes by a single storage method. The least loaded node in the group is selected as the control node. This approach reduces the risk of the control node degradation.

If the control node is failed, the remaining group of virtual machines has all the information about each other, which allows choosing a new control node automatically. Each control node also carries out cooperation with control nodes from other groups to maintain up-to-date information on the state of the entire system.

Thus, the system of software-defined storages is constructed as a hierarchy that includes three basic levels: the level of local access, the level of the controlled group, and the level of data exchange within the whole system. In our model, the description of cloud applications consists of task descriptions and data source descriptions specifying directions and methods of data transfer as well as required resources.

The data obtained allows us to proceed to a description of the model of virtualization service-oriented cloud applications on the basis of containers.

#### 4. A model of the virtualization of a service-oriented cloud applications based on containers

Let us describe the formalized structure and communications of cloud applications and services to describe the model of virtualization using the method of containerization.

Every cloud application can be described as a set of components, which are the following union of sets:

$$App_i = Lib \cup Qu \cup StgData \cup Service \cup PM \quad (15)$$

where Lib is a set of operating system libraries used in the application; Qu is a set of queues formed at the requests of the users accessing the application; StgData is set of storage systems used for placing data for cloud applications; Service is a set of services that uses cloud applications in the course of their work; PM is set of methods for placing cloud applications in the software-defined infrastructure of the data center.

In this study, we consider three methods of placing cloud applications. The first method is based on using virtual machines –  $P_{vm}$ . The next method involves the use of containers placed on physical compute nodes –  $P_d$ . The last method uses a hybrid approach based on containerization inside a virtual machine –  $P_{dvm}$ .

A service-oriented application ( $App_i$ ) placed in a software-defined infrastructure of data center is a set of instances running in the cloud platform  $Vapp \in App_i$ . Thus, at each moment of time, the cloud application may be represented as a dynamic weighted directed graph:

$$App_i(t) = (Vapp, Eapp, FlowApp(e_a, t), Iapp(vapp)_i) \quad (16)$$

where Vapp is a set of nodes that represent instances of cloud applications placed in the software-defined infrastructure of the data center; Eapp is the maximum total number of network connections forming the graph of the arc in the process of balancing the requests between instances Vapp; FlowApp( $e_a, t$ ) is the function that determines the number of transmitted data on the arc  $e_a \in Eapp$  at time  $t \geq 0$ . If FlowApp( $e_a, t$ ) = 0, no arc  $e_a$  at time  $t$ ; Iapp(vapp)<sub>i</sub> is a set of the characteristics of an instance of the cloud applications  $vapp \in Vapp$ .

In turn, each cloud service in the software-defined infrastructure of the data center can be described by the following set of parameters:

$$Service_i = \{AgrIP, NameServ, Ma, PM, Format\} \quad (17)$$

AgrIP is the address of computing node for requests aggregation to cloud service; NameServ is a name of the cloud service placed in the infrastructure of the virtual data center; Ma is a set of supported methods of access to the service; PM is a set of methods for placing cloud service in the software-defined infrastructure of the data center; Format is the data format.

Cloud service is a set of instances running in the virtual data center infrastructure. Thus, it can be represented as a dynamic weighted directed graph:

$$Service_i(t) = (Vserv, Eserv, FlowServ(e_s, t), Iserv(vserv)_i) \quad (18)$$

where Vserv is a set of vertices, which are the running instances of the cloud service;

Eserv is the most complete set of arcs (network connections) allowed between multiple cloud applications vertices (application-level) and service instances  $vserv \in Vserv$  and deployed in a virtual data center; FlowServ( $e_s, t$ ) is a function that determines the number of transmitted data on the arc  $e_s \in Eserv$  at time  $t \geq 0$ . If FlowServ( $e_s, t$ ) = 0, no arc  $e_s$  at time  $t$ ; Iserv(vserv)<sub>i</sub> is a set of the characteristics of an instance of the cloud service  $vserv \in Vserv$ .

Every instance of a running cloud application  $vapp \in Vapp$  or a cloud service  $vserv \in Vserv$  has the vectors:

$$Iapp(vapp) = n_i(t), m_i(t), u_i(t), \Delta t_i, p_i; \quad (19)$$

$$Iserv(vserv)_i = n_i(t), m_i(t), u_i(t), \Delta t_i, p_i \quad (20)$$

where  $n_i(t)$  is the number of requests flows that are processed for one instance at the moment of time  $t$ ;

$m_i(t)$  is the volume of consumed memory for one instance of application for placing on the computing node at the time  $t$ ;

$u_i(t)$  is the average load cores on the computing node for one instance of application at the time  $t$ ;

$\Delta t$  is the average time of response to the incoming flow of requests for the instance;

$p \in PM$  is a method of placing an instance in the virtual data center infrastructure.

The developed model describes the mapping of the service-oriented cloud applications in the virtual data center infrastructure using different methods of placement.

Let us describe the flow of user requests coming to the service-oriented cloud applications by the following functional:

$$Fur = (U, AppLayer, Q) \quad (21)$$

where U is a set of users; AppLayer = {App<sub>1</sub>, ..., App<sub>m</sub>} is a set of service-oriented cloud applications; Q is a set of user requests.

To ensure the efficient use of the resources of the virtual data center and the required quality of service to users, we formulate the optimization problem. The flow of user requests should be distributed efficiently between the running instances of the service-oriented cloud applications.

$$Fur(t) : Q \rightarrow Vapp \quad (22)$$

At the same time, the copies of applications and services should be optimally placed in the virtual data center infrastructure.

$$App_i(t) : Vapp \rightarrow PM \quad (23)$$

It was found that the consumption of basic resources of the virtual data center using different methods of the placement of a set of service-oriented cloud applications has a different weight. To take into account this feature, we introduce the model of optimizing the weighting factors  $k_1, k_2, k_3$  for each type of placement. Then, the function of resource consumption will be:



$$Rvapp_i = \sum_{l=1}^L k_l Rapp_i \quad (24)$$

$$Rvserv_i = \sum_{l=1}^L k_l Rserv_i \quad (25)$$

$Rapp_i$  and  $Rserv_i$  – is a basic weight of resource-intensive applications and services, respectively.

We use the expression “optimal placement” to denote the minimum number of the instances of running applications and services. This ensures minimal consumption in the virtual data center resources. Besides, this approach supports the response time within the allowable value for serving maximum number of users per unit time. This can be formalized as follows:

$$\begin{aligned} \sum_{i=1}^N \sum_{j=m}^M vapp_i^j Rvapp_i &\rightarrow \min \\ \sum_{i=1}^N \sum_{j=m}^M vserv_i^j Rvserv_i &\rightarrow \min, \\ \sum_{i=1}^D Fu_i &\rightarrow \max \end{aligned} \quad (26)$$

where  $i=1 \dots D$  is a number of applications received in the interval of time.

This will minimize the number of concurrent computing devices in the virtual data center infrastructure and maximize processing user requests at a given time interval  $\Delta T$ .

To achieve these requirements, we should observe a number of functional limitations.

The time of response to user's request is limited and must not exceed the permissible value. The following restrictions apply to architecture of applications. The response time of the request queue must be less than the maximum response time to a request to the application. Otherwise, the request not will be serviced. Another limitation is the request time of cloud applications and service response time to a request data for application from storage or services.

$$Tu_{resp} \leq Tu_{resp}^{\max} \quad (27)$$

$$Tqu_{resp} + Tapp + Tserv < Tu_{resp}^{\max} \quad (28)$$

## 5. The algorithm of optimizing the launch and deployment of applications and services in the virtual data center infrastructure using different placement methods

The models presented allow us to choose the most suitable methods of placing the instance of cloud applications and services in the virtual data center infrastructure based on the current load and the incoming flow of requests. The main task of the distribution of cloud applications and services is to choice the number of instances in time interval, which is formulated as making a plan. When accessing to the service-oriented cloud applications, it is especially important to prepare a plan. The load on the compute nodes may vary greatly over relatively short time intervals and depend on the method of placement of the virtual data center infrastructure. To solve the optimization problem, we developed an algorithm to monitor the virtual data center infrastructure and schedule and launch applications and services. It is based on a biased random-key genetic algorithm (BRKGA). However, in comparison with the BRKGA, the algorithm uses the heuristic analysis of request flows and their classification depending on the application placed in the virtual data center.

The enlarged algorithm has the following steps.

Step 1. Evaluate the incoming flow of requests to cloud applications. Group the requests by type of application. Rank the type of application by the number of requests.

Step 2. Count the number of running instances of each cloud application and determine the amount of used cloud services. Determine the load on the physical computing nodes. Rank the cloud applications and services on the load generated by the virtual data center infrastructure.

Step 3. Based on the data obtained in step 1 and 2, compare the data and determine the applications and services that require scaling.

Step 4. For applications and services that are not involved in the processing, to implement release of the resources . Add the minimum number of instances of using containers.

Step 5. Find the applications and services that require scaling and which creating the maximum load on the infrastructure to evaluate the method of placement.

Step 6. Distribute the most loaded applications and services using a hybrid method of placing (containers deployed in a virtual machine).

Step 7. Translate less loaded applications and services, which require scaling, into operation in the virtual machine.

Step 8. Move virtual machines to the least loaded nodes.

The approach used in the proposed algorithm of controlling service-oriented applications takes into account the way of accommodation and organizes the work of the virtual data center. It also takes into account the incoming flow of user requests while adjusting the number of running instances of applications and services.

## 6. Experimental part

The aim of the experimental research is to define the effectiveness the algorithm for placing service-oriented cloud applications in the virtual data center infrastructure.

To evaluate the performance of applications, we have used the flows of different intensity. In the first case, flows create minimum load capacity (to evaluate response time and delays, which make data center infrastructure) (experiment 1). In the second case, we have created workload applications placed in the data center virtual infrastructure, traditional for each application. Thus, we can to evaluate the application response time (experiment 2). In the third case (experiment 3), we have applied the developed algorithm for load balancing between the instances of applications and services. We have defined the consumption of resources by each of the running instances; therefore, we can predict the required resources for a third computing experiment.

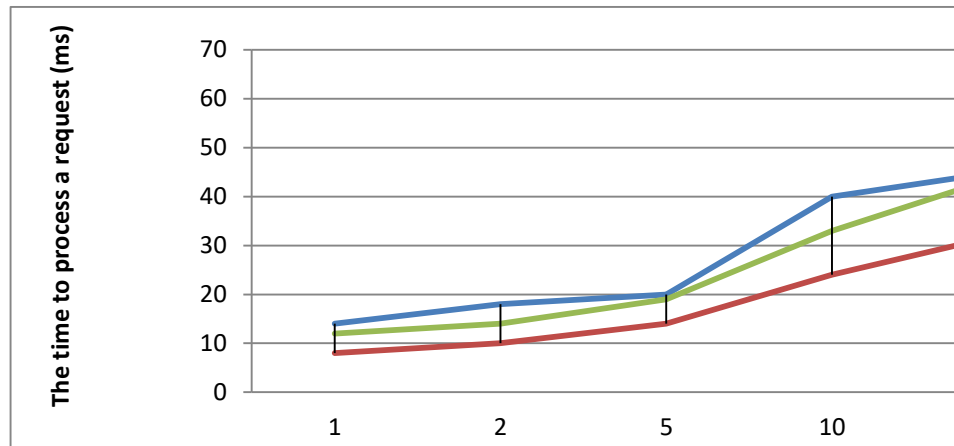


Fig. 1. The result of computing experiment.

The research has shown that the static placement of containers on the physical nodes is not effective because it does not allow redistributing the load quickly. In addition, the movement of the container to another computing unit leads to a loss of the current connections. The placement of applications based on virtual machines due to the flexibility of load balancing showed better results; however, the load on computing nodes has increased considerably due to the additional overhead associated with the use of virtual machines. In this research, the most effective placement was the use of containers inside the virtual machines. It is possible to increase the density of the placement of applications and managed services and software within the data center. Besides, we can place containers as well as data services and network applications in close proximity to each other. Thus, we reduce the time of response to users' requests by applications and increase the efficiency of the system.

## 7. Discussion

Traditional approaches to route traffic based on load-balancing are reactive. They use simple classical algorithms for distributed computing tasks First Fit or Best Fit. Such algorithms as [10–13] First Come First Served Scan, Most Processors First Served Scan, and Shortest Job First Scan are popular too. Their main disadvantage is poor utilization of a computer system due to a large number of windows in the task launch schedule and problem with “hanging up” when their service is postponed indefinitely due to tasks of higher priority. The solution proposed by D. Lifka from Argonne National Laboratory is usually applied as an alternative method of load distribution between nodes. It is based on the aggressive variant of Backfill algorithm [10, 11, 13] and has two conflicting goals – a more efficient use of computing resources by filling the empty windows schedule and prevention of problems with “hanging up” due to redundancy mechanism. D. Feytelson and A. Weil offered a conservative variant of Backfill algorithm [11]. Further, various modifications have been created by B. Lawson and E. Smyrni [13], D. Perkovic and P. Keleher [14], S. Srinivasan [15]. The main drawback of these algorithms is the time lag during calculation, which is not acceptable for critical services at the time of failure.

In addition to the traditional reactive fault-tolerant technology, such as replication and redundancy to ensure reliability of networked storage cloud platforms, a group of scientists from Nankai University proposed an approach based on the Markov model, which provides secure storage of data without excessive redundancy [16]. However, a significant drawback of this model is the lack of classification and analysis of the types and sources of data to be placed in their consumption. Nevertheless, the model demonstrates a proactive approach that gives certain advantages to achieve the desired resiliency of cloud storage.

Reliability and availability of applications and services play an important role in the assessment of its cloud platform performance. A major shortcoming of existing software reliability solutions in the data center infrastructure is the use of traditional data flow routing methods. In this work, we offer to use the software-defined network technology to adjust the network to the current load of the applications and services that are hosted in a cloud platform before they start using pre-computed and installation routes of transmission (in case of known oriented acyclic graph task dependencies and communication schemes). The principles of a software-defined network first emerged in research laboratories at Stanford and Berkeley, and are currently being developed by the Open Network Foundation consortium, GENI project, the European project OFELIA [17] and the Russian University Consortium for the Development of Software-Defined Network Technology with Orenburg State University as its member.

Centralized decision on the organization of data center heterogeneous infrastructure proposed in the papers has some drawbacks including reliability support, cost of obtaining a complete and current network conditions, low scalability [18, 19]. We assume that the development of a fully decentralized solution is the best option; however, in this case, there is a problem of interaction between the controllers of autonomous systems. We are going to address this issue within a framework of our research using SDX technology, which will be extended to exchange not only information about the network, but also distributed sections and condition of cloud services and applications.

The algorithms for routing data flows in a software-defined network in case of track selection published in scientific sources do not take into account the need to ensure the QoS parameters for the previously installed and routed data flows [21, 22]. We are going to do it within a framework of the developed methods of adaptive network communications routing.

The existing QoS algorithms to provide a software-defined network are also quite inefficient. The paper [20] describes an approach to dynamic routing of multimedia flows transmission that provide a guaranteed maximum delay via the LARAC algorithm (Lagrangian Relaxation Based Aggregated) [22]. However, the authors consider only the cases of single delays on each network connection and do not take into account the minimum guaranteed bandwidth. A similar approach is described in the paper [21]; the authors pose and solve the optimization problem for the transfer of multimedia traffic without losses on alternative routes, leaving the shortcuts for common data.

The researchers from Stanford have offered an algorithm for adaptive control of QoS Shortest Span First, which enables to calculate the optimal priorities for each flow mathematically, to minimize crosstalk influence of flows on delay, to manage priorities dynamically depending on the current situation, and to lay the flow of data transmission through specific port queues [23].

We are going to formulate optimization problems for laying routes with QoS constraints and load balancing within a framework of adaptive routing methods of network communications cloud services and applications developed in this research. In their solution, we may use heuristics similar to the Shortest Span First algorithm. Besides, we will account for the distributed nature of a cloud platform.

The analysis of scientific sources on the topic of the study has shown that:

- a) so far, there are no effective algorithmic solutions for planning virtual machines, cloud services, application-oriented accounting topology of the computer system, and communication tasks schemes;
- b) the existing solutions for managing distributed scientific computing on multi-cloud platforms plan computing tasks without subsequent adjustment of network to their communication schemes and use traditional routing methods;
- c) the existing methods of data flow routing can be enhanced by taking into account the QoS requirements and distributed nature of a heterogeneous cloud platform.

This demonstrates the novelty of the solutions offered by the project.

Thus, the development of new methods and algorithms to improve the efficiency of cloud computing with the use of heterogeneous cloud platforms is a crucial task.

## 8. Conclusion

We propose an efficient algorithm for placing applications and services in the infrastructure of a virtual data center. The optimization of placing service-oriented cloud applications based on the VM template and containers with disabilities infrastructure of the virtual data center is reduced to packing in containers. We also generalize the well-known heuristic and deterministic Karmakar-Karp's algorithms. We have developed an efficient algorithm to placing VM by neural network optimization. If we compare the exact algorithm with the developed algorithm, we will find that its approximate solutions do not differ much from the exact solutions.

Thus, the use of the algorithm provides a 12-15% profit compared to conventional methods. This is extremely effective in case of high intensity of requests.

## Acknowledgements

The research has been supported by the Russian Foundation of Basic Research (grants 16-37-60086 mol\_a\_dk, 16-07-01004 a), and the President of the Russian Federation within the grant for state support of young Russian scientists (MK-1624.2017.9).

## References

- [1] Bein D, Bein W, Venigella S. Cloud Storage and Online Bin Packing. Proc. of the 5th Intern. Symp. on Intelligent Distributed Computing 2011; 63–68.
- [2] Nagendram, S. Efficient Resource Scheduling in Data Centers using MRIS / S. Nagendram, J.V. Lakshmi, D.V. Rao // Indian J. of Computer Science and Engineering. – 2011. – Vol. 2. Issue 5. – P. 764–769.
- [3] Arzuaga, E. Quantifying load imbalance on virtualized enterprise servers / E. Arzuaga, D.R. Kaeli // Proc. of the first joint WOSP/SIPEW international conference on Performance engineering. – 2010. – P. 235–242.
- [4] Mishra, M. On theory of VM placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach / M. Mishra, A.Sahoo // IEEE International Conference Cloud Computing. – 2011. – P. 275–282.
- [5] Bolodurina I, Parfenov D. Development and research of models of organization storages based on the software-defined infrastructure. Proc. 39th International Conference on Telecommunications and Signal Processing 2016; 1–6. DOI: 10.1109/TSP.2016.7760818.
- [6] Singh A, Korupolu M, Mohapatra D. Server-storage virtualization: integration and load balancing in Data Centers. Proc. of the ACM/IEEE Conf. on Supercomputing 2012; 1–12.
- [7] Plakunov A, Kostenko V. Data center resource mapping algorithm based on the ant colony optimization. Proc. of Science and Technology Conference (Modern Networking Technologies), 2014; 1–6. DOI: 10.1109/MoNeTeC.2014.6995596.
- [8] Darabseh A, Al-Ayyoub M, Jararweh Y, Benkhelifa E, Vouk M, Rindos A. SDStorage: A Software Defined Storage Experimental Framework. Proc. of Cloud Engineering (IC2E). Tempe: IEEE Press, 2015; 341–346.

- [9] Bolodurina I, Parfenov D. Approaches to the effective use of limited computing resources in multimedia applications in the educational institutions. WCSE 2015-IPCE, 2015.
- [10] Garey M, Graham R. Bounds for multiprocessor scheduling with resource constraints. *SIAM Journal on Computing* 1975; 4(2): 187–200. DOI: 10.1137/0204015.
- [11] Arndt O, Freisleben B, Kielmann T, Thilo F. A comparative study of online scheduling algorithms for networks of workstations. *Cluster Computing* 2000; 4(2): 95–112. DOI: 10.1023/A:1019024019093.
- [12] Feitelson D, Weil A. Utilization and predictability in scheduling the IBM SP2 with backfilling. *Parallel Processing Symposium 1998*; 542–546. DOI: 10.1109/IPPS.1998.669970.
- [13] Lawson B, Smirmi E. Multiple-queue Backfilling Scheduling with Priorities and Reservations for Parallel Systems. *Lecture Notes in Computer Science* 2002; 2537: 40–47. DOI: 10.1007/3-540-36180-4\_5.
- [14] Perkovic D, Keleher P. Randomization, Speculation, and Adaptation in Batch Schedulers. *Supercomputing ACM/IEEE Conference 2000*; 7–18. DOI: 10.1109/SC.2000.10041.
- [15] Srinivasan S, Kettimuthu R. Selective Reservation Strategies for Backfill Job Scheduling. *Lecture Notes in Computer Science* 2002; 2357: 55–71. DOI: 10.1007/3-540-36180-4\_4.
- [16] Jing L, Mingze L, Gang W, Xiaoguang L, Zhongwei L, Huijun T. Global reliability evaluation for cloud storage systems with proactive fault tolerance. *Lecture Notes in Computer Science* 2015; 9531: 189–203. DOI: 10.1007/978-3-319-27140-8\_14.
- [17] OFELIA: OpenFlow in Europe. URL: <http://www.fp7-ofelia.eu> (27.02.2017).
- [18] Mambretti J, Chen J, Yeh F. Software-Defined Network Exchanges (SDXs) and Infra-structure (SDI): Emerging innovations in SDN and SDI interdomain multi-layer services and capabilities. *Proc. of Science and Technology Conference (Modern Networking Technologies) 2014*; 1–6. DOI: 10.1109/MoNeTeC.2014.6995590.
- [19] Lin T, Kang J, Bannazadeh H. Enabling SDN Applications on Software-Defined Infrastructure. *Network Operations and Management Symposium. IEEE Network Operations and Management Symposium (NOMS) 2014*; P. 1–7. DOI: 10.1109/NOMS.2014.6838226.
- [20] Ibanez G, Naous J, Rojas E, Rivera D, Schuymer T. A Small Data Center Network of ARP-Path Bridges made of Openflow Switches. *36th IEEE Conference on Local Computer Networks 2011*; 15–23.
- [21] Shimonishi H, Ochiai H, Enomoto E, Iwata A. Building Hierarchical Switch Network Using OpenFlow. *International Conference on Intelligent Networking and Collaborative Systems 2009*; 391–394. DOI: 10.1109/INCOS.2009.66.
- [22] Egilmez H. OpenQoS: An OpenFlow controller design for multimedia delivery with end-to-end quality of service over software-defined networks. *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC) 2012*; 1–6.
- [23] Kim W, Sharma P, Lee J, Banerjee S, Tourrilhes J, Lee S, Yalagandula P. Automated and Scalable QoS Control for Network Convergence. *Internet network management conference on Research on enterprise networking 2010*; 1–1.
- [24] Bolodurina I, Parfenov D. Development and research of models of organization distributed cloud computing based on the software-defined infrastructure. *Procedia Computer Science* 2017; 103: 569–576. DOI: 10.1016/j.procs.2017.01.064.

# The elaboration of numerical simulation error light pulse propagation in a waveguide of circular cross-section

A.A. Degtuarev<sup>1</sup>, A.V. Kukleva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We considered the problem of estimating the error in the solution of the wave equation recorded using infinite series Fourier-Bessel. The algorithm that adjusts the number of elements in a partial sum of infinite series, based on the assessment of the series balance. The application of the algorithm made it possible, without loss of accuracy, to substantially reduce the number of summable elements of the series in the numerical simulation of the light pulse propagation in a circular cross-section.

*Keywords:* wave equation; Fourier-Bessel series; evaluation of the residual series; numerical simulations; pulse of light; computational experiment; redundancy of partial sum components

---

## 1. Introduction

During the development of an application program for the numerical simulation of a physical process, it is important to investigate the actual error of the method used on special test cases. As test cases typically use such examples that can be resolved by an alternative method with high sufficiently precision, allowing to calculate the error of numerical method [1, 2].

This work is devoted to study the error of test value problem for the wave equation describing the propagation process of the light pulse in a waveguide in circular cross section. To elaboration the error estimate, we used remainder of the Fourier-Bessel. To check the quality of the balance assessment in the series we used the technique of computational experiment, which allows determine the degree of redundancy among several elements needed to sum to achieve the necessary precision [3].

In solving problems from numerical simulation propagation of a light pulse in a medium, various mathematical descriptions of the pulse [4-6]. In this paper, we considered two options describe different degrees of smoothness pulse function.

## 2. Mathematical model of light pulse propagation in a waveguide of circular cross-section

To describe the process of light pulse propagation we will consider the following boundary value problem:

$$\left\{ \begin{array}{l} \frac{\partial^2 E}{\partial t^2} = \frac{c^2}{n^2} \left( \frac{\partial^2 E}{\partial r^2} + \frac{1}{r} \frac{\partial E}{\partial r} + \frac{\partial^2 E}{\partial z^2} \right), \quad r \in (0; R], \quad z \in [0; L], \quad t \in [0; T]; \\ E|_{t=0} = 0, \quad r \in (0; R], \quad z \in [0; L]; \\ \left. \frac{\partial E}{\partial t} \right|_{t=0} = 0, \quad r \in (0; R], \quad z \in [0; L]; \\ E|_{z=0} = \psi(r, t), \quad r \in (0; R], \quad t \in [0; T]; \\ \left. \frac{\partial E}{\partial z} \right|_{z=L} = 0, \quad r \in (0; R], \quad t \in [0; T]; \\ E|_{r=R} = 0, \quad z \in [0; L], \quad t \in [0; T], \end{array} \right.$$

where  $E$  is a dielectric field intensity,  $c$  is a wave propagation speed in vacuum,  $n$  is a refractive index material of the waveguide,  $R$  and  $L$  is the radius and length of the waveguide,  $T$  is the duration of the dissemination process,  $\psi(r, t)$  is the function describing the pulse shape.

It is assumed when  $r = R$  an ideally conducting shell bound the waveguide, and the medium is not perturbed at the initial instant of time.

Here are the following two variants of kinetic moment:

$$\psi_1(r, t) = \varphi(r) \gamma(t) \sin \omega t, \quad \psi_2(r, t) = \varphi(r) \gamma(t) \sin \omega t \sin^2 \omega^* t,$$

where  $\gamma(t) = \begin{cases} 1, & t \in [0; t^*]; \\ 0, & t \in (t^*; T_t], \end{cases}$   $\omega = \frac{2\pi c}{\lambda}$ ,  $\omega^* = \frac{2\pi c}{\lambda j}$ ,  $t^*$  is the pulse duration at the entrance of the waveguide,  $\lambda$  is the length

of disturbing wave in vacuum,  $j$  a positive integer.  $\psi_1(r, t)$  a piecewise smooth function at variable  $t$ , because derivative has function jump in  $t = 0$ ,  $t = t^*$ . Function  $\psi_2(r, t)$  has the smoothness of a second-order variable  $t$ .

### 3. Exact solution of boundary value problem

Application of the separation variables method [5] allows getting solution of boundary-value problem for the wave equation, it can be thought of as infinite series Fourier-Bessel. For example, when describing an impulse function  $\psi_1(r, t)$  and using  $\varphi(r) = J_0(\lambda_1 r)$  the solution would be:

$$E(r, z, t) = J_0(\lambda_1 r) \left[ \sum_{k=0}^{\infty} c_k \sin(\nu_k z) \frac{\omega \sin(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \omega_k \sin(\omega t)(\hat{\omega}^2 - \omega^2)}{\omega_k (\omega_k^2 - \omega^2)} + \sin(\omega t) \right], \text{ if } t \in [0; t^*];$$

$$E(r, z, t) = J_0(\lambda_1 r) \sum_{k=0}^{\infty} \sin(\nu_k z) \left( a_1(t^*) \cos(\omega_k(t-t^*)) + \frac{a_2(t^*)}{\omega_k} \sin(\omega_k(t-t^*)) \right), \text{ if } t \in (t^*; T].$$

When writing these formulas, we use the following notation:

$$\lambda_1 = \frac{\mu_1}{R}, \quad c_k = \frac{4}{\pi(2k+1)}, \quad \nu_k = \frac{\pi(2k+1)}{2L}, \quad \omega_k = \frac{c}{n} \sqrt{\nu_k^2 + \lambda_1^2}, \quad \hat{\omega} = \frac{c}{n} \lambda_1,$$

$$a_1(t) = c_k \left[ \frac{\omega \sin(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \omega_k \sin(\omega t)(\hat{\omega}^2 - \omega^2)}{\omega_k (\omega_k^2 - \omega^2)} + \sin(\omega t) \right],$$

$$a_2(t) = c_k \omega \left[ \frac{\cos(\omega_k t)(\hat{\omega}^2 - \omega_k^2) - \cos(\omega t)(\hat{\omega}^2 - \omega^2)}{(\omega_k^2 - \omega^2)} + \cos(\omega t) \right].$$

Graph of the cross section of a pulse by a plane  $r = 1 \mu\text{m}$  in the process of its propagation in the waveguide has shown in fig. 1.

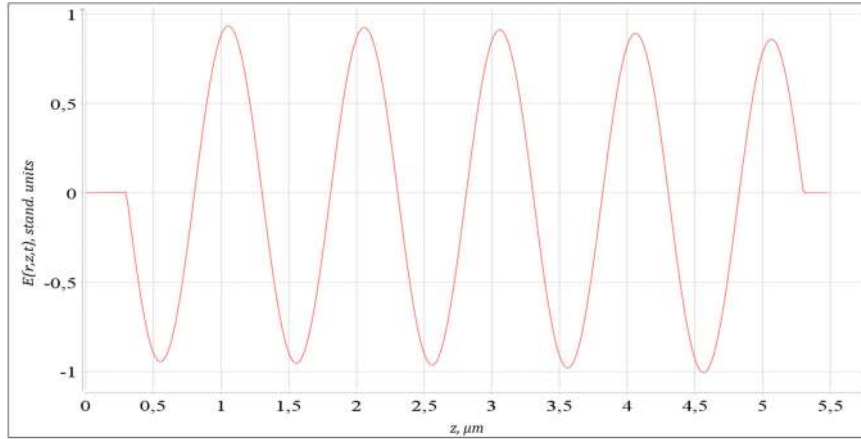


Fig. 1. Modeling the distribution piecewise smooth impulse in wave conductor, separation  $r = 1 \mu\text{m}$ .

For the case of smooth pulse described by function  $\psi_2(r, t)$  if  $\hat{\omega}^* = \frac{\omega}{10}$  and,  $\varphi(r) = J_0(\lambda_1 r)$  solution of boundary-value problem is as follows:

$$E(r, z, t) = J_0(\lambda_1 r) \left[ \sum_{k=0}^{\infty} \frac{c_k}{\omega_k} \sin(\nu_k z) (0.5a_3(t) + a_4(t) + a_5(t)) + \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right) \right], \text{ if } t \in [0; t^*];$$

$$E(r, z, t) = J_0(\lambda_1 r) \sum_{k=0}^{\infty} \sin(\nu_k z) \left( a_6(t^*) \cos(\omega_k(t-t^*)) + \frac{a_7(t^*)}{\omega_k} \sin(\omega_k(t-t^*)) \right), \text{ if } t \in (t^*; T].$$

In the last formulas, we used the following notations:

$$a_3(t) = \frac{\hat{\omega}^2 - \omega^2}{\omega^2 - \omega_k^2} (\omega \sin \omega_k t - \omega_k \sin \omega t),$$

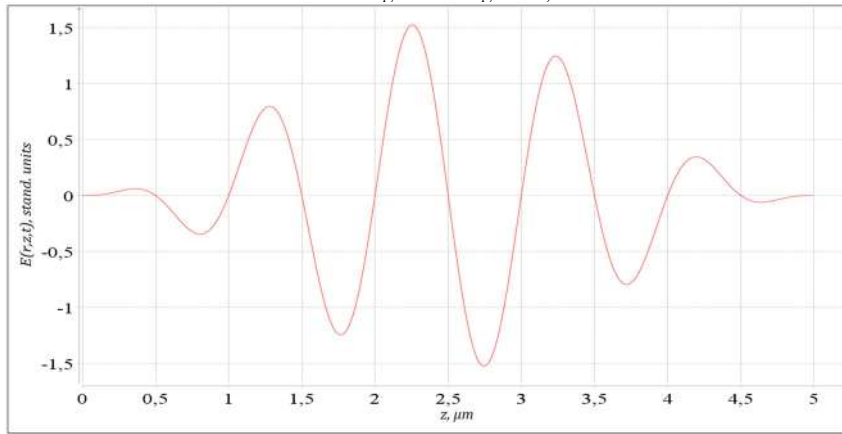
$$a_4(t) = 5 \frac{0.16\omega^2 - 0.25\hat{\omega}^2}{16\omega^2 - 25\omega_k^2} \left( 4\omega \sin \omega_k t - 5\omega_k \sin \frac{4}{5}\omega t \right), \quad a_5(t) = 5 \frac{0.36\omega^2 - 0.25\hat{\omega}^2}{36\omega^2 - 25\omega_k^2} \left( 6\omega \sin \omega_k t - 5\omega_k \sin \frac{6}{5}\omega t \right),$$

$$a_6(t) = \frac{c_k}{\omega_k} (0.5a_3(t^*) + a_4(t^*) + a_5(t^*)) + c_k \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right),$$

$$a_7(t) = \frac{c_k}{\omega_k} (0.5a_3'(t^*) + a_4'(t^*) + a_5'(t^*)) + c_k \left( \sin(\omega t) \sin^2\left(\frac{\omega t}{10}\right) \right),$$

$\mu_1$  is a root of an equation  $J_0(\mu R) = 0$ .

The process of propagating a piecewise-smooth pulse has shown in figure 2.


 Fig.2. Modeling of smooth pulse in wave conductor, separation  $r = 1 \mu m$ .

#### 4. Series truncation error control

A computer program simulating the spread of pulse truncation of the infinite series implied above.

If we can get an estimate a balance number of  $E(r, z, t) = \sum_{k=1}^N u_k(r, z, t)$  in the form of

$$|R_N| = \left| \sum_{k=N+1}^{\infty} u_k(r, z, t) \right| \leq \Phi(N),$$

where  $\Phi(N)$  is the positive monotonically decreasing function if  $N \rightarrow +\infty$ , this assessment can be used to control the truncation error. To do this, we need only find  $N(\varepsilon)$ , is the least value  $N$ , satisfy the inequality  $\Phi(N) \leq \varepsilon$ , and for approximate calculation of function values  $E(x, z, t)$  use a partial amount  $E_{N(\varepsilon)} = \sum_{n=1}^{N(\varepsilon)} e_n(x, z, t)$ .

In this case, the actual error of the calculated value of a function  $E$  at the selected point does not exceed the required level  $\varepsilon$ , that is

$$\varepsilon_{\text{fact}} = |E - E_{N(\varepsilon)}| = |R_{N(\varepsilon)}| \leq \Phi(N(\varepsilon)) \leq \varepsilon.$$

For the above two ways to specify the light pulse residues had been received by the relevant rows with the following  $t^*$  and  $w^*$ :

$$w^* = \frac{\pi c}{5\lambda}, \quad t^* = \frac{10\lambda}{c}.$$

In the case of piecewise smooth impulse, that described function  $\psi_1(r, t)$ , assessed takes the following form:

$$|E_N(r, z, t)| \leq \frac{8Ln}{\pi(2N+1)} \left( \frac{1.003}{\lambda} + \frac{2nL(\omega^2 - \hat{\omega}^2)}{\pi^2 c^2 \left( 3 + 2N - \frac{4nL}{\lambda} \right)} \right),$$

as for the case of smooth pulse, that described function  $\psi_2(r, t)$ , assessed takes the following form:

$$|E_N(r, z, t)| \leq \frac{0.16n^2 L^2 \omega^2}{c^2 \pi^3 (2N+1)^2}.$$

It should be noted that recorded higher truncation error estimates infinite series are uniform for all independent variables.

#### 5. The method of refinement of the number of summable elements of a series using a computational experiment

Proposed evaluation are not ideal because they are using strict inequalities, and also they are uniform for all independent variables. That is why using of estimates results in adding more elements than is necessary to achieve the required accuracy. In this case, it is advisable to apply a technique, which reduces the degree of redundancy terms in the partial sum, and in so doing guarantees the achievement of required accuracy [3].

Let  $N$  positive integer, satisfies the inequality  $N \leq N(\varepsilon_1)$ , where  $\varepsilon_1 < \varepsilon$ , number  $N(\varepsilon_1)$  found by the rule described in paragraph 4. Then for partial amount  $E_N$  the actual error will satisfy the inequality:

$$\varepsilon_{\text{fact}}(N) = |E - E_N| \leq |E - E_{N(\varepsilon_1)}| + |E_{N(\varepsilon_1)} - E_N|.$$

Changing  $N$  within the boundaries  $N(\varepsilon_1) \geq N \geq 1$ , find lowest value  $N(\varepsilon_2)$ , when running the inequality  $|E_{N(\varepsilon_1)} - E_N| \leq \varepsilon_2$ , where  $\varepsilon_2 = \varepsilon - \varepsilon_1$ .

For this choice  $\varepsilon_2$  and equity of the previous inequality, the actual error  $\varepsilon_{fact}(N(\varepsilon_2))$  do not exceed value  $\varepsilon$ .

Thus, to reduce the number of summands in the partial amount, we must:

- 1) Specify the number of  $\varepsilon_1 < \varepsilon$  and then find the value  $N(\varepsilon_1)$ , that the smallest value  $N$ , satisfy the inequality  $\Phi(N) \leq \varepsilon_1$ .
- 2) Changing a variable  $N$  from the value  $N(\varepsilon_1)$  downward, find the smallest of its value that satisfies the inequality  $|E_{N(\varepsilon_1)} - E_N| \leq \varepsilon_2$ . The resulting value is  $N(\varepsilon_2)$ .
- 3) Changing value with sample spacing  $\varepsilon_1$  and  $\varepsilon_2$  so, to  $\varepsilon_1 + \varepsilon_2 = \varepsilon$ , run the steps 1) and 2) again.
- 4) Of all the values  $N(\varepsilon_2)$ , obtained in step 3), select the smallest.

As a result of the use of this algorithm, it can be expected that the number of summable elements  $N(\varepsilon_2)$  in the partial sum will be reduced significantly as compared with the number of  $N(\varepsilon)$  while maintaining safeguards for accuracy, i.e.

$\varepsilon_{fact}(N(\varepsilon_2)) \leq \varepsilon$ . In tables 1 and 2 are the results of computational experiments, aimed at reducing the number of summands in partial amounts. The calculations have been carried out with the following parameters:

$$\lambda = 1 \mu m, n = 1, L = 7 \mu m, R = 5 \mu m, c = 3 \cdot 10^{14} \mu m / s, r = 1 \mu m, z = 1 \mu m, t = \frac{tc}{n} \mu m.$$

Asked value  $\varepsilon$  in increments of the maximum value of the amplitude of the wave.

Table 1. The dependence of the summands number  $N(\varepsilon)$  and  $N(\varepsilon_2)$  of coordinate  $t$  with different values  $\varepsilon$  for piecewise smooth impulse.

$\varepsilon$	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$
$N(\varepsilon)$	131	1019	9844	98079	980434
$t, \mu m$	$N(\varepsilon_2)$				
0.9	13	48	231	3116	9906
0.999	37	306	1241	6774	26632
0.99999	37	312	3072	35599	126836
1	37	312	3075	37713	377122
1.00001	37	312	3072	35599	126836
$\varepsilon$	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$
1.001	37	306	1241	6774	26633
1.1	16	68	320	3119	9906
1.7	19	34	124	1286	4086
2.5	15	32	96	928	984
4	13	25	66	612	643
5.1	16	30	75	649	1436
5.9	17	28	324	3116	3258
5.999	47	355	1262	6422	9906
5.99999	47	466	4672	35599	26632
6	47	467	4672	37713	126836
6.00001	47	465	4671	35599	377122
6.001	47	383	1461	6423	126836
6.1	17	30	360	3119	26634

Table 2. The dependence of the summands number  $N(\varepsilon)$  and  $N(\varepsilon_2)$  of coordinate  $t$  with different values  $\varepsilon$  for smooth pulse.

$\varepsilon$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$	$10^{-6}$
$N(\varepsilon)$	21	36	113	357	1128
$t, \mu m$	$N(\varepsilon_2)$				
0.9	15	18	28	62	132
0.999	15	18	28	61	136
0.99999	14	17	28	62	126
1	15	18	27	67	141
1.000001	10	21	37	91	186
1.001	10	22	42	101	211
1.1	15	17	33	61	132
1.7	10	17	37	81	181
2.5	13	15	26	67	146
4	15	22	46	101	216



From the table it can be seen that the number of summands, using uniform assessments for the respective series truncation allows you to get only the rough partial sums of lengths. These values are repeatedly exceed the values obtained from the application of the above algorithm. As can be seen from table 1, to calculate the tension of the electric field in the foreground and background areas of wave fronts requires a much larger number of terms, for example, in the range  $1.7 \mu\text{m} \leq t \leq 5.1 \mu\text{m}$  order enough 4086 parts to achieve precision  $10^{-5}$ , while in the range  $0.9 \mu\text{m} \leq t \leq 1.1 \mu\text{m}$  we want 377122 parts. This increase in the number of summands is a consequence of the weak function breaks  $\psi_1(r, t)$ , significantly slowing down the convergence of series. For the case of smooth pulse, function description  $\psi_2(r, t)$  the uneven distribution of values  $N(\varepsilon_2)$  for different  $t$  turns out to be negligible.

## 6. Conclusion

Developed and implemented programmatically algorithm provides adjustment of the partial sums length of infinite series, obtained in the course of solving boundary value problem for the wave equation. For practical application of the algorithm, it is of fundamental importance to first obtain an upper estimate for the remainder of the Fourier series that determines the solution of the boundary value problem.

The application of developed algorithm for specific series that describe the distribution of momentum in circular waveguide section allowed multiple times (from 3 up to 1500 times and more for Piecewise-smooth momentum and from 2 to 5 times for the case of smooth pulse) to reduce length of the partial sums of the series.

## References

- [1] Feng X. A high-order compact scheme for the one-dimensional Helmholtz equation with a discontinuous coefficient. *International Journal of Computer Mathematics* 2012; 1: 1–7.
- [2] Degtyarev AA, Kozlova ES. Investigation of accuracy of numerical solution of the one-way helmholtz equation by method of computational experiment. *Computer Optics* 2012; 36(1): 36–45.
- [3] Degtyarev AA, Praslova MO. Estimation of the error in the solution of the wave equation in the problem of modeling the propagation of a light pulse in a planar waveguide. *Proceedings of the International Conference Information Technology and Nanotechnology*. Samara, Samara National Research University, 2016; 852–859.
- [4] Kozlova ES, Kotlyar VV. Simulation of ultrafast 2d light pulse. *Computer Optics* 2012; 36(2): 158–164.
- [5] Kotlyar VV, Kozlova ES. Simulations of sommerfeld and brillouin precursors in the medium with frequency dispersion using numerical method of solving wave equations. *Computer Optics* 2013; 37(2): 146–154.
- [6] Fuchs U, Zeitner U, Tunnermann A. Ultra-short pulse propagation in complex optical system. *Optics Express* 2005; 13(10): 3852–3861.
- [7] Tikhonov AN, Samarskiy A.A. *Equations of mathematical physics*. M.: Nauka, 1972; 736 p.

# The calculation of the spatial spectrum of multidimensional fractals using the fast Fourier transform

O.A. Mossoulina<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The Fast Fourier Transform was applied to spatial spectrum modeling of a one-dimensional fractal (Cantor set), a two-dimensional fractal (Sierpinski carpet), and a three-dimensional fractal (Menger sponge). A spectrum is developed for different levels. The spatial spectrum was also obtained and modeled for various filling parameters. The ParaView software package was used for 3D modeling.

*Keywords:* cantor set; Sierpinski carpet; Sierpinski carpet; fast Fourier transform; 3D modeling

---

## 1. Introduction

Many natural phenomena have distinctive features, which are often associated with fractal structures. Visually, fractals represent a geometric figure, replication of which is exactly the same at every scale [1]. This ability is called self-similarity. Fractals are interesting because of widespread presence in natural formations [1-3]. In this case, natural fractals are called "statistical", and artificial "exact". Statistical fractals can be observed in various polymers, biological structures, electrical circuits, galactic clusters and fluctuations in exchange prices [4]. Exact fractals are generated from mathematical approach [5]. Can these precise mathematical abstractions be found in physical reality? Yes, it is optical fractals [3]. This concept includes "diffractals" (diffraction pattern on fractal lattice) [6, 7], eigen modes of unstable resonators [8], distributions in nonlinear optics [3, 9].

Particularly interesting can be the coincidence of certain properties of "accurate" and "statistical" fractals [10], such as aerosols, smoke, moire [11-13], which is very important applied to optical signal transmission through a heterogeneous or random medium [14-17]. Examination of diffraction on fractal lattice [6, 7, 18-20] can solve other important problems - the formation of periodically self-reproducing fields [21-26], the creation of multi focus [27-30] or specified longitudinal distributions [31-33], and in achromatic depicting systems [34-37].

One of the most important characteristics of fractals is the spatial spectrum [38-41], which are also important in the analysis of crystal structures [42-44]. Taking into account possible multidimensionality of fractals, the calculation of the spatial spectrum can lead to problems associated with computational complexity, which depends on the technical capabilities of modern computers. The solution to the problem can be the usage of the fast calculation algorithm. Within this paper, the fast transformation is used to develop the spatial spectrum of multidimensional fractals with different characteristics.

## 2. The calculation of the spatial spectrum of multidimensional fractals

The first stage of the modeling is the implementation of a one-dimensional case. We take a unit segment  $E_0 = [0,1]$ . The next segment is formed according to the rule  $E_1 = [0,a] \cup [b,1]$ , where  $a$  and  $b$  are the fractal parameters specified in the range of  $(0,1)$ , whereby  $a < b$  and  $a + b = 1$ . We continue until reaching the desired order of the fractal. The intersection of all segments will be a simulated fractal.

$$E = \bigcap_{i=1}^n E_i, \quad (1)$$

where  $n$  is the order of the fractal.

If the parameters are set to  $a = \frac{1}{3}$  and  $b = \frac{2}{3}$ , then we get the Cantor set.

For programming is used a vector consisting of units, which is successively filled with zeros, according to input parameters and order.

To simulation for two-dimensional case, we used a similar implementation with some corrections. We took the unit square  $E_0 = [0,1] \times [0,1]$  and the next one will take form of  $E_1 = ([0,a_1] \cup [b_1,1]) \times ([0,a_2] \cup [b_2,1])$ , where  $a_1, a_2, b_1$  and  $b_2$  are fractal parameters specified in the range of  $(0,1)$ , whereby  $a_1 < b_1, a_2 < b_2$  and  $a_1 + b_1 = 1, a_2 + b_2 = 1$ . The simulated fractal can be

found by the previously applied for the one-dimensional case formula (1). If we set the parameters  $a_1 = \frac{1}{3}, a_2 = \frac{1}{3}, b_1 = \frac{2}{3}$  and

$b_2 = \frac{2}{3}$  we get a fractal called the Sierpinski carpet (Fig. 1).

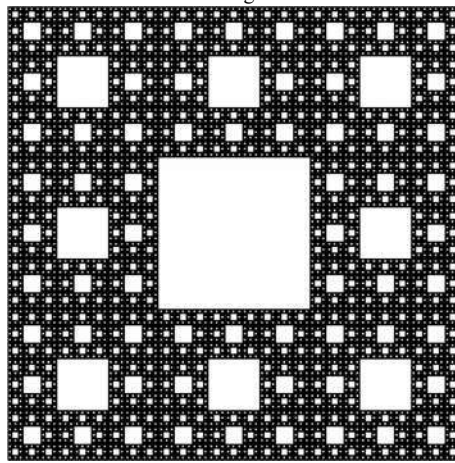


Fig. 1. Fractal (Sierpinsky carpet).

The three-dimensional case is implemented reciprocally to the two-dimensional case. The unit cubes  $E_0 = [0,1] \times [0,1] \times [0,1]$  and  $E_1 = ([0, a_1] \cup [b_1, 1]) \times ([0, a_2] \cup [b_2, 1]) \times ([0, a_3] \cup [b_3, 1])$  was taken, whereby  $a_1, a_2, a_3, b_1, b_2$  and  $b_3$  are fractal parameters specified in the range of  $(0,1)$ , whereby  $a_1 < b_1, a_2 < b_2, a_3 < b_3$  and  $a_1 + b_1 = 1, a_2 + b_2 = 1, a_3 + b_3 = 1$ . If we set the parameters  $a_1 = \frac{1}{3}, a_2 = \frac{1}{3}, a_3 = \frac{1}{3}, b_1 = \frac{2}{3}, b_2 = \frac{2}{3}$  and  $b_3 = \frac{2}{3}$  we get a three-dimensional fractal called Menger sponge (Fig. 2 a), the boundary section of which is a Sierpinsky carpet.

If we set the parameters  $a_1 = \frac{1}{3}, a_2 = \frac{3}{8}, a_3 = \frac{1}{3}, b_1 = \frac{2}{3}, b_2 = \frac{5}{8}$  and  $b_3 = \frac{2}{3}$  we get a scalable three-dimensional fractal (Fig. 3 a).

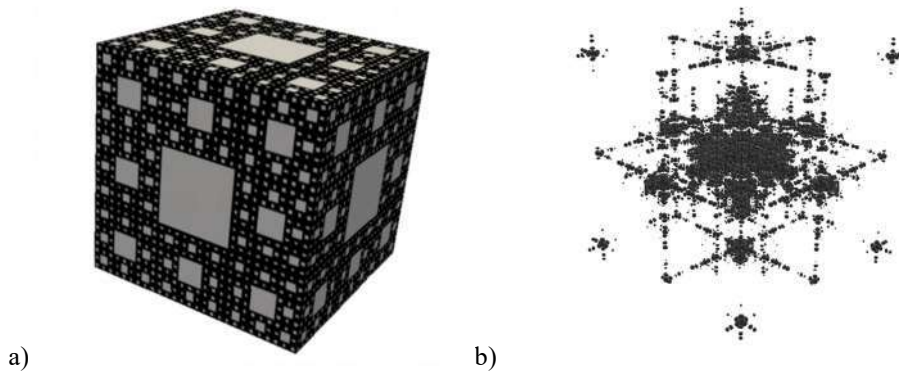


Fig. 2. a) Three-dimensional fractal (Menger sponge), b) the spatial spectrum of a three-dimensional fractal.

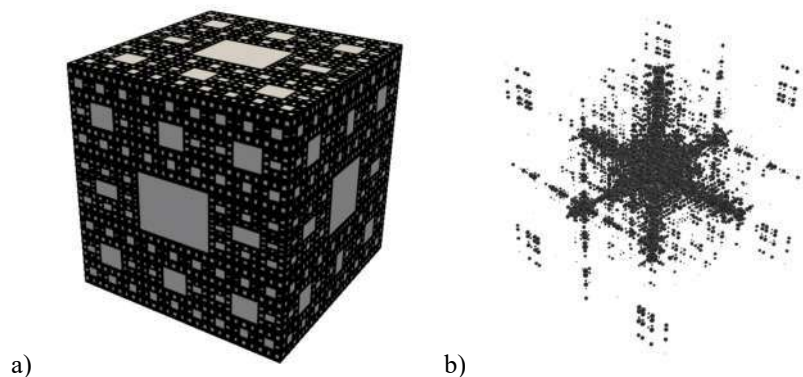


Fig. 3. a) Three-dimensional scalable fractal (Menger sponge), b) the spatial spectrum of a three-dimensional scalable fractal.

The Fast Fourier Transform was used to generate the spatial spectrum.

$$F(\mathbf{u}) = \mathfrak{F}[f(\mathbf{x})](\mathbf{u}) = \int_{R^n} f(\mathbf{x}) \exp(-2\pi i \mathbf{x} \mathbf{u}) d^n \mathbf{x}, \tag{2}$$

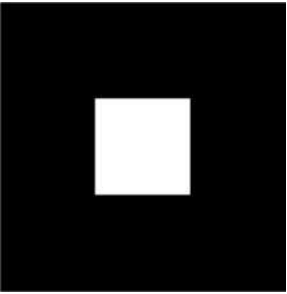
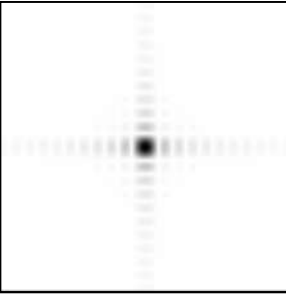
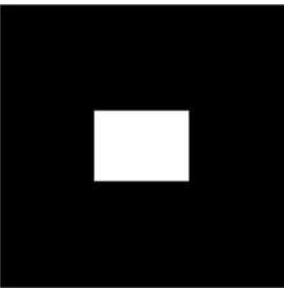
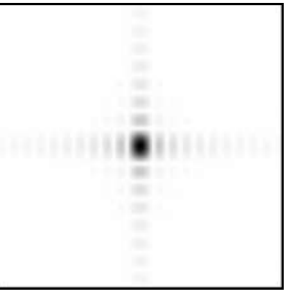
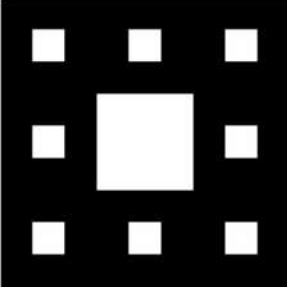
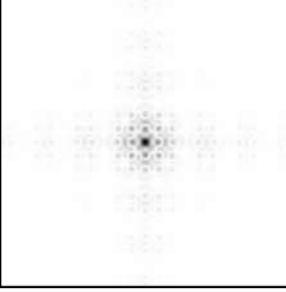
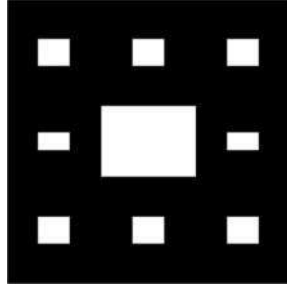
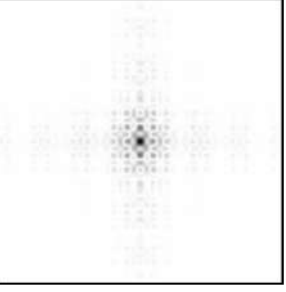
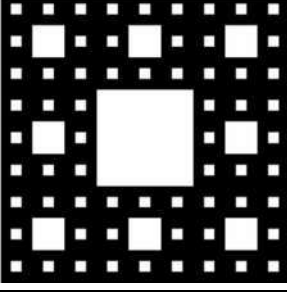
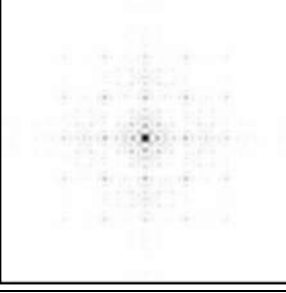
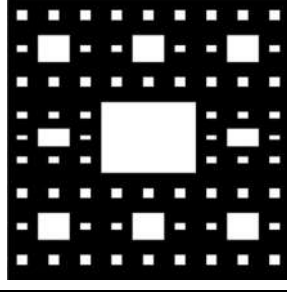
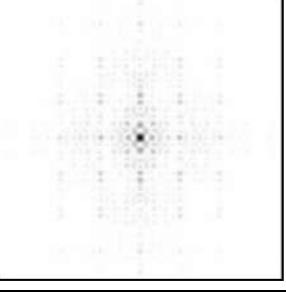
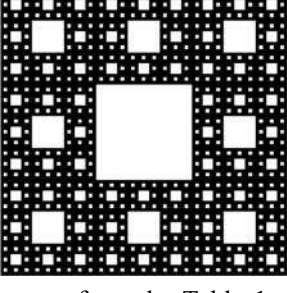
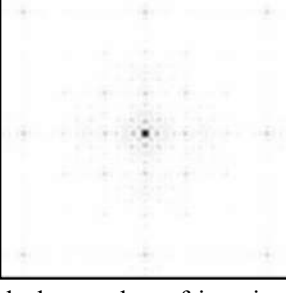
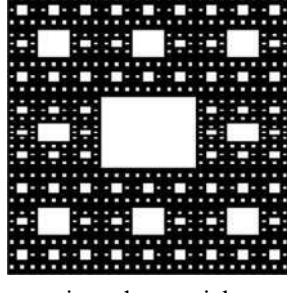
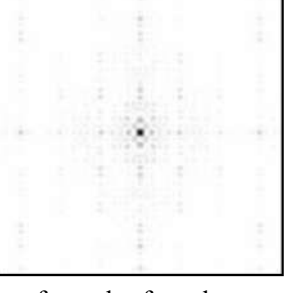
whereby  $f(\mathbf{x})$  is the input function specified as a vector, which is a binary representation of the fractal,

$F(\mathbf{u})$  is the output function,

$\mathfrak{F}[\cdot]$  is the Fourier transform operator.

The spatial spectrum was obtained from a two-dimensional fractal structure (Sierpinski carpet). The results for the different number of iterations and scale are presented in Table 1.

Table 1. Variability of the spectrum in relations to the number of iterations and scale.

Number of iterations	Fractal	Spectrum	Fractal	Spectrum
2				
3				
4				
5				

As can be seen from the Table 1, with the number of iteration increasing, the spatial spectrum from the fractal structure becomes more complex and the energy at higher frequencies increases. However, the pattern of the spectrum maintains a regular structure, which is also characteristic of crystalline structures [42-44].

### 3. Conclusion

As a result of the work, the spatial spectrum was calculated and visualized from a two-dimensional (Sierpinski carpet) and a three-dimensional (Menger sponge) fractal structure using the Fast Fourier Transform algorithm.

### Acknowledgements

The work was supported by the Ministry of Education and Science of the Russian Federation.

### References

[1] Mandelbrot BB. The Fractal Geometry of Nature. New York: W.H. Freeman and Company, 1983; 468 p.  
 [2] Barnsley M. Fractals Everywhere. Academic. Boston: Mass, 1988; 534 p.  
 [3] Segev M, Soljagic M, Dudley JM. Fractal optics and beyond. Nature Photonics 2012; 6(4): 209–210.  
 [4] Addison PS. Fractals and Chaos. An illustrated course IOP, 1997; 256 p.  
 [5] Feder J. Fractals. New York: Springer Science & Business Media, 2013; 282 p.  
 [6] Berry MV. Diffraction. J. Phys. A: Math. Gen. 1979; 12: 781–797.

- [7] Berry MV, Klein S. Integer, fractional and fractal Talbot effects. *J. Mod. Opt.* 1996; 43(10): 2139–2164.
- [8] Karman GP, McDonald GS, New GHC, Woerdman JP. Laser optics: Fractal modes in unstable resonators. *Nature* 1999; 402(6758): 138 p.
- [9] Gabitov IR, Manakov SV. Propagation of Ultrashort Optical Pulses in Degenerate Laser Amplifiers. *Phys. Rev. Lett.* 1983; 50(7): 495 p.
- [10] Peitgen HO, Jurgens H, Saupe D. *Chaos and fractals: new frontiers of science*. 2nd edn. New York: Springer, 2004; 864 p.
- [11] Forrest S, Witten TA. Long-range correlations in smoke-particle aggregates. *J. Phys. A*. 1979; 12(5): L109 p.
- [12] Berry MV, Percival IC. Optics of fractal clusters such as a smoke. *Journal of Modern Optics* 1986; 33(5): 577–591.
- [13] Oster G, Wasserman M, Zwerling C. Theoretical interpretation of moire patterns. *J. Opt. Soc. Am.* 1964; 54(2): 169–175.
- [14] Striletz AS, Khonina SN. Matching and study of methods based on differential and integral operators of laser propagation in media with small inhomogeneities. *Computer Optics* 2008; 32(1): 33–38.
- [15] Khonina SN, Golub I. Creating order with the help of randomness: generating transversely random, longitudinally invariant vector optical fields. *Optics Letters* 2015; 40(17): 4070–4073.
- [16] Soifer VA, Korotkova O, Khonina SN, Shchepakina EA. Vortex beams in turbulent media. *Computer Optics* 2016; 40(5): 605–624. DOI: 10.18287/2412-6179-2016-40-5-605-624.
- [17] Porfirev AP, Kirilenko MS, Khonina SN, Skidanov RV, Soifer VA. Study of propagation of vortex beams in aerosol optical medium. *Applied Optics* 2017; 56(11): E8–E15.
- [18] Sakurada Y, Uozumi J, Asakura T. Fresnel diffraction by 1-D regular fractals. *Pure Appl* 1992; 1: 29–35.
- [19] Jaggard AD, Jaggard DL. Cantor ring diffractals. *Opt. Commun.* 1998; 158(1): 141–148.
- [20] Szwaykowski P. Self-imaging in polar coordinates. *J. Opt. Soc. Am. A* 1988; 5(2): 185–191.
- [21] Hou B, Xu G, Wen W, Wong GKL. Diffraction by an optical fractal grating. *Appl. Phys. Lett.* 2004; 85(25): 6125–6127.
- [22] Mendez DC, Lehman M. Talbot effect with Cantor transmittances. *Optik* 2004; 115(10): 439–442.
- [23] Kotlyar VV, Soifer VA, Khonina SN. An algorithm for the generation of laser beams with longitudinal periodicity: rotating images. *Journal of Modern Optics* 1997; 44(7): 1409–1416.
- [24] Khonina SN, Kotlyar VV, Soifer VA. Light beams with periodic properties. *Methods for Computer Design of Diffractive Optical Elements*. Ed. Soifer VA. New York: Wiley & Sons, Inc., 2002; 535–605.
- [25] Almazov AA, Khonina SN. Periodic self-reproduction of multi-mode laser beams in graded-index optical fibers. *Optical Memory and Neural Networks* 2004; 13(1): 63–70.
- [26] Khonina SN, Volotovskiy SG. Self-reproduction of multimode laser fields in weakly guiding stepped-index fibers. *Optical Memory and Neural Networks* 2007; 16(3): 167–177.
- [27] Saavedra G, Furlan WD, Monsoriu JA. Fractal zone plates. *Optics Letters* 2003; 28(12): 971–973.
- [28] Mihailescu M, Preda AM, Sobetkii A, Petcu AC. Fractal-like diffractive arrangement with multiple focal points. *Opto-electronics review* 2009; 17(4): 330–337.
- [29] Kotlyar VV, Khonina SN, Soifer VA. Diffraction computation of focusator into longitudinal segment and multifocal lens. *Proceedings of SPIE* 1993; 1780: 263–272.
- [30] Soifer VA, Doskolovich LL, Kazanskiy NL. Multifocal diffractive elements. *Optical Engineering* 1994; 33(11): 3610–3615.
- [31] Khonina SN, Kotlyar VV, Soifer VA. Calculation of the focusators into a longitudinal linesegment and study of a focal area. *Journal of Modern Optics* 1993; 40(5): 761–769.
- [32] Khonina SN, Ustinov AV. Design lenses forming paraxial longitudinal distribution according to their spatial spectra. *Computer Optics* 2013; 37(2): 193–202.
- [33] Khonina SN, Ustinov AV. Lenses to form a longitudinal distribution matched with special functions. *Optics Communications* 2015; 341: 114–121.
- [34] Wang YX, Yun WB, Jacobsen C. Achromatic Fresnel optics for wideband extreme-ultraviolet and X-ray imaging. *Nature* 2003; 424(6944): 50–53.
- [35] Furlan WD, Saavedra G, Monsoriu JA. White-light imaging with fractal zone plates. *Opt. Lett.* 2007; 32(15): 2109–2111.
- [36] Andersen G, Tullson D. Broadband antihole photon sieve telescope. *Phys. R. A* 2007; 46(18): 3706–3708.
- [37] Khonina SN, Ustinov AV, Skidanov RV, Morozov AA. Comparative study of the spectral characteristics of aspheric lenses. *Computer Optics* 2015; 39(3): 363–369. DOI: – 10.18287/0134-2452-2015-39-3-363-369.
- [38] Allain C, Cloitre M. Spatial spectrum of a general family of self-similar arrays. *Phys. Rev.* 1987; 36(12): 5751–5757.
- [39] Uozumi J, Kimura H, Asakura T. Fraunhofer diffraction by Koch fractals. *J. Mod. Opt.* 1990; 37(6): 1011–1031.
- [40] Zunino L, Garavaglia M. Fraunhofer diffraction by cantor fractals with variable lacunarity. *J. Mod. Opt.* 2003; 50(5): 717–727.
- [41] Horvath P, Smid P, Vaskova I, Hrabovsky M. Koch fractals in physical optics and their Fraunhofer diffraction patterns. *Optik* 2010; 121(2): 206–213.
- [42] Dal Negro L, Boriskina SV. Deterministic aperiodic nanostructures for photonics and plasmonics applications. *Laser Photonics Rev.* 2011; 1–41.
- [43] Kharitonov SI, Volotovskiy SG, Khonina SN, Kazanskiy NL. A differential method for calculating x-ray diffraction by crystals: the scalar theory. *Computer Optics* 2015; 39(4): 469–479.
- [44] Kharitonov SI, Kazanskiy NL, Volotovskiy SG, Khonina SN. Calculating x-ray diffraction on crystals by means of the differential method. *International Society for Optics and Photonics* 2016; 10 p.

# Study of a singularly perturbed tuberculosis model

E. Tropkina<sup>1</sup>, E. Shchepakina<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

## Abstract

A detailed analysis of the dynamic model of the tuberculosis epidemic was conducted. It is shown that the dynamic model contains several time scales and can be represented in a singularly perturbed form. With help of the integral manifolds theory and the reduction principle, the reduction of the modeling system was justified and carried out. This approach allow us to replace the original system by another system of a lower order on an integral manifold whose dimension is equal to that of the slow subsystem and which, at the same time, preserves the essential properties of the original system. The conditions for the stabilization of the epidemiology based on the selection of necessary treatment and preventive measures are determined.

*Keywords:* singular perturbations; integral manifold; order reduction; stability; epidemiology; tuberculosis

## 1. Introduction

It is well known that tuberculosis is a deadly disease, the fight against which is still relevant today. In 1882, Robert Koch, along with the discovery of the tubercle bacillus, found that this disease is transmitted by aerogenic means [1]. Consequently, people who frequently contact individuals with an active form of tuberculosis (the infectious stage of the disease) have a much higher risk of infection. Majority of infected people remain latent carriers throughout their life. The average duration of the latent period (the period of latent infection) ranges from several months to dozens of years. However, the risk of progression to the active form of tuberculosis increases dramatically in the presence of concomitant infections that weaken the immune system. In the absence of treatment for tuberculosis of respiratory organs, mortality is about 50%.

Currently, about 3 million people die from tuberculosis every year worldwide [2]. But in most cases, tuberculosis is curable. The current methods of treating this disease require long-term courses of treatment (from six months to several years), the violation of which often leads to the return of the disease and the development of drug resistance.

Treatment of drug-resistant tuberculosis is much more dangerous for the patient, less successful and more costly than treatment of the disease caused by common strains of the pathogen. All this shows the necessity of developing a new approach of identifying and treating patients with tuberculosis, as well as timely and effective treatment.

One of the most effective methods for solving such problems is the construction and study of a mathematical model describing the processes of infection spread in the population, the development of the disease and the impact of anti-tuberculosis measures. Based on this study we can determine effective measures to combat this dangerous phenomenon.

In the present work, a detailed analysis of the dynamic model of the tuberculosis epidemic based on the cluster approach has been carried out. The presence of several time scales made it possible to apply the geometric theory of singular perturbations for its qualitative study. This approach made it possible to determine the conditions for the stabilization of epidemiological conditions based on the selection of the necessary treatment and preventive measures.

## 2. Cluster model of the tuberculosis epidemic

Mycobacterium drops of tuberculosis get into the air when coughing or sneezing infected people. Tuberculosis bacillus, spreading by such drops, lives in the air for a short period of time (about two hours) and, therefore, it is believed that casual contact with persons with active tuberculosis (infected individuals) rarely lead to the spread of the disease, and that most relapses are the result of prolonged and close contacts with primary carriers of infection. Latently infected individuals become infectious after some, usually long, time period. This period of transition to an active form of infection is called latent. Latent periods vary from several months to dozens of years. Most infected individuals never go to the active form of tuberculosis. On the other hand, the average length of the infection period is relatively short (several months). This indicator is decreasing in countries with affordable treatment.

A common scheme for analyzing the spread of the tuberculosis epidemic is based on the division of the population into certain classes. Consider the basic mathematical model of the spread of tuberculosis [2]:

$$\begin{cases} \frac{dS}{dt} = \Lambda - \mu \cdot S + \beta_1 \cdot S \cdot \frac{I}{N}, \\ \frac{dE}{dt} = \beta_1 \cdot S \cdot \frac{I}{N} - (\mu + k + r_1) \cdot E + \beta_2 \cdot E \cdot \frac{I}{N} + \beta_3 \cdot R \cdot \frac{I}{N}, \\ \frac{dI}{dt} = k \cdot E + \beta_2 \cdot E \cdot \frac{I}{N} - (\mu + d + r_2) \cdot I, \\ \frac{dR}{dt} = r_2 \cdot I + r_1 \cdot E - \mu \cdot R - \beta_3 \cdot R \cdot \frac{I}{N}. \end{cases} \quad (1)$$

Model (1) was derived under assumption that the total population is divided into four epidemiological classes: susceptible ( $S$ ), i.e., uninfected, but susceptible to infection individuals, those in whose body the pathogens of tuberculosis have not yet penetrated; carriers of latent infection ( $E$ ), i.e., individuals in whose body tuberculosis pathogens are present, in equilibrium with the immune system, such individuals are characterized by the absence of any external manifestations of the disease; infected individuals ( $I$ ) with clinical manifestations of tuberculosis caused by sufficiently extensive tissue damage as a result of the activity of mycobacteria in their organisms; recovery individuals ( $R$ ) who have completed treatment and recovered from the disease. The parameter  $\Lambda$  reflects the influx of young people into the model population. The parameters  $\beta_1, \beta_2, \beta_3$  are the transmission rates of tuberculosis infection for the respective classes;  $\mu$  is natural mortality rate;  $k$  is the tuberculosis progression rate;  $d$  is the tuberculosis induced mortality rate;  $r_1, r_2$  denote the treatment rates for latent class and infectious class, respectively;  $N = S + E + I + R$  is the total population size. The mixing of classes in this model is homogeneous, that is, there are no assumed differences between individuals while tuberculosis transmission depends on the rate of infection.

The basic reproduction number, one of the most important characteristics in mathematical biology, defined as the average number of secondary cases produced by typical infected individuals, mainly in a susceptible population, is described by

$$\mathfrak{R}_0^{HM} = \frac{\beta_1}{(\mu + d + r_2)} \cdot \frac{k}{(k + \mu + r_1)} = Q_0 \cdot \frac{k}{(k + \mu + r_1)},$$

$$Q_0 = \frac{\beta_1}{\gamma}, \quad \gamma = \mu + d + r_2,$$

where  $Q_0$  is the number of secondary latent infected individuals produced by a typical infected individual during the mean infectious period  $1/\gamma$ ,  $f = k/(k + \mu + r_1)$  shows the probability of survival during the transition from the latent to the active infectious stage.

It is assumed that only individuals who have frequent and prolonged interactions with infected people have a high risk of tuberculosis infection. Newly infected individuals activate clusters (groups of individuals who come into regular and close contact with people in an active form of tuberculosis), increasing the risk of tuberculosis infection for susceptible individuals of each cluster [2].

Let us make a number of variables changes. We set a constant  $n$  be the average size of the cluster; the risk of infection with tuberculosis in the cluster will be determined by the parameter  $\beta$ . Further, the population of uninfected people in the cluster will be given by  $N_1(t) = n \cdot I(t)$ , where  $N_1$  includes two subpopulations, namely, susceptible  $S_1$  and latent infected  $E_1$ , i.e.,

$$N_1(t) = n \cdot I(t) = S_1(t) + E_1(t).$$

The population of persons who do not belong to the cluster at the time  $t$  is denoted as  $N_2$ . This population consists only of the sensitive ( $S_2$ ) and the latent infected ( $E_2$ ) individuals. The subpopulations of the sick are not included in this model for the sake of simplicity. We assume that  $n \cdot k \cdot E_2 (S_2 / N_2)$  individuals go to the class  $S_1$  per unit of time, while the individuals  $n \cdot k \cdot E_2 (E_2 / N_2)$  go to the class  $E_1$  per unit time. In addition, since infected individuals are cured or die (with a speed  $\gamma \cdot I$ ), that  $\gamma \cdot I$  is the rate at which clusters become inactive (or die). We suppose that  $n \cdot \gamma \cdot I \cdot S_1 / N$  individuals returns to the class  $S_2$  and  $n \cdot \gamma \cdot I \cdot E_1 / N$  returns to the class  $E_2$  for the unit of time, respectively. Assuming a low level of distribution of individuals with an active form of tuberculosis, we consider the case  $N_1 \ll N_2$ , hence we can neglect the birth rate in the population.

The above assumptions lead (1) to the following basic cluster model [2]:

$$\begin{cases} \frac{dS_1}{dt} = -(\beta + \gamma) \cdot S_1 + \frac{S_2}{N_2} \cdot n \cdot k \cdot E_2, \\ \frac{dE_1}{dt} = \beta \cdot S_1 - \gamma \cdot E_1 + \frac{E_2}{N_2} \cdot n \cdot k \cdot E_2, \\ \frac{dI}{dt} = k \cdot E_2 - \gamma \cdot I, \\ \frac{dS_2}{dt} = \Lambda - \mu \cdot S_2 + \gamma \cdot S_1 - \frac{S_2}{N_2} \cdot n \cdot k \cdot E_2, \\ \frac{dE_2}{dt} = \gamma \cdot E_1 - (\mu + k) \cdot E_2 - \frac{E_2}{N_2} \cdot n \cdot k \cdot E_2. \end{cases} \quad (2)$$

The basic reproduction number for system (2) is:

$$\mathfrak{R}_0^c = \frac{\beta \cdot n}{(\beta + \gamma)} \cdot \frac{k}{(\mu + k)} = Q_0 \cdot f.$$

Hence,  $Q_0 = \beta \cdot n / (\beta + \gamma)$  is the expected number of infections produced by one infected individual in his cluster. Only a part  $f = k / (\mu + k)$  among the infected individuals will survive during the latent period.

### 3. Dimensionless model

Disease and dynamics of populations have characteristic time scales. Dynamics of influenza in humans is "super fast" at the individual and community levels in comparison with the dynamics of the carriers of infection (people). This is because the

average life expectancy of an infected person is approximately 4000-8000 times the average duration of an influenza infection and, therefore, the largest outbreaks of influenza occur in local communities before any significant demographic change can occur (several months). Consequently, when studying the dynamics of the flu epidemic, two typical time scales are often used: the time scale of the disease and the life span of the carrier of the infection [3].

Tuberculosis is usually described as a slow disease due to its long and varied distribution of the latent period and due to the short and relatively narrow distribution of its infectious period [1]. Most latently infected with tuberculosis do not become actively infected, i.e., there is no transition of latent form to active. Some become actively infected during a five-year period, while others become active only after a longer period of time (perhaps decades). On the other hand, infected individuals remain so for relatively short periods of time, in part because of the use of antibiotics (an average of six months). Since secondary infections are formed from infected individuals, individuals with an active form of tuberculosis have a relatively short period for possible infection of other people. Consequently, the infection of tuberculosis-susceptible individuals occurs on the same time scale as the recovery of persons with active tuberculosis. The disease progresses, moving from the latent stage to the active one. This occurs on a time scale that is of the same order as the average lifespan of the carrier of the infection (human).

Tuberculosis can be acquired at random (individual level), that is, as a result of accidental contacts, or through members of an epidemiologically active cluster that includes at least one actively infected person. The choice of these levels of distribution is not accidental, it is associated with the observed statistical data of the spread of tuberculosis. Since in the original system there is no clear separation into fast and slow variables, but the process speeds have different orders, it is necessary to bring the system (2) to a dimensionless form. To do this, we introduce the following dimensionless variables and parameters:

$$\tau = k \cdot t, \quad dt = \frac{d\tau}{k}, \quad x_1 = \frac{S_2}{\Omega}, \quad x_2 = \frac{E_2}{\Omega}, \quad y_1 = \frac{\beta + \gamma}{k} \cdot \frac{S_1}{\Omega},$$

$$y_2 = \frac{\beta + \gamma}{k} \cdot \frac{E_1}{\Omega}, \quad y_3 = \frac{\beta + \gamma}{k} \cdot \frac{I}{\Omega}, \quad \varepsilon = \frac{k}{\beta + \gamma}, \quad m = \frac{\beta}{\beta + \gamma}, \quad B = \frac{\mu}{k},$$

where  $\varepsilon$  is a small positive parameter. With new variables system (2) has a form:

$$\begin{cases} \frac{dx_1}{d\tau} = B \cdot (1 - x_1) + (1 - m) \cdot y_1 - n \cdot \frac{x_1 \cdot x_2}{x_1 + x_2}, \\ \frac{dx_2}{d\tau} = (1 - m) \cdot y_2 - (1 + B) \cdot x_2 - n \cdot \frac{x_2^2}{x_1 + x_2}, \\ \varepsilon \cdot \frac{dy_1}{d\tau} = -y_1 + n \cdot \frac{x_1 \cdot x_2}{x_1 + x_2}, \\ \varepsilon \cdot \frac{dy_2}{d\tau} = m \cdot y_1 - (1 - m) \cdot y_2 + n \cdot \frac{x_2^2}{x_1 + x_2}, \\ \varepsilon \cdot \frac{dy_3}{d\tau} = x_2 - (1 - m) \cdot y_3, \end{cases} \quad (3)$$

where  $y_1$ ,  $y_2$  and  $y_3$  are the fast variables, while  $x_1$  and  $x_2$  are the slow variables;  $x_1$  corresponds to a population of sensitive individuals not belonging to the cluster;  $x_2$  is a population of latently infected individuals not belonging to the cluster;  $y_1$  reflects a population of susceptible people in the cluster;  $y_2$  is a latently infected population;  $y_3$  is the population of infected individuals belonging to the cluster. The generated system [4, 5] for (3) is:

$$\begin{cases} \frac{dx_1}{d\tau} = B \cdot (1 - x_1) + (1 - m) \cdot y_1 - n \cdot \frac{x_1 \cdot x_2}{x_1 + x_2}, \\ \frac{dx_2}{d\tau} = (1 - m) \cdot y_2 - (1 + B) \cdot x_2 - n \cdot \frac{x_2^2}{x_1 + x_2}, \\ 0 = -y_1 + n \cdot \frac{x_1 \cdot x_2}{x_1 + x_2}, \\ 0 = m \cdot y_1 - (1 - m) \cdot y_2 + n \cdot \frac{x_2^2}{x_1 + x_2}, \\ 0 = x_2 - (1 - m) \cdot y_3. \end{cases} \quad (4)$$

The last three equations in (4) determine the zeroth order approximation of the slow manifold (the slow surface) of system (3). From these equations we have the expressions for  $y_1$ ,  $y_2$  and  $y_3$ :



$$\begin{aligned}
 y_1(t) &= n \cdot \frac{x(t)_1 \cdot x_2(t)}{x_1(t) + x_2(t)}, \\
 y_2(t) &= n \cdot \frac{x_2(t)}{x_1(t) + x_2(t)} \cdot \frac{Q_0 \cdot x_1(t) + n \cdot x_2(t)}{1 - m}, \\
 y_3(t) &= \frac{x_2(t)}{(1 - m)}.
 \end{aligned} \tag{5}$$

Since the Jacobi matrix of the fast subsystem of system (3):

$$D = \begin{pmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} & \frac{\partial g_1}{\partial y_3} \\ \frac{\partial g_2}{\partial y_1} & \frac{\partial g_2}{\partial y_2} & \frac{\partial g_2}{\partial y_3} \\ \frac{\partial g_3}{\partial y_1} & \frac{\partial g_3}{\partial y_2} & \frac{\partial g_3}{\partial y_3} \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ m & -(1 - m) & 0 \\ 0 & 0 & -(1 - m) \end{pmatrix}$$

has the negative eigenvalues, then the slow surface (5) is stable [4, 5], hence, we can replace the original system (3) by the reduced one. The reduced system is the projection of the original system on the slow surface (5) with preservation of the essential qualitative features of the dynamics of the complete system (see, for example, [6-17]).

The reduced system has the form

$$\begin{cases} \frac{dx_1}{d\tau} = B \cdot (1 - x_1) - Q_0 \cdot \frac{x_1 \cdot x_2}{x_1 + x_2}, \\ \frac{dx_2}{d\tau} = Q_0 \cdot \frac{x_1 \cdot x_2}{x_1 + x_2} - (1 + B) \cdot x_2, \end{cases} \tag{6}$$

where  $Q_0 = n \cdot m = \beta \cdot n / (\beta + \gamma)$  is a number of secondary infections produced by one infected individual in a population each member of which is susceptible.

System (6) is a homogeneous mixed model in which the infection spread parameter is a function of the parameters:  $Q_0 = \beta \cdot n / (\beta + \gamma)$ . Recall that the basic reproduction number is defined as  $\mathfrak{R}_0^c = Q_0 \cdot k / (\mu + k)$ . It is easy to see that  $\mathfrak{R}_0^c$  is the threshold parameter for the dynamic model (6).

#### 4. Analysis

System (6) has two equilibriums  $P_1$  and  $P_2$ :

$$P_1(1, 0), \quad P_2 \left( \frac{B}{Q_0 - 1}, -\frac{B - B^2 + BQ_0}{(1 + B)(Q_0 - 1)} \right).$$

Our goal is to find such conditions on the parameters of the system, reflecting the methods of treatment and preventive measures, in which the infected and latent infected populations are stabilized at the minimum value. From a mathematical point of view, this means that the equilibrium of the system is globally asymptotically stable, and their coordinates  $x_1$ ,  $x_2$  should be as small as possible.

It should be noted that the coordinates of the point  $P_1$  correspond to the following situation in real life: all persons are susceptible to tuberculosis, but not infected (either in active or latent form). In other words, this is the most favorable situation from the point of view of epidemiology. Therefore, if this point is globally asymptotically stable, then this will be the most favorable outcome.

The Jacobi matrix of the system (6)

$$J = \begin{pmatrix} -B - Q_0 \cdot \frac{x_2^2}{(x_1 + x_2)^2} & -Q_0 \cdot \frac{x_1^2}{(x_1 + x_2)^2} \\ Q_0 \cdot \frac{x_2^2}{(x_1 + x_2)^2} & -(1 + B) + Q_0 \cdot \frac{x_1^2}{(x_1 + x_2)^2} \end{pmatrix} \tag{7}$$

at the point  $P_1(1, 0)$  is

$$J_{P_1} = \begin{pmatrix} -B & -Q_0 \\ 0 & -(1 + B) + Q_0 \end{pmatrix}.$$

with the eigenvalues  $\lambda_1 = -B$ ,  $\lambda_2 = -1 - B + Q_0$ . Thus, according to the reduction principle, the condition  $Q_0 < 1 + B$  is the criteria for the globally asymptotic stability of the point  $P_1$ . This condition can be written in terms of the main reproductive number as  $\mathfrak{R}_0^c \leq 1$ .

The Jacobi matrix (7) at the equilibrium point  $P_2$  has the form

$$J_{P_2} = \begin{pmatrix} -\frac{B^2 - B(Q_0 - 2) + (Q_0 - 1)^2}{Q_0} & -\frac{(1+B)^2}{Q_0} \\ \frac{(1+B-Q_0)^2}{Q_0} & \frac{(1+B)(1+B-Q_0)}{Q_0} \end{pmatrix}.$$

For the asymptotic stability of the equilibrium  $P_2$ , it is necessary and sufficient that the trace of this matrix be negative and the determinant positive, i.e.,

$$\begin{cases} B^2 - B(Q_0 - 2) + (Q_0 - 1)^2 + (1+B)(Q_0 - 1 - B) > 0, \\ (B^2 - B(Q_0 - 2) + (Q_0 - 1)^2)(1+B)(Q_0 - 1 - B) + (1+B)^2(Q_0 - 1 - B)^2 > 0. \end{cases}$$

The last system can be written as

$$\begin{cases} (Q_0 - 1)Q_0 > 0, \\ (1+B)(Q_0 - 1 - B)(Q_0 - 1)Q_0 > 0. \end{cases}$$

Taking into account the physical meaning of the parameters, this condition is equivalent to the inequality  $Q_0 > 1 + B$ . Thus, we have the following statement.

**Theorem.** If  $\mathfrak{R}_0^c \leq 1$  a disease-free equilibrium  $P_1(1, 0)$  (i.e., the point determining the absence of infection in the population) is globally asymptotically stable. If  $\mathfrak{R}_0^c > 1$ , that point  $P_1(1, 0)$  is unstable and the equilibrium

$$P_2 \left( \frac{B}{Q_0 - 1}, -\frac{B - B^2 + BQ_0}{(1+B)(Q_0 - 1)} \right)$$

is globally asymptotically stable.

It should be noted that in [1] this statement was obtained on the basis of the Hopfenshtadt theorem.

Although the singular point  $P_2$  corresponds to the situation when the infected individuals in the population are present, but for typical values of the parameters, the ratio of the quantity of latently infected individuals to the amount of susceptible individuals is insignificant. Hence, the situation when this point is asymptotically stable is not so bad from the point of view of the real situation. In other words, the latently infected individuals will exist but their quantity will not be so high, that is, the epidemic threshold will not be reached.

### 5. Correctness of model reduction

The solutions of the original and reduced systems were plotted with the help of Wolfram Mathematica 10.3 and NetBeans 8.0.1. Figures 1-4 show the graphs for systems (3) and (6) with the following parameters values:

$$\mu = 1 / 60, \beta = 2, \Lambda = \mu \times 10^5, n = 20, \gamma = 1.$$

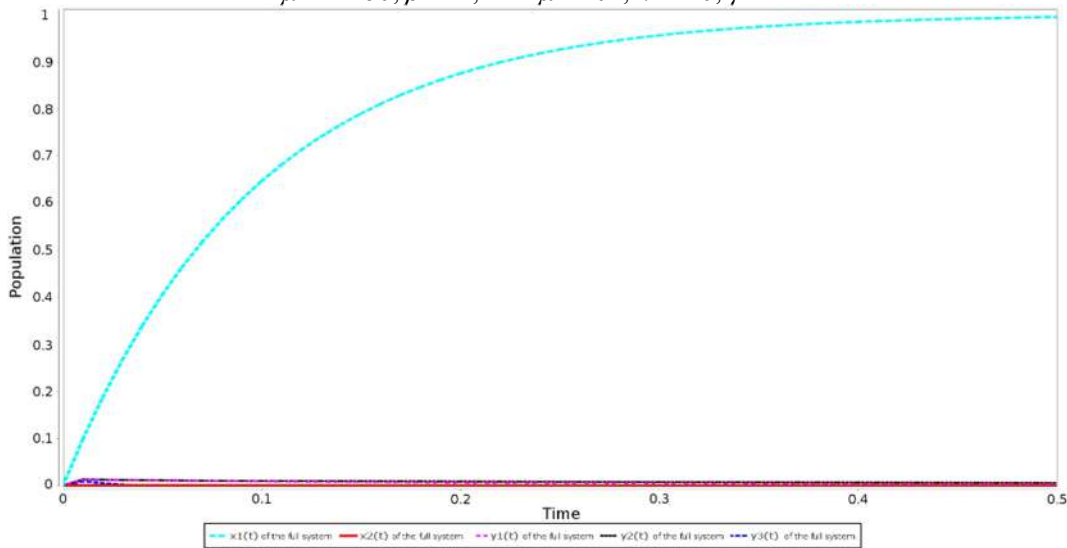


Fig. 1. The solutions of system (3) with  $\epsilon=0.00053$ .

Using a program written in the Java language in the NetBeans 8.0.1 environment, the errors in the deviation of the plots of the solutions to the complete and reduced systems were determined. For the three cases considered, the errors between the graphs of the solutions of the original and reduced systems are 0.00010, 0.00012 and 0.00015, respectively. In Figure 5 one can see the solutions of the original and reduced systems. The figure clearly demonstrates the almost complete coincidence of the solutions plots, which confirms the correctness of the reduction performed. Thus, the conclusions drawn about the qualitative behavior of the solutions to the reduced system can be transferred to the original model (3).

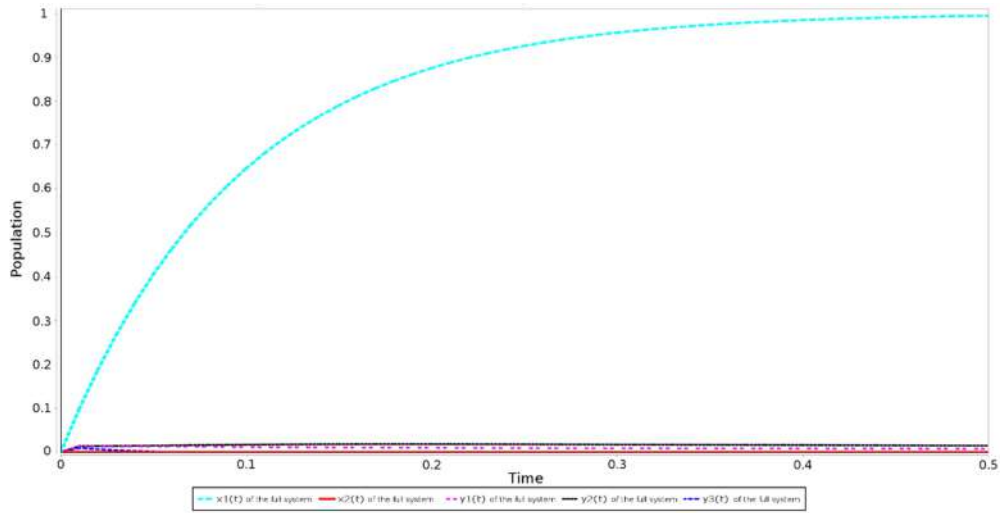


Fig. 2. The solutions of system (3) with  $\epsilon=0.00204$ .

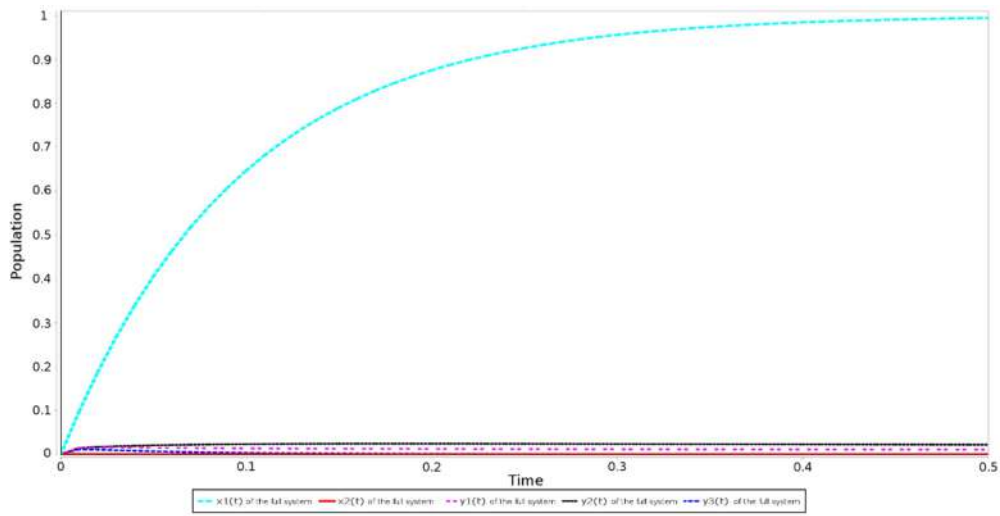


Fig. 3. The solutions of system (3) with  $\epsilon=0.00476$ .

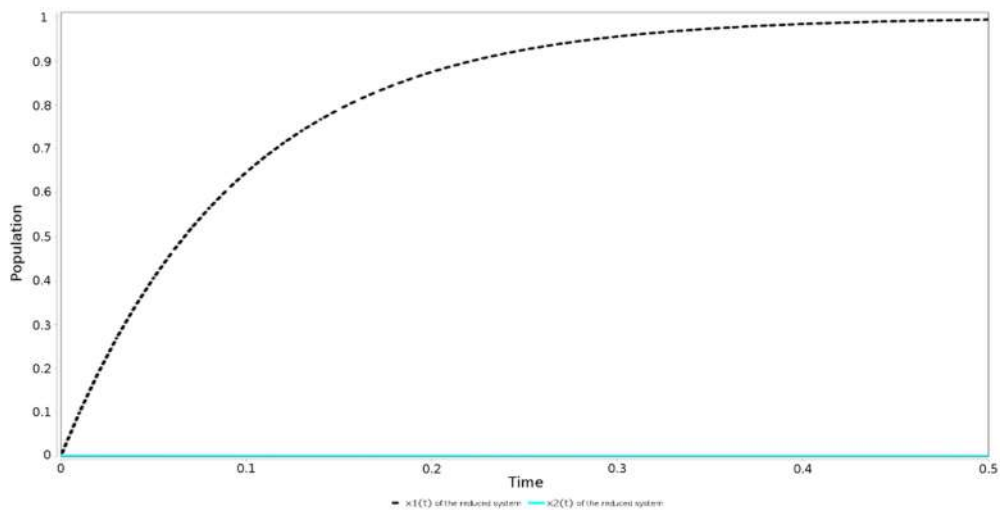


Fig. 4. The solutions of system (6).

## 6. Conclusion

In the present work, the tuberculosis model has been investigated via methods of qualitative analysis. It has been shown that the model has several time scales, so it can be represented as a singularly perturbed system of ODEs. The reduction of the

system has been carried out, as a result of which, the original system of five differential equations was replaced by its projection on the slow integral manifold. It should be noted that, due to the stability of the slow integral manifold, this reduction is correct, and the reduced system of two differential equations preserves the essential properties of the original model. The conditions under which the system has the globally asymptotically stable equilibrium have been found. This result means that under the appropriate selection of treatment and preventive measures, the spread of the tuberculosis epidemic can be completely suppressed.

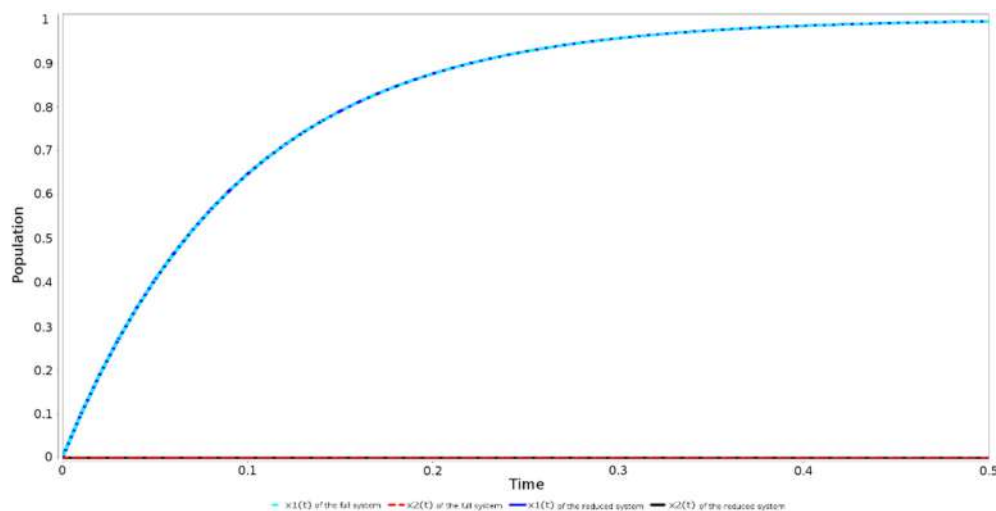


Fig. 5. The solutions  $x_1(t)$ ,  $x_2(t)$  of systems (3) and (6) with  $\varepsilon=0.00053$ .

## Acknowledgements

This study was supported by the Russian Foundation for Basic Research and Samara region (grant 16-41-630529-p) and the Ministry of Education and Science of the Russian Federation as part of a program of increasing the competitiveness of SSAU in the period 2013–2020.

## References

- [1] Song B, Castillo-Chavez C, Aparicio JP. Tuberculosis models with fast and slow dynamics: the role of close and casual contacts. *Mathematical Biosciences* 2002; 180: 187–205.
- [2] Castillo-Chavez C, Song B. Dynamical models of tuberculosis and their applications. *Mathematical biosciences and engineering* 2004; 361–404.
- [3] Vynnycky E, Fine PE. The long-term dynamics of tuberculosis and other diseases with long serial intervals: implications of and for changing reproduction numbers. *Epidemiol Infect* 1998; 121(2): 309–324.
- [4] Sobolev VA, Shchepakina EA. *Model Reduction and Critical Phenomena in Macrokinetics*. Moscow: Energoatomizdat Publisher, 2010; 320 p. (in Russian)
- [5] Shchepakina E, Sobolev V, Mortell MP. *Singular Perturbations: Introduction to System Order Reduction Methods with Applications*. Springer Lecture Notes in Mathematics 2014; 2114: 212 + XIII p.
- [6] Strygin VV, Sobolev VA. Effect of geometric and kinetic parameters and energy dissipation on orientation stability of dual-spin satellites. *Cosmic Research* 1976; 14(3): 331–335.
- [7] Gavin C, Pokrovskii A, Prentice M, Sobolev V. Dynamics of a Lotka-Volterra type model with applications to marine phage population dynamics. *J. Phys.: Conf. Series* 2006; 55(1): 80–93.
- [8] Pokrovskii A, Shchepakina E, Sobolev V. Canard doublet in a Lotka-Volterra type model. *J. Phys.: Conf. Series* 2008; 138: 012019.
- [9] Sazhin SS, Shchepakina E, Sobolev V. Order reduction of a non-Lipschitzian model of monodisperse spray ignition. *Mathematical and Computer Modelling* 2010; 52(3-4): 529–537.
- [10] Pokrovskii A, Rachinskii D, Sobolev V, Zhezherun A. Topological degree in analysis of canard-type trajectories in 3-D systems. *Applicable Analysis* 2011; 90(7): 1123–1139.
- [11] Sobolev VA, Tropkina EA. Asymptotic expansions of slow invariant manifolds and reduction of chemical kinetics models. *Computational Mathematics and Mathematical Physics* 2012; 52(1): 75–89.
- [12] Korobeinikov A, Archibasov A, Sobolev V. Order reduction for an RNA virus evolution model. *Journal Mathematical Biosciences and Engineering* 2015; 12(5): 1007–1016.
- [13] Archibasov AA, Sobolev VA, Korobeinikov A. Asymptotic expansions of solutions in a singularly perturbed model of virus evolution. *Computational Mathematics and Mathematical Physics* 2015; 55(2): 240–250.
- [14] Korobeinikov A, Archibasov A, Sobolev V. Multi-scale problem in the model of RNA virus evolution. *J. Phys.: Conf. Series* 2016; 727(1): 012007.
- [15] Sobolev VA. Canards and the effect of apparent disappearance. *CEUR Workshop Proceedings* 2015; 1490: 190–197.
- [16] Lapshova MA, Shchepakina EA. Study of the dynamical model of HIV. *CEUR Workshop Proceedings* 2016; 1638: 600–609.
- [17] Korobeinikov A, Shchepakina E, Sobolev V. Paradox of enrichment and system order reduction: bacteriophages dynamics as case study. *Math Med Biol.* 2016; 33(3): 359–369.

# Forecasting models generation of the electronic means quality

R.O. Mishanov<sup>1</sup>, S.V. Tyulevin<sup>2</sup>, M.N. Piganov<sup>1</sup>, E.S. Erantseva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>JSC SRC Progress, 18 Zemetsa street, 443009, Samara, Russia

---

## Abstract

The article describes the results of forecasting models generation of quality and reliability indicators of the electronic means. In the learning process variants of normalizing and centering of controlled parameters are described. Much attention is given to the methods of the Theory of Pattern Recognition and extrapolation methods. This paper gives information about the advanced technique of the models generation and individual forecasting of electronic means for the space equipment. The verification of derived models is investigated in detail. Special emphasis is paid to the analysis of the models efficiency.

*Keywords:* forecasting model; electronic means; verification; learning; informative parameters; analysis

---

## 1. Introduction

A realization of increasing requirements to the quality and reliability of the radio-electronic means and electronic components (EC) is ensured by the improvement of their design, manufacturing technology, controlling methods and testing. In addition, some hidden defects are not detected by the existing system of technological control and testing methods. The decisive influence on the reliability of hidden defects determines the development of works on the investigation of mechanisms and the causes of failures. However, a special interest is caused by using methods and means of flaw detection and physicochemical analysis.

Despite the effectiveness of work in this direction, the complexity and high cost of their implementation caused the necessity to search for and develop methods and means to identify hidden defects of the EC, which correspond to the pace of modern batch production. In addition, about 30% of defects and failures of EC cannot be controlled by these methods and means [1].

Thus, methods of testing and forecasting reliability and other quality indicators based on the informative parameters are being developed [2-8], which are reposed on the assumption of the existence of a stochastic connection between reliability and initial values of the informative parameters set of the product. The choice of the informative parameters set has a decisive influence on the validity of testing and forecasting. Ensuring the presence of informative parameters in the initial set is assigned to the researcher and in most cases is a very difficult task.

Ensuring the quality and reliability of space electronics requires a wide implementation of new methods of diagnostic nondestructive testing (NDT) [9-15]. For their development, it is necessary to establish the dependencies of the main reliability indicators on the physical properties and parameters of the devices, on the physicochemical processes occurring in them, and on the physical nature of the failures mechanisms [16].

One of the promising directions in the development of effective and economically acceptable methods for assessing the quality and reliability is to forecast their future state.

Forecasting failures of the devices can be carried out at various stages of their life cycle (control, testing, application, operation). The individual forecasting (IF) provides the greatest accuracy. Its meaning is to estimate the potential reliability of each instance using the forecasting model and information about the value of the informative parameter or results of monitoring the instances [17]. A structural IF model is required to generate an operator (mathematical model), an algorithm, an individual forecasting technique, and a hardware quality management. Such a model is generated in the form of an enlarged technological scheme with a description of the functions performed by the component parts [18].

A new structural forecasting model was proposed to increase the accuracy of the IF. It includes the following interrelated steps:

- analysis of the IF methods;
- physical and technical analysis of the failures;
- preliminary selection of the informative parameters and selection of the forecasting parameters;
- development of the investigation test technique;
- learning experiment;
- final selection of the informative parameters;
- selection of the IF method;
- algorithm development;
- program development;
- evaluation of the software product quality;
- development of the forecast model (the IF operator);
- evaluation of the IF operator models quality;
- development of working technique;
- verification of the model;

- attestation of the technique;
- operational forecasting;
- optimization of the model;
- refinement of the IF model;
- clarifying learning experiment;
- development or selection of new informative parameters;
- definition of levels;
- development of the recommendations;
- technological process (TP);
- parameter checkout of the radio-electronic means;
- change of the design and technology option;
- refinement of the technique;
- verification of the updated technique;
- heuristic forecasting or a rejection.

## 2. Development of the IF operators based on the regression models

The IF task including the value estimation of the forecasting parameter with a large number of the informative parameters was solved using the regression models. A problem statement was reduced to the determination of the operator  $H_x$ .

When the linear model of the connection between  $\tilde{y}$  and  $x_i$  is adopted the estimation of the forecasting parameter value of the  $j^{\text{th}}$  element is defined by [19]:

$$y^{*(j)}(t_f) = H_x \left[ \{x_i^{(j)}\} \right] = B_0 + B_1 x_1^{(j)} + B_2 x_2^{(j)} + \dots + B_i x_i^{(j)} + \dots + B_k x_k^{(j)}, \quad (1)$$

where  $x_i^{(j)}$  – the value of the  $i^{\text{th}}$  attribute of the  $j^{\text{th}}$  element;  $B_i$  – constant coefficients.

To find the coefficients  $B_i$  in a linear regression model, it is more convenient to turn the initial data to the centered and normalized values  $\tilde{x}_{ic}$ , which were determined by:

$$\tilde{x}_{ic} = \frac{\tilde{x}_i - M^*[\tilde{x}_i]}{D^{*1/2}[\tilde{x}_i]}.$$

$M^*[x_i]$  and  $D^*[x_i]$  are the estimates of the expected value and standard deviation of the random variable  $\tilde{x}_i$  calculated from the learning experiment data:

$$M^*[\tilde{x}_i] = \frac{1}{n} \sum_{j=1}^n x_i^{(j)};$$

$$D^{*1/2}[\tilde{x}_i] = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_i^{(j)} - M^*[\tilde{x}_i])^2}.$$

The idea of representing the connection between the forecasting parameter and informative parameters in the form of a regression model is as follows [20].

The coefficients  $b_i$  always can be found for any centered and normalized values  $\tilde{y}_{ic}$  and  $\tilde{x}_{ic}$  while the equation (2) has meaning regardless of the distribution law of random variables.

$$\tilde{y}_c = b_1 \tilde{x}_{1c} + b_2 \tilde{x}_{2c} + \dots + b_k \tilde{x}_{kc} + \Delta\tilde{y}, \quad (2)$$

In this equation  $b_i$  are the constant coefficients of the regression model with centered and normalized values of the random variables;  $\Delta\tilde{y}$  – a forecasting error.

If the values of the coefficients  $b_i$  are found, the estimation of the forecasting parameter value can be determined from the expression (2). The coefficients  $b_i$  must be such that the error variance  $D[\Delta\tilde{y}]$  is minimal, and the expected value of the error  $M[\Delta\tilde{y}]$  equals zero, i. e.

$$D[\Delta\tilde{y}] \rightarrow \min, \quad M[\Delta\tilde{y}] = 0.$$

If the error variance does not exceed the allowable value, the forecasting operator can be recommended to estimate the value of the forecasting parameter of new instances. In this case, having measured the values of its characteristics for the  $m^{\text{th}}$  instance and substituting them into expression (1), we obtain the estimate:

$$y^{*(m)}(t_f) = B_0 + B_1 x_1^{(m)} + B_2 x_2^{(m)} + \dots + B_k x_k^{(m)}.$$

The estimation of the forecasting error will be more accurate than the larger sample size is used in learning experiment. In this case the estimates of the expectation value, the standard deviation and the correlation coefficient will be found more accurately. For CMOS chips and stabilitrons the forecasting operators were obtained (Table 1).

Table 1. The forecasting models of study samples.

Number of sample	Forecasting model (IF operator)
Sample №44	$\frac{\Delta I_{ic}}{I_{ic}} = -29,53 + 29,11 t_p^+ - 51,07 U_s$
Sample №45	$\Delta U_s = -46,94 + 42,04 K_T + 0,096 R_d$

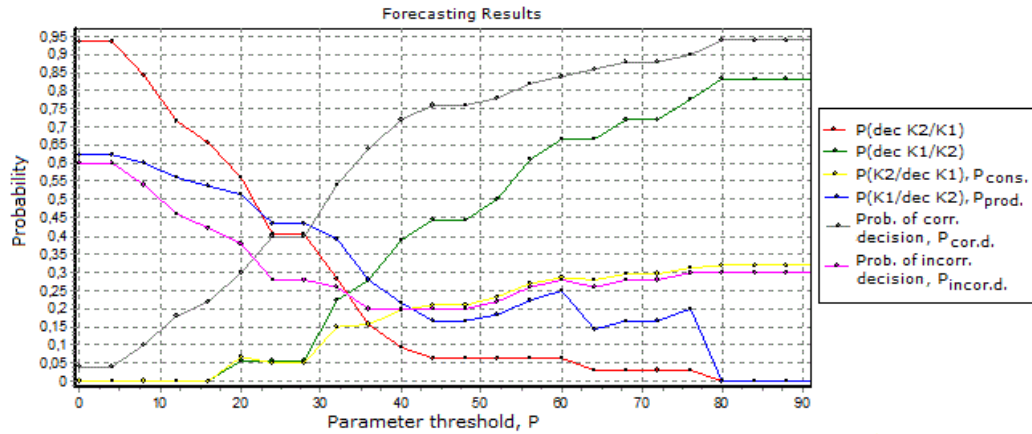


Fig. 1. The dependence of the probabilistic characteristics on the threshold  $P$  of the regression function of the CMOS chips.

$\Delta I_{lc}/I_{lc}$  – a leakage current drift,  $t_p^+$  – a rise time of the signal,  $U_s$  – a supply voltage,  $\Delta U_s$  – a stabilized voltage drift,  $K_T$  – a temperature coefficient of stabilization,  $R_d$  – a differential resistance.

Figure 1 shows the influence of the threshold  $P$  on the forecasting efficiency of the CMOS chips.

The analysis of this model have shown that the forecasting operator for the CMOS chips provides the optimal value of the forecasting indicators at the threshold  $P = 35$ . In this case the risk of the incorrect decision  $P_{inc.d}$  equals 0,22; Consumer's risk ( $\beta$ -Risk)  $P_{cons.}$  equals 0,18; Producer's risk ( $\alpha$ -Risk)  $P_{prod.}$  equals 0,13. The minimum value of the  $P_{cons.}$  equals 0 when  $P = 0 \dots 16$ ,  $P_{inc.d} = 0,6 \dots 0,42$ ;  $P_{prod.} = 0,63 \dots 0,54$ . The minimum value of the  $P_{prod.}$  equals 0 when  $P = 80 \dots 90$ ,  $P_{inc.d} = 0,3$ ;  $P_{cons.} = 0,32 \dots 0,33$ .

Figure 2 shows the influence of the threshold  $P$  on the forecasting efficiency of the stabilitrons.

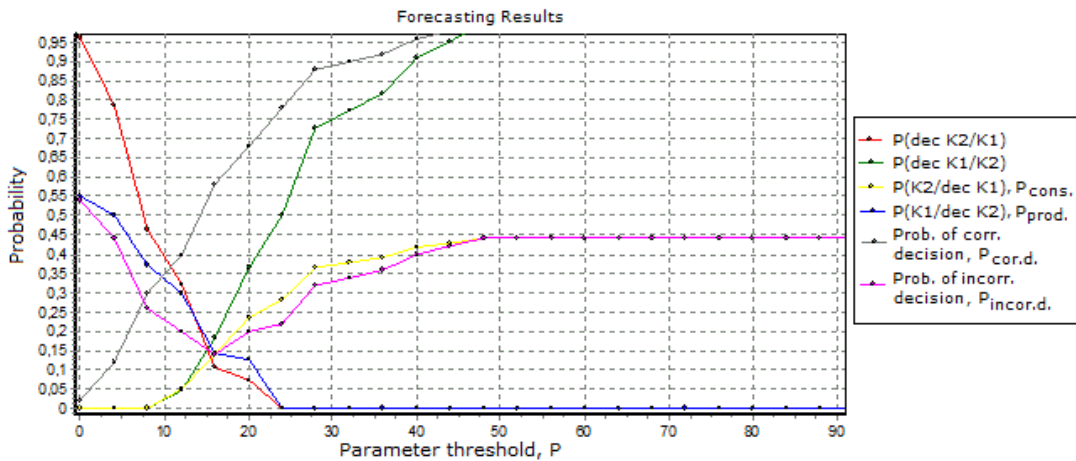


Fig. 2. The dependence of the probabilistic characteristics on the threshold  $P$  of the regression function of the stabilitrons.

The analysis of this model have shown that the forecasting operator for the stabilitrons provides the optimal value of the forecasting indicators at the threshold  $P = 16$ . In this case the risk of the incorrect decision  $P_{inc.d}$  equals 0,15; Consumer's risk ( $\beta$ -Risk)  $P_{cons.}$  equals 0,14; Producer's risk ( $\alpha$ -Risk)  $P_{prod.}$  equals 0,14. The minimum value of the  $P_{cons.}$  equals 0 when  $P = 0 \dots 8$ ,  $P_{inc.d} = 0,54 \dots 0,26$ ;  $P_{prod.} = 0,55 \dots 0,37$ . The minimum value of the  $P_{prod.}$  equals 0 when  $P = 24 \dots 90$ ,  $P_{inc.d} = 0,22 \dots 0,44$ ;  $P_{cons.} = 0,29 \dots 0,44$ .

### 3. The models verification

The method of discriminant functions was used for the models verification.

In general terms the problem formulation of such forecasting reduces to find the operator  $H_{xcl}$ . It is desirable to have the simplest model, when the hyperplane is a surface that divides the space into two regions.

The equation of the  $(k-1)$ -dimensional hyperplane in the  $k$ -dimensional feature space has the form:

$$g(x_1, x_2, \dots, x_k) = B_1 x_1 + B_2 x_2 + \dots + B_k x_k = P_d,$$

where  $P_d, B_1, B_2, \dots, B_k$  – constant coefficients that define the position of the hyperplane in the  $k$ -dimensional space.

Then the discriminant function takes the form:

$$g(x_1, x_2, \dots, x_k) = B_1 \tilde{x}_1 + B_2 \tilde{x}_2 + \dots + B_k \tilde{x}_k.$$

In this function the dimension of the coefficients  $B_i$  is inverse to the dimension of the corresponding characteristics  $\tilde{x}_i$ .

It was required to find those values of the coefficients  $P_d$  and  $B_i$ , which in the best way (in the sense of a misclassifications minimum) would specify the position of this hyperplane in the feature space. Since the sample size is limited the estimates  $\beta_i$  were determined.

The following approach was used to find the estimates of the coefficients  $\beta_i$ . According to the learning experiment, the actual class is known, to which each of  $n$  copies belongs –  $K_s^{(j)}$ . It is possible to find the estimates of conditional expected value and conditional variance of each  $i^{\text{th}}$  attribute  $x_i$ :

$$M^*[\tilde{x}_i/K_1] = \frac{1}{n_1} \sum_{\substack{j=1 \\ j \in K_1}}^{n_1} x_i^{(j)},$$

$$D^*[\tilde{x}_i/K_1] = \frac{1}{n_1 - 1} \sum_{\substack{j=1 \\ j \in K_1}}^{n_1} \{x_i^{(j)} - D[\tilde{x}_i/K_1]\}^2,$$

$$M^*[\tilde{x}_i/K_2] = \frac{1}{n_2} \sum_{\substack{j=1 \\ j \in K_2}}^{n_2} x_i^{(j)},$$

$$D^*[\tilde{x}_i/K_2] = \frac{1}{n_2 - 1} \sum_{\substack{j=1 \\ j \in K_2}}^{n_2} \{x_i^{(j)} - D[\tilde{x}_i/K_2]\}^2.$$

$n_1$  and  $n_2$  – number of the instances, which belong to the class  $K_1$  and  $K_2$ , respectively, so that  $n_1 + n_2 = n$ .

Using theorems on the numerical characteristics of random variables, the estimates of the conditional expected values of random variable were determined as:

$$G = g(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_k).$$

If the instance belongs to the class  $K_1$ :

$$M^*[G/K_1] = \sum_{i=1}^k \beta_i M^*[\tilde{x}_i/K_1] \quad (3)$$

and to the class  $K_2$ :

$$M^*[G/K_2] = \sum_{i=1}^k \beta_i M^*[\tilde{x}_i/K_2]. \quad (4)$$

If the attributes are not correlated the corresponding estimates of conditional variances are equal:

$$D^*[G/K_1] = \sum_{i=1}^k \beta_i^2 D^*[\tilde{x}_i/K_1]; \quad (5)$$

$$D^*[G/K_2] = \sum_{i=1}^k \beta_i^2 D^*[\tilde{x}_i/K_2]; \quad (6)$$

If the classes are well separated, then  $M^*[G/K_1]$  and  $M^*[G/K_2]$  will differ significantly, i.e.  $D^*[G/K_1]$  and  $D^*[G/K_2]$  are small. Therefore, as an optimization criterion for finding estimates of the coefficients  $\beta_i$ , we used an expression of the form:

$$\frac{M^*[G/K_1] - M^*[G/K_2]}{\sqrt{D^*[G/K_1] + D^*[G/K_2]}} \rightarrow \text{extr.} \quad (7)$$

After substituting in the expression (7) the estimates of the conditional expected values and conditional variances of the random variable  $G$ , determined by the expressions (3) - (6), we obtain the function:

$$V(\beta_1, \dots, \beta_k) = \left| \frac{\sum_{i=1}^k \beta_i M^*[\tilde{x}_i/K_1] - \sum_{i=1}^k \beta_i M^*[\tilde{x}_i/K_2]}{\sqrt{\sum_{i=1}^k \beta_i^2 D^*[\tilde{x}_i/K_1] - \sum_{i=1}^k \beta_i^2 D^*[\tilde{x}_i/K_2]}} \right|. \quad (8)$$

Taking partial derivatives  $\partial V / \partial \beta_i$  and equating them to zero, we obtain a system of  $k$  algebraic equations with  $k$  unknown coefficients  $\beta_1, \beta_2, \dots, \beta_k$  for finding optimal estimates  $\beta_{i \text{ opt}}$ . The obtained coefficients  $\beta_{i \text{ opt}}$  will determine the best slope of the hyperplane in the feature space.

Then we find the threshold value  $P_d$  for the discriminant function  $g(x_1, x_2, \dots, x_k)$ , which specifies the best position of the separating hyperplane. Obviously, the following condition must be satisfied:

$$M^*[G/K_1] > P_d > M^*[G/K_2]$$

or

$$M^*[G/K_1] < P_d < M^*[G/K_2].$$

When the threshold is changed, the risk of the incorrect decisions will change. The value of the threshold was found by several recalculations of the probability of incorrect decisions from the data of the learning experiment for various  $P_d$  and by choosing one of them at which the risk of incorrect decisions turned out to be the least.

If the obtained risk does not exceed the permissible value, the previously found operator can be used forecast the class of new instances (which not participating in the learning experiment). For this, the values of the attributes  $x_i^{(m)}$  of the new  $m^{\text{th}}$  instance are measured and the discriminant function has the form:

$$G^{(m)} = g(x_1^{(m)}, x_2^{(m)}, \dots, x_k^{(m)}) = \sum_{i=1}^k \beta_i x_i^{(m)}.$$



If  $M^*[G/K_1] > M^*[G/K_2]$  and  $G^{(m)} \geq P_d$ , then a decision is to relegate the  $m^{\text{th}}$  instance to the class  $K_1$ ,  $G^{(m)} < P_d$ , then a decision is to relegate it to the class  $K_2$ .

The method of discriminant functions made it possible to obtain the forecasting operators (Table 2):

Table 2. The forecasting models of study samples.

Number of sample	Forecasting model (IF operator)
Sample №44	$P_d = \frac{\Delta I_{lc}}{I_{lc}} + 0,76t_p^+ + 0,5U_s$
Sample №45	$P_d = \Delta U_s + 0,75K_T + 0,28R_d$

Figure 3 and 4 show the dependencies of the probabilistic characteristics on the discriminant function threshold  $P_d$  for the CMOS chips and stabilitrons.

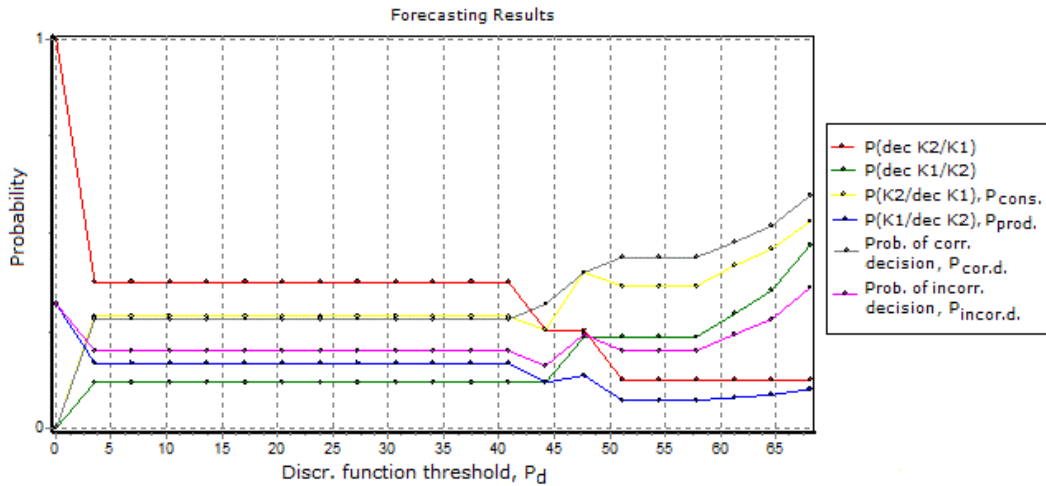


Fig. 3. The influence of the threshold  $P_d$  on the performance characteristics of the IF operator for the CMOS chips.

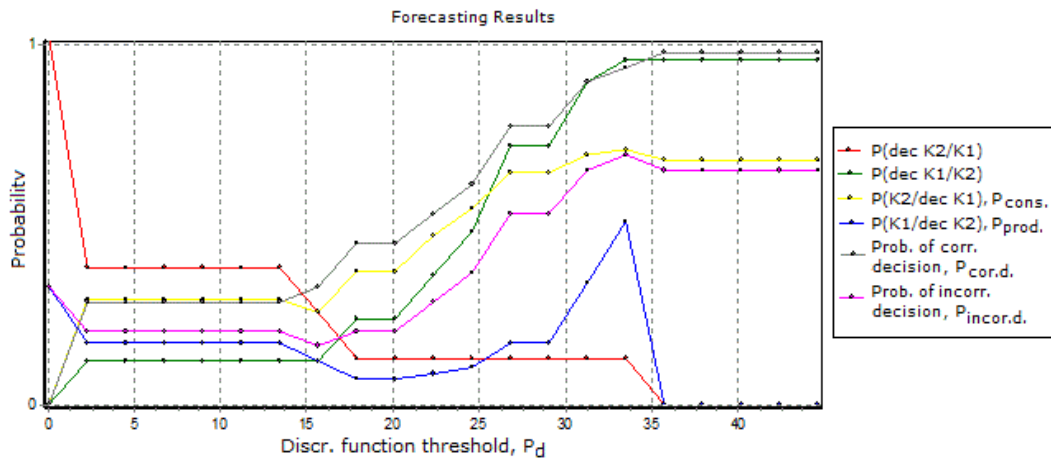


Fig. 4. The influence of the threshold  $P_d$  on the performance characteristics of the IF operator for the stabilitrons.

The optimal values of the forecasting indicators for the CMOS chips are at the threshold  $P_d = 44$ . In this case the risk of the incorrect decision  $P_{inc.d} = 0,17$ ; Consumer's risk ( $\beta$ -Risk)  $P_{cons.} = 0,27$ ; Producer's risk ( $\alpha$ -Risk)  $P_{prod.} = 0,13$ . The minimum value of the  $P_{cons.}$  equals 0,27 when  $P_d = 44$ . The minimum value of the  $P_{prod.}$  equals 0 when  $P_d = 57$ ;  $P_{inc.d} = 0,21$ ;  $P_{cons.} = 0,37$ .

The optimal values of the forecasting indicators for the stabilitrons are at the threshold  $P_d = 16$ . In this case the risk of the incorrect decision  $P_{inc.d} = 0,18$ ; Consumer's risk ( $\beta$ -Risk)  $P_{cons.} = 0,25$ ; Producer's risk ( $\alpha$ -Risk)  $P_{prod.} = 0,13$ . The minimum value of the  $P_{cons.}$  equals 0,25 when  $P_d = 16$ . The minimum value of the  $P_{prod.}$  equals 0 when  $P_d \geq 36$ ;  $P_{inc.d} = 0,52$ ;  $P_{cons.} = 0,57$ .

#### 4. Conclusion

The method of regression models was chosen for the forecasting models generation of the spacecraft electronic means. The CMOS chips and the stabilitrons were used as the electronic means. The forecasting models allow to provide the IF with the probability of correct decisions  $P_{cor.d} = 0,78$  for the chips and  $P_{cor.d} = 0,85$  for the stabilitrons. The method of discriminant functions was used to verify obtained models. They gave close to the initial models probabilities of the incorrect decisions: for the chips  $P_{inc.d} = 0,22$  and 0,17; for the stabilitrons  $P_{inc.d} = 0,15$  and 0,18.

Consequently, these models can be used at the stage of operational forecasting.

## References

- [1] Berezhnoy VP, Yusov YP, Khodnevich SP. Electrophysical diagnosis of the elements of radio electronic means. Moscow: CRI Electronica Publisher, 1990; 304 p.
- [2] Zhadnov V. Reliability forecasting of electronic means with mechanical elements. Ekaterinburg: Publishing house of LCC Fort-Dialog-Iset, 2014; 172 p.
- [3] Tyulevin SV, Piganov MN, Erantseva ES. To the problem of forecasting the quality indices of spacecraft elements. Reliability and quality of complex systems 2014; 1(5): 9–17.
- [4] Berenshtein GV, Dyachenko AM. Quality forecasting of the microcircuits based on the analysis of the internal stress. Physical basis of reliability and degradation of semiconductor devices. Chisinau 1991; 4: 36.
- [5] Luchino AI, Savina AS. Possibility investigation of the individual forecasting of the transistor long-eternity by the method of pattern recognition. Electronic Technology 1976; 10: 3–9.
- [6] Mishanov RO, MN. Piganov. Individual forecasting of quality characteristics by an extrapolation method for the stabilitrons and the integrated circuits. The experience of designing and application of CAD systems in Microelectronics (CADSM 2015): Proceeding XIII international conference. Ukraine, Lviv, 2015; 242–244.
- [7] Piganov MN, Tyulevin SV, Erantseva ES. Individual prognosis of quality indicators of space equipment elements. The experience of designing and application of CAD systems in microelectronics (CADSM 2015): Proceeding XIII international conference. Ukraine, Lviv, 2015; 367–371.
- [8] Mishanov RO, Piganov MN. Generation of the forecasting quality model of the semiconductor devices by extrapolation. Proceedings of the Samara Scientific Center of the Russian Academy of Sciences 2014; 6(4/3): 594–599.
- [9] Mishanov RO, Piganov MN. Technology of diagnostic for non-destructive control of the bipolar integrated circuits. Sense. Enable. Spitse: proceedings of the 2<sup>nd</sup> international scientific symposium. Russia, St. Peterburg, 2015; 38–41.
- [10] Sergeev VA, Yudin VV. Quality control of the digital integrated circuits by the parameters of the thermal bond matrix. Proceedings of the Institution of Higher Education. Electronics 2009; 6: 72–78.
- [11] Piganov MN, Tyulevin SV, Erantseva ES, Mishanov RO. Apparatus diagnostic for non-destructive control chip CMOS-Type. European science and technology: materials of the VIII international research and practice conference. Germany: Munich, 2014; 398–401.
- [12] Watchik R, Bucelot T, Li G. J. Appl. Phys. 1998; 9: 4734–4740.
- [13] Jonson JB. The Schottky effect in box frequency circuit. Phys. rev, 1925; 26: 71–85.
- [14] Chang MH, Das D, Varde PV, Pecht M. Light emitting diodes reliability review. Microelectronics Reliability 2012; 5: 762–782.
- [15] Kuba J. Application of low temperature infailure diagnostics of semiconductor devices. Power Semic. Hybrid Device – th Jnt. Spring Semin. Electrotechnol. Prenet, 1985; 31–34.
- [16] Pryanikov VS. Forecasting failures of semiconductor devices. Moscow: Energy, 1978; 122 p.
- [17] Piganov MN, Tyulevin SV. The reliability forecasting of the radio-electronic means. St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control System 2009; 1: 175–182.
- [18] Tyulevin SV, Piganov MN. Structural model of the individual forecasting of space equipment parameters. Vestnik of Samara State Aerospace University 2008; 1: 92–96.
- [19] Piganov MN. Technological fundamentals of quality assurance of microassemblies. Samara: SSAU Publishing house, 1999; 231 p.
- [20] Tyulevin SV. Method of the individual forecasting of the space radio-electronic means reliability. Actual problems of radio-electronics and telecommunications: materials of the Russian Scientific and Technical Conference. Samara: SSAU Publishing house, 2007; 162–163.

# About scarce resources allocation in conditions of incomplete information

N.L. Dodonova<sup>1</sup>, O.A. Kuznetsova<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The article examines the problem of the efficient allocation of resources in conditions of incomplete information concerning the parameters of agents' utility functions. Through business game the results are modeled and compared in conditions of incomplete information concerning the agents' utility functions. We experimentally prove the inexpediency of information distortion of the agents' effectiveness when using a non-manipulative distribution mechanism in a multi-step game.

*Keywords:* game theory; reflexive games; incomplete information; information structure; information management; distribution mechanisms; behavior models; utility functions; nontransferable utility; fuzzy logic

---

## 1. Introduction

The problem of effective resource allocation occurs in various applied problems [4]. If the resource value is limited and the participants interests do not coincide, a conflict situation arises. Interaction of participants in this case can be considered as a game. The description of several agents interaction includes the following parameters:

- the multitude of agents;
- agent preferences (he/she is assumed that each agent is interested in maximizing his profits);
- set of permissible actions;
- awareness of agents (at the time of making decisions about the chosen action);
- the order of functioning (the sequence of actions).

These parameters set the game. The game purpose is to define the multitude of active agents' actions. It means finding an equilibrium situation.

Decision-making models, behavior models, the equilibrium concept have been studied in game theory for more than 100 years. The review of the results is given, for example, in [3].

Basically, it is assumed that participants have the same information about the parameters of the game. A class of reflexive games in which agent awareness is not a common knowledge and agents make decisions according to their perceptions of opponents' preferences. Their permissible actions are described in [8].

Obviously, in the situation of incomplete information, participants' behavior patterns. Indeed, if the agent assumes that his rivals are "strong" players, he/she will stick to one behavior pattern; If he/she thinks that opponents are "weak" players, then the behavior pattern can change.

The process and result of the agent's thinking about the values of uncertain parameters, and what about these competitors think about these parameters, are called information reflection [8]. The players' perception hierarchy is represented by the form of information structure tree. The research and analysis of the game information structure allows to determine the conditions for the information equilibrium, as well as to set the information management task—to create the information structure creation that implements the equilibrium situation that is most beneficial to the Resources Allocation Center.

This paper is devoted to the investigation of the appropriateness of information distortion about the agents' effectiveness in different behavior models in the situation of incomplete awareness of the participants about the parameters of the game.

It is assumed that the game participants can distort information about their target functions parameters, posing as "strong" or "weak" players. A hypothesis that the information distortion about the values effectiveness, with the possibility of requests further distortion, does not have a significant effect on the limited resource distribution between the players, is confirmed experimentally.

The study was carried out using an original Fuzzy Logic Model (FLM) [6] and the Best Response Model (BRM) [1].

## 2. Basic concepts and parameters

Let us consider the problem of distributing the resource  $R$  between  $n$  players.  $R$  be a distributable resource;  $N$  is the number of players;

$u(x_i) = bx_i - a_i x_i^2, a > 0, b > 0$  is the utility function of the  $i$ -th player.

Obviously, the player will get the maximum profit at the point  $x_i^* = \frac{b}{2a_i}$ .

In the case  $\sum_{i=1}^n x_i^* > R$ , the conflict situation develops and players are forced to fight for the resource.

If  $S(x_1, x_2, \dots, x_n) = \sum_{i=1}^n (bx_i - a_i x_i^2)$  is the players total profit and there exists a restriction  $\sum_{i=1}^n x_i = R$ , then it is easy to show that  $S(x_1, x_2, \dots, x_n)$  reaches a maximum at the point  $(x_1^0, x_2^0, \dots, x_n^0)$ , where

$$x_i^0 = \frac{a_1 a_2 \dots a_{i-1} a_{i+1} \dots a_n}{\sum_{t \neq j} a_t a_j}, i = \overline{1, n}, j = \overline{1, n}.$$

If  $x_i^* \neq x_i^0$  then the  $i$ -th player will be interested in increasing his profit.

As a mathematical model of the described interest conflict situation, we will use the business game for resource allocation  $R$  between  $n$  players with a reverse priority mechanism. At each step of the game the participant makes an request  $s_i$  to the resource. The request is satisfied by the Resource Allocation Center in the volume

$$x(s_i) = \frac{\frac{A_i}{s_i}}{\sum_{j=1}^n \frac{A_j}{s_j}} R, i = \overline{1, n}, \text{ where } A_i = u(x_i^*).$$

The winning is determined by the player's profit from the resource obtained in the last step.

It should be note that the resource distribution is based on the knowledge of the values  $A_i = u(x_i^*)$  of each of the participants. In a sense,  $A_i$  can be interpreted as the utility limit of the  $i$ -th player. Let us suppose that the true values of  $A_i$  are not known to the Center (the cost factor  $a_i$  is known only to the player) and for the distribution of the resource the players themselves inform the Center of the value  $A_i$ . In this case, the player has the opportunity to exaggerate, downplay the limit of its usefulness or to convey its true meaning. Also, at each step, players report the value of the required resource, which is adjusted by the players in order to obtain the desired amount.

How will the distribution of the resource change in conditions of incomplete information of the Center about the usefulness of the players? Is it possible in such conditions to maximize the profit of an individual player and the total utility of the players? Is it profitable for participants to hide the true meaning of the limit of their usefulness?

### Purpose of the study

The purpose of this study is to compare the participants profit size in a business game on the resource distribution in different information levels of the players parameters conditions.

We tasks:

- to conduct a computational experiment in incomplete information conditions about the needs of players in resources, using different participants' behavior models;
- to conduct a computational experiment in incomplete information conditions about the players target functions and their resource needs, using different participants' behavior models;
- to conduct a comparative analysis of the results.

### 3. Description of the experiment

For carrying out the computing experiment two models will be used:

- Best Response Model (BRM);
- Fuzzy Logic Model (FLM).

The BRM [7] assumes that at the  $k+1$  step of the game, the bid value  $s_i^{k+1}$  must be such that  $x(s_i^{k+1}) = x_i^*$ . If the remaining players do not change their bids, then the volume of the request can be calculated from the condition

$$x(s_i^{k+1}) = \frac{\frac{A_i}{s_i^{k+1}}}{\sum_{j=1}^n \frac{A_j}{s_j^k} - \frac{A_i}{s_i^{k+1}}} R = x_i^*,$$

then

$$s_i^{k+1} = \frac{\frac{A_i}{x_i^*}}{\sum_{j=1}^n \frac{A_j}{s_j^k} - \frac{A_i}{s_i^k}} (R - x_i^*)$$

The FLM [6] uses the following input data:

$$\alpha_i = \frac{x(s_i)}{x_i^*}$$

$\alpha_i$  is the degree of satisfaction of the request;

$N$  is the proportion of players with  $\alpha_i \geq 1$ .

The rules base, which gives an assessment of the attractiveness of the player's actions, consists of the possible actions:

- to increase the request,
- to lower the request,
- not to change the request.

The rules base has the form:

R1. If the degree of the request satisfaction  $\alpha_i$  is small and the players share  $N$  is low, then the declining of the request attractiveness is great.

R2. If the degree of the request satisfaction  $\alpha_i$  is small and the players proportion  $N$  is high, then the attractiveness not to change the request is great.

R3. If the degree of the requests satisfaction  $\alpha_i$  is close to 1 and the players share  $N$  is low, then the declining of the request attractiveness is great.

R4. If the degree of the requests satisfaction  $\alpha_i$  is close to 1 and the share of players  $N$  is high, then the attractiveness of the bid increase is great.

R5. If the degree of the request satisfaction  $\alpha_i$  is large and the players share  $N$  is low, then the attractiveness not to change the request is great.

R6. If the degree of the requests satisfaction  $\alpha_i$  is large and the share of players  $N$  is high, then the attractiveness of the bid increase is great.

As a result of FLM, the evaluation  $\lambda \in [0,1]$  of the attractiveness of player actions is given. The player may increase the bid ( $P\uparrow$ ), lower the bid ( $P\downarrow$ ) or not to change the request ( $P0$ ).

Special software was developed for the experiment in the program environment O-Tree [9].

In the course of study, various combinations of the input parameters considered in Table 1 were considered. In each experiment, a series of 10 steps was conducted.

Table 1. Experiments input parameters combinations.

№ experiment	Utility function $u(x_i) = bx_i - a_ix_i^2$	Deficiency of resource $\left  R - \sum_{i=1}^n x_i^* \right $	Relative location of $x_i^*$	Behavior model
1	the same	small	the same	BRM
2	the same	large	the same	BRM
3	different	small	narrow spread	BRM
4	different	small	wide spread	BRM
5	different	large	narrow spread	BRM
6	different	large	wide spread	BRM
7	the same	small	the same	FLM
8	the same	large	the same	FLM
9	different	small	narrow spread	FLM
10	different	small	wide spread	FLM
11	different	large	narrow spread	FLM
12	different	large	wide spread	FLM

The form of utility function determines whether the player should maximize the amount of the resource he receives, or optimize it.

Deficiency of resource implies various tensions in the game and level request distortion.

Relative location of  $x_i^*$  means that the optimal resource values in different functions have the same deviation from equal distribution. In this case players have the same chance to be winner.

Behavior model means that players use special rules for their actions.

#### 4. Results and Discussion

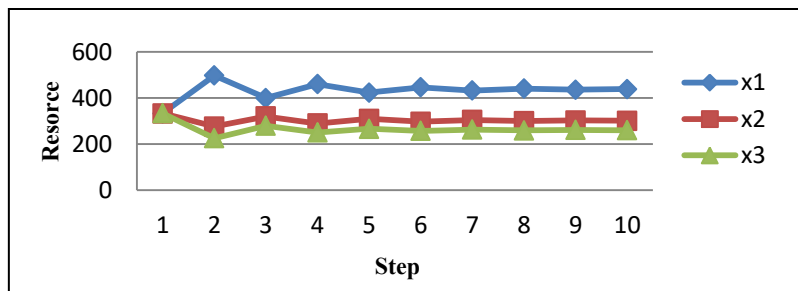


Fig. 1. Agents report the exact value of their effectiveness.

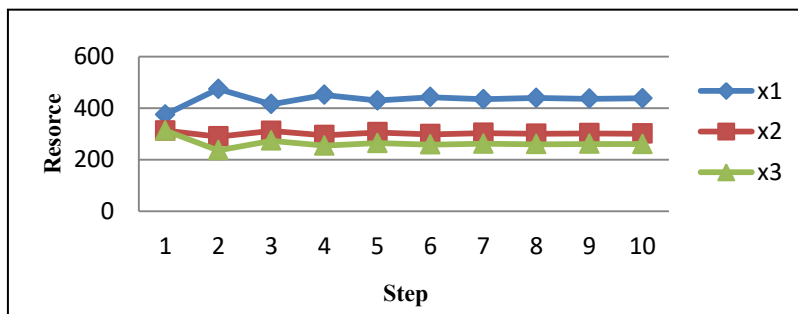


Fig. 2. The first agent overestimates the importance of its effectiveness.

The dynamics of resource allocation is presented on Figures 1, 2, 3.

Here  $x_1$  is the value of the resource allocated to the first player,  $x_2$  is the value of the resource allocated to the second player,  $x_3$  is the value of the resource allocated to the third player.

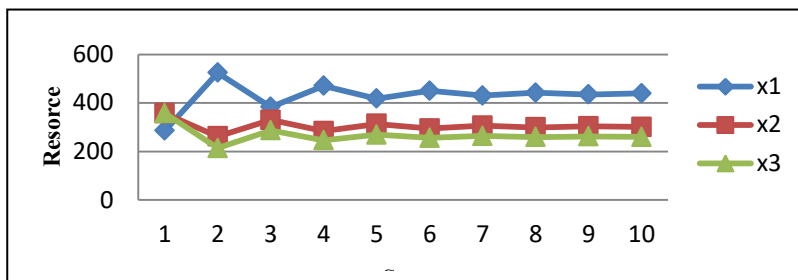


Fig. 3. The first agent underestimates the importance of its effectiveness.

In all cases, the deviation of the obtained resource from the optimal individual indicator is approximately the same.

Table 2 shows the relative deviations of the resource obtained by agents in cases of reliable reporting of information on effectiveness, overestimation of the first agent effectiveness, underestimation of the first agent effectiveness in the first step.

Table 2. The relative deviations of the resource obtained by agents in the first step.

	№1	№2	№3
x1	0,33	0,25	0,43
x2	0,07	0,12	0,00
x3	-0,07	0,00	-0,15

№ 1. In the first step all players provide reliable information about their own effectiveness and the amount of the required resource. In the next steps distorting the value of the resource request is distorted in accordance with the chosen behavior model.

№ 2. In the first step the player 1 overstates the information on its own efficiency by 20%, other players provide reliable information about their own effectiveness and all players report reliable information about the amount of the required resource. In the next steps distorting the value of the resource request in accordance with the chosen behavior model.

№ 3. In the first step the player 1 understates information about its own efficiency by 20% other players provide reliable information about their own effectiveness and all players report reliable information about the amount of the required resource. In the next steps distorting the value of the resource request in accordance with the chosen behavior model.

Table 3 shows the resources relative deviations obtained by agents in cases of reliable reporting of information on effectiveness, overestimation of the first agent effectiveness, underestimation of the first agent effectiveness in the tenth step.

Table 3. The relative deviations of the resource obtained by agents in the tenth step.

	№1	№2	№3
x1	0,12	0,12	0,12
x2	0,16	0,16	0,16
x3	0,17	0,17	0,16

The results of the calculations presented in the tables 2, 3. The information distortion about efficiency leads to a significant change in the distribution results in the first step. As we can see from the results of the calculations presented in the Table 3, the information distortion about efficiency does not lead to a change in the distribution results in the tenth step.

Figure 4 presents the averaged values of the relative deviations from the optimal resource values in games with BRM (exact information about the effectiveness of players, distorted information about the effectiveness of players).

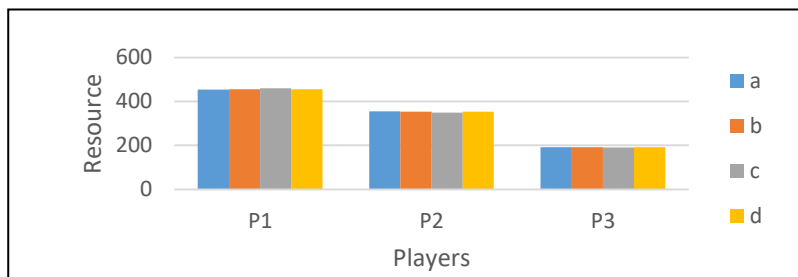


Fig. 4. The resource distribution.

*a* is the players provide reliable information about the maximum of their profits;

*b* is the players overestimate the value of their maximum profit;

*c* is the players underestimate the value of their maximum profit;

*d* is the players distort information about the maximum of their profits.

P1 is the first player, P2 is the second player, P1 is the third player.

Figure 5 shows the averaged values of the total utility of participants in games with BRM and FLM (exact information about the effectiveness of players, distorted information about the effectiveness of players)

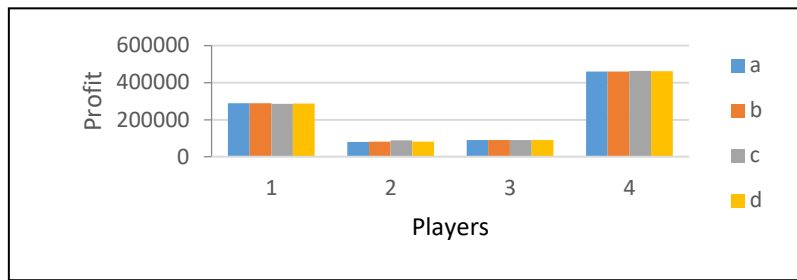


Fig. 5. The averaged values of the total utility.

Table 4 presents numerical data on the resource distribution among participants, the magnitude of individual and total profits.

Table 4. The resource distribution among participants.

Player	The value of the resource distribution				Profit of player			
	a	b	c	d	a	b	c	d
P1	453,57	455,4782	460,1372	455,9209	289962,8	288575,2	285095,7	288250,2
P2	355,1272	353,4823	349,4531	353,096	79677,78	82215,97	88318,84	82808,03
P3	191,3028	191,0396	190,4097	190,9831	89831,48	90110,22	90772,57	90169,83
Total profit					459472	460901,4	464187,1	461228,1

You can see that difference between total profit in the different experiments consists less of than 1%. We consider this deviation to be insignificant.

## 5. Conclusion

The article discusses the effectiveness of distorting information about the agents' effectiveness in incomplete information conditions in the limited resource distribution problem.

Experiments were performed by robots with various input parameters combinations. A comparative analysis of the games results with reliable and inaccurate information about the players effectiveness was carried out. The agents' profit and the system total profit are calculated.

The conducted experiments showed that a single distortion of information about the effectiveness of players, with constant distortion of players 'requests, does not affect the distribution of players' profits.

## References

- [1] Burkov VN, Danev B, Enaleev AK. Large systems: modeling of organizational mechanisms. M.: Science, 1989; 248 p.
- [2] Arifovic J, Ledyard J. A Behavioral Model for Mechanism Design: Individual Evolutionary Learning. Journal of Economic Behavior & Organization 2010.
- [3] Avtonomov VS. Model of man in economics. St. Petersburg: Economic School, 1998; 230 p.
- [4] Bogomolnaia A, Moulin H, Sandmirskiy F, Yanovskaya EB. Dividing Goods and Bads Under Additive Utilities. NRU Higher School of Economics. Series EC "Economics" 2016; 153.
- [5] Geraskin MI. Transferable utility distribution algorithm for multicriteria control in strongly coupled system with priorities. CEUR Workshop Proceedings 2016; 1638: 542—551. DOI: 10.18287/1613-0073-2016-1638-542- 551.
- [6] Geraskin MI, Egorova VV. The algorithm for dynamic optimization of the production cycle when custom planning in industry. CEUR Workshop Proceedings 2016; 1638: 552—568. DOI: 10.18287/1613-0073-2015-1638-552- 568.
- [7] Dodonov MV, Dodonova NL, Kuznetsova OA, Elistratov AA. Model of a decision support system using fuzzy logic in business resource allocation games with nontransferable utility. Management of large systems. Materials XIII Vseros. School-conf. young scientists. Moscow: IMP RAS, 2016.
- [8] Korgin NA. Representing sequential resource allocation mechanism in form of strategy-proof mechanism of multi-criteria active expertise. Management of large systems. Moscow: IMP RAS 2012; 36: 186—208.
- [9] Novikov DA, Chkhartishvili AG. Reflective games. Moscow: SINTEG, 2003; 160 p.
- [10] A software platform for economics experiments URL: <http://www.otree.org/>.

# A model of milling process based on Morlet wavelets decomposition of vibroacoustic signals

A.I. Khaymovich<sup>1</sup>, S.A. Prokhorov<sup>1</sup>, A.A. Stolbova<sup>1</sup>, A.I. Kondratyev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The paper considers the problem of online monitoring the condition of cutting tools to avoid its unexpected failure. To approach this problem we proposed a model of milling process based on Morlet decomposition of vibroacoustic signals. In addition, using the wavelets scalogram, we imposed a new condition that helps to improve early wear detection of the cutting tool. The findings of this research reveal the advantages of the proposed model compared to the previously reported models that rely on Haar wavelets and Short-time Fourier transform.

*Keywords:* milling process; acoustic emission; wear detection; Morlet wavelet decomposition

## 1. Introduction

The increasing demands for the characteristics of modern gas turbine engines make it necessary to improve the accuracy and reliability of their manufacture. This improvement permits to increase the durability of critically important components such as rotating turbine discs. The processing characteristics sharply deteriorate at high mechanical strength at high temperatures as well as low thermal conductivity of Ti / Ni-based alloys [1-5]. Cutting off parts from nickel-base heat-resistant alloys (for example, Inconel 718, Udimed 720) leads to both a rapid wear of the cutting tool and tool surface [1, 11-16], which can be generally called surface anomalies. These surface anomalies are the result of the bad processing characteristics of nickel-base alloys and the trend of rapid tool wear at cutting regardless of the types of machining operations [11, 12, 14-22]. Aircraft engine manufacturers are developing a monitoring system to detect anomalies in the processing and to react against it [34].

The procedure behind most monitoring systems consists of the following steps. First, it is a need to measure parameters second, these parameters need to be analyzed by means of specific methods such as wavelet decomposition, Short-time Fourier transform (STFT) and etc. One of the efficient methods of spectral analysis is the wavelet transformation (decomposition), the advantage of which is the possibility to analyze non-stationary signals. The wavelets frequently used in practice are described in [8, 34, 37].

The main purpose of this study is to develop a model of milling process based on Morlet decomposition of vibroacoustic signals and, thus, to propose tool wear condition. This condition is of use in solving the problem of identifying both non-stationary modes and early tool wear.

## 2. Problem statement

STFT assumes the stationarity of signals during a given time interval [19-22]. It can be expressed by

$$w(\tau, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-j\omega t} f(t) h(t - \tau) dt, \quad (1)$$

where  $f(t)$  is a given signal,  $h(t)$  is a Hanning window [28],  $\tau$  is a time delay.

The main drawback of STFT is the assumption of stationarity (permanence) of the signal on the time interval of the window. This issue increase errors in the analysis for such dynamic processes as milling process.

Wigner [29, 30] and later Cohen [21] improved the classical Fourier transform (T-F). Results of the Wigner distribution can comprise a cross-interference, because of signal is multicomponent.

Cohen [21] introduced the general class of distribution function in T-F as

$$w(t, \omega) = \frac{1}{2\pi} \iiint e^{-j(\theta + \tau\omega + \theta\omega)} f(\mu + \tau/2) f^*(\mu - \tau/2) \phi(\theta, \tau) d\mu d\tau d\theta, \quad (2)$$

where  $f^*(m)$  is the complex conjugate value,  $\phi(\theta, \tau)$  is a kernel function,  $\theta$  is a distribution parameter (in frequency domain).

Choi and Williams [31] made an improvement on Wigner distribution (WD). The Choi-Williams distribution (CWD) is

$$w(t, \omega) = \frac{1}{4\pi^{3/2}} \iint \frac{1}{\sqrt{\tau^2/\sigma}} e^{[-(u-t)^2/4\tau^2/\sigma - j\tau\omega]} f(\mu + \tau/2) f^*(\mu - \tau/2) d\mu d\tau. \quad (3)$$

If  $\sigma$  is large, CWD approaches to “plan” Wigner distribution. As  $\sigma$  reduces, cross interference decreases [32].

Zhao–Atlas–Marks distribution (ZAMD) [33] reduces the cross interference comprised in multicomponent signals. ZAMD is useful in modeling of small spectral peaks and analyze non-stationary multicomponent signals [32]. ZAMD has a kernel represented by (4),  $q$  is permanent.

$$\phi(\theta, \tau) = g(\tau) \tau \frac{\sin(q\theta\tau)}{q\theta\tau}. \quad (4)$$



As a result, power spectral density is defined by

$$w(t, \omega) = \frac{1}{4\pi a} \int_{-\infty}^{+\infty} g(\tau) e^{-j\tau\omega} \int_{1-|\tau|/a}^{1+|\tau|/a} f(\mu + \tau/2) f^*(\mu - \tau/2) d\mu d\tau. \quad (5)$$

Formant analysis [33] is used to analyze a vibroacoustic signals because these signals have multi-frequency components connected with different anomalies while cutting [35, 6].

The efficiency of time-frequency methods is presented in Fig. 1 [7].

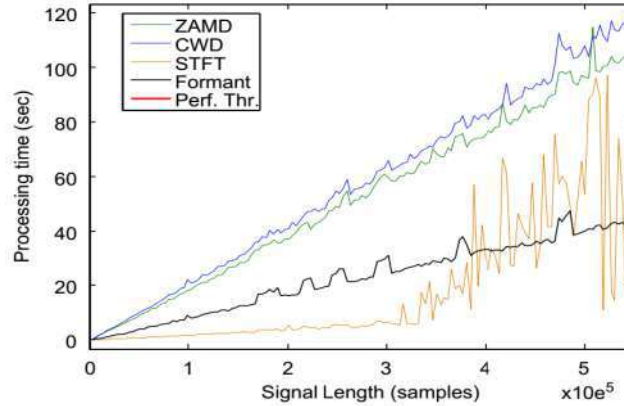


Fig. 1. Comparative efficiency of the STFT, CWD, ZAMD methods and formant-analysis [31].

One of the first and simplest wavelets is the discrete Haar wavelet:

$$\psi(t) = \begin{cases} 1, & 0 \leq t < 1/2, \\ -1, & 1/2 \leq t < 1, \\ 0, & t \notin [0,1). \end{cases} \quad (7)$$

The informative parameter characterizing the cutting tool (CT) wear is the dispersion of the detail coefficients of the Haar wavelet decomposition of AE signal. This parameter is insensitive to changes in processing modes [31]. The minimum duration of the analyzed sample is 0.1 s. Wear identification of cutting tool is carried out according to the energy value of the  $j$ -th detail factors. For Haar wavelet decomposition, it is advisable to take  $3 < j < 6$ . The forecast of CT wear in real time is in correction of the base model estimation from the results of current measurements of the AE signal parameters by an additive component obtained on the basis of extrapolation of the residual function. The study [35] proposes the adaptation of the suggested method for molding conditions by automatic window selection of a fragment of the AE signal which falls on the cutter tooth.

The main drawback of the Haar wavelet is the asymmetric and non-smooth, consequently, an infinite alternation of "petals" arises in the frequency domain due to sharp boundaries in the time domain. The complex Morlet wavelet does not suffer from these drawbacks.

### 3. A model of milling process based on Morlet wavelets decomposition of vibroacoustic signals

Wavelet transformation coefficients can be defined as [10, 36, 37]:

$$W_{\psi}(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt, \quad (6)$$

where  $f(t)$  is a random process,  $\psi(t)$  is a chosen wavelet,  $a \neq 0$  is a scale parameter,  $b \geq 0$  is a shift parameter.

Morlet wavelet is given by

$$\psi(t) = \exp(-jkt) \exp\left(-\frac{t^2}{2}\right), \quad (8)$$

where  $j$  is the imaginary unit, parameter  $k = 2\pi$  [37] controls the time-frequency resolution.

The graphical results of wavelet transformation can be calculated by

$$w_{i,j} = |W_{\psi}(a_i, b_j)|^2, \quad (9)$$

where  $i = 0, \dots, N_a - 1$ ,  $j = 0, \dots, N_b - 1$ ,  $N_a$  is a counting scale,  $N_b$  is a counting shift.

The scalograms are obtained from (9) as

$$y_i = \frac{1}{N_b} \sum_{j=0}^{N_b-1} w_{i,j}, \quad (10)$$

We propose to use the equation (11) to calculate area under curve of scalograms:

$$s = \Delta\omega \left( \frac{y_0 + y_{N-1}}{2} + \sum_{i=1}^{N-2} y_i \right), \quad (11)$$

where  $\Delta\omega$  is a frequency of quantization interval,  $y$  is a scalogram,  $N$  is a counting rate of scalograms.

We use a new identification criterion (12) to analyze processing parameters. This criterion is a cross-factor  $CF_{\Delta\omega}$  of the spectral energy density in the frequency bands  $\Delta\omega_{\max} \subset \Delta\omega_{\Sigma}$  of every local maximum of scalograms. We built the scalograms in the frequency intervals  $\Delta\omega_{\Sigma}$ .

$$CF_{\Delta\omega_{\max}} = \frac{\Delta\omega_{\Sigma} \int_{\Delta\omega_{\max}} w_{i,j} d\omega}{\Delta\omega_{\max} \int_{\Delta\omega_{\Sigma}} w_{i,j} d\omega}. \quad (12)$$

To identify wear the following equations were considered:

$$k_{\Delta\omega_{\max}} = \frac{CF_{\Delta\omega_{\max}}(t_0)}{CF_{\Delta\omega_{\max}}(t_d)}, \quad (13)$$

where  $t_0$  is the time of tool work without wear out,  $t_d$  is the time of tool work with wear out.

In accordance with equations (11-13), the calculation of the wear identification coefficient can be made by:

$$k_{\Delta\omega_{\max}} = \frac{s_{\Sigma}(t_d) \cdot s_{\Delta\omega_{\max}}(t_0)}{s_{\Sigma}(t_0) \cdot s_{\Delta\omega_{\max}}(t_d)}. \quad (14)$$

## 4. Results

### 4.1. Experiments design

The phenomena explained by the dislocation theory, of deformation distortions of the crystal lattice, friction, the formation and extension of cracks, phase transformations leads to AE. In metal cutting, the processes arised at an interaction between the part and tool are the most important sources of AE [23].

We register acoustic emission and power cutting of milling by the lateral and end surfaces of the milling tool. The main system element for measuring power cutting is the piezo-multicomponent dynamometer Kistler – Type 9257B (Switzerland) This dynamometer was installed at the base of the machining center Micron UCP 800. We use the LTR22 analog to frequency converter to record vibroacoustic signals with the microphone sensor (OCTAFON-110).

The connection scheme of the experimental setup for data collection is shown in Fig. 3.

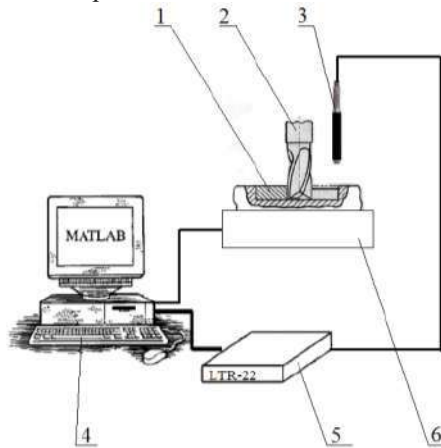


Fig. 2. Scheme of AE parameter measurement: 1 - sample, 2 – milling cutter, 3 – microphone- vibration meter, 4 – PC with software IIK, 5 – crane system LTR22, 6 – dynamometric table built up on the machine platen.

We used the four-tooth carbide monolithic milling tool by Seco JHP 780120E2R15Q0Z4-M64 with a diameter of 12 mm. In the experiments, we used new milling tools and tools with worn teeth, Fig. 4.



Fig. 3. Milling cutters for carrying out the research.

The machining process with variable allowance was simulated to analyze the influence of the cutting depth on the acoustic emission parameters and the stability of the wear identification technique. The processed sample of steel 45 was a blank part with a stepwise increase in allowance during milling (Fig. 4). A special groove on the surface of the blank part is designed to simulate intermittent cutting.

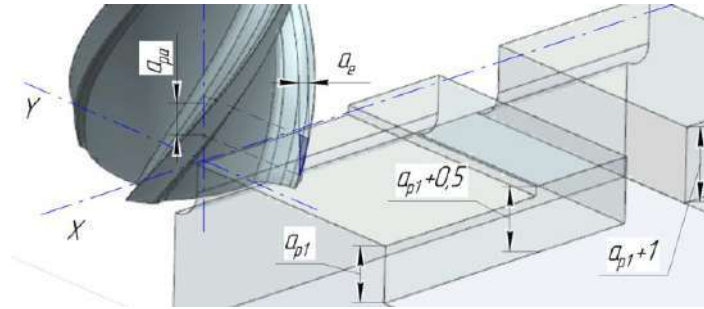


Fig. 4. Experimental sample.

The cutting conditions for the experiments are given in Table 1.

Table 1. Technological cutting parameters for material Steel 45.

Cutting speed 50 m/min			
№ exp.	F, mm/tooth	$A_p$ , mm	$A_e$ , mm
1			0,2
2	0,05	2	0,3
3			0,4

4.2. Experiment results

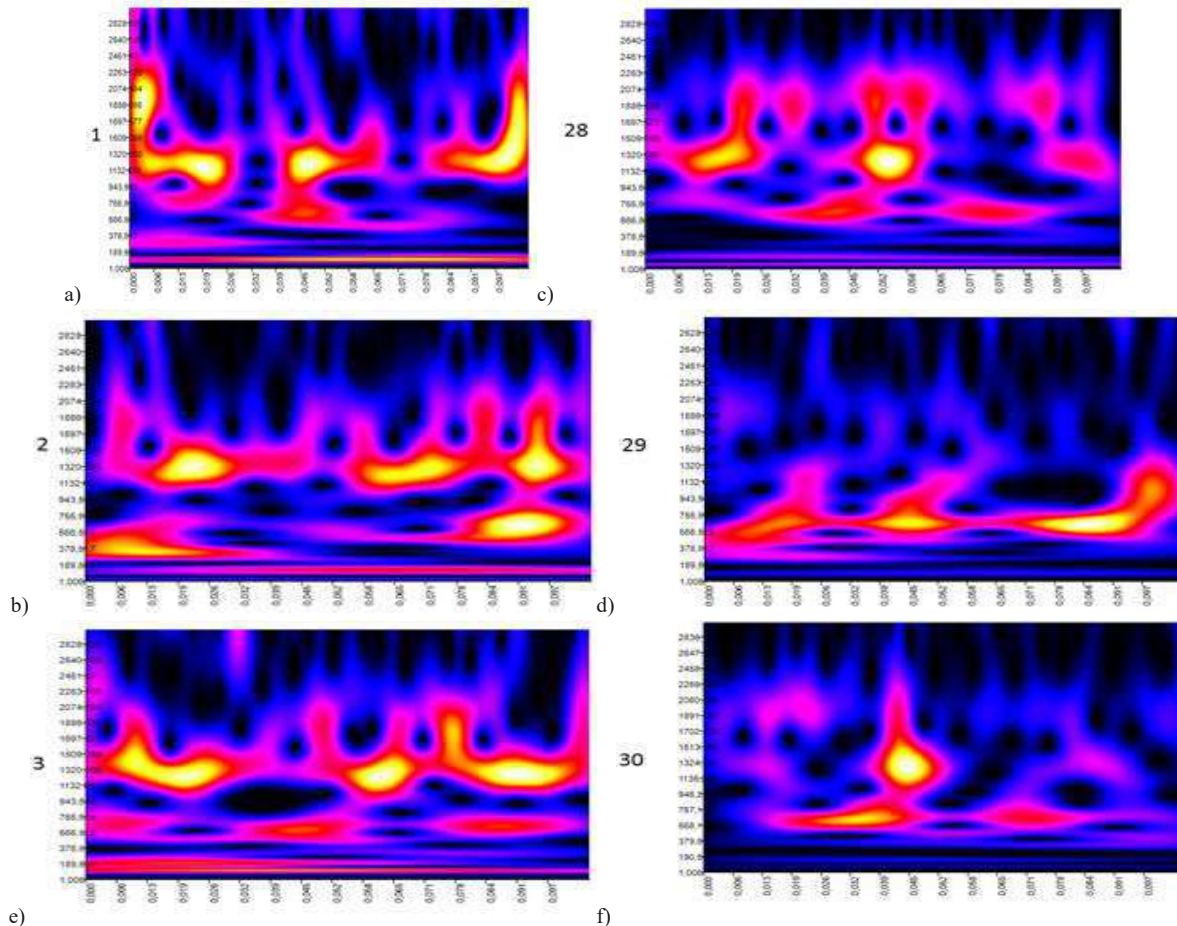


Fig.5. Wavelet spectrum of analyzed signals.

We use six different AE signals to analyze the cutting process with a multi-tooth tool. The signals denoted by the numbers 1, 2, 3 and 28, 29, 30 correspond to the regimes of Table 1 and are obtained by examining the new tool (a, b, c) and the worn tool (r, d, e). Fig. 5 shows the wavelet spectrum calculated by (9), where the X-axis of the wavelet spectrum graph represents the time in seconds, and the Y-axis represents the frequency in rad/s. The larger the value of the spectrum is, the lighter the pattern is.

Fig. 6 shows the scalogramms of the analyzed signals, which were obtained on the basis of the wavelet spectrum by (10).

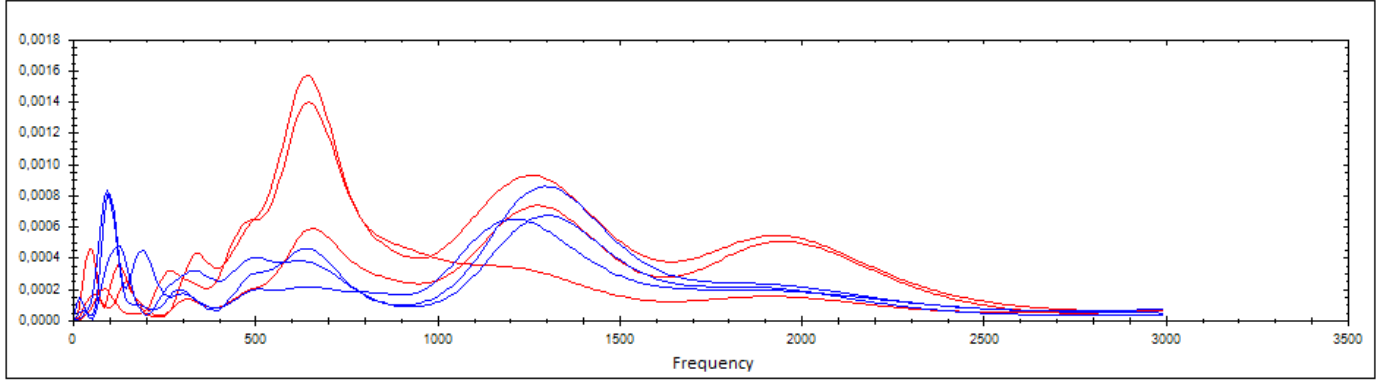


Fig.6. Scalogramms of analyzed signals.

The blue color shows the scalogramms of the signals corresponding to the state of the new tool, and the red one shows the worn tool.

The analysis of scalogramm of acoustic signal shows that it is possible to distinguish 3 characteristic maxima localized in the following frequency bands (in rad/s):  $\Delta\omega_{low} = 550 - 750$ ,  $\Delta\omega_{mid} = 1200 - 1500$ ,  $\Delta\omega_{hi} = 1950 - 2100$ .

The values of local maximum were calculated by (11). Results are shown in Table 2.

Table 2. The area of local maximum of scalogramms.

Frequency bands of local maximum $\Delta\omega_{max}$ , rad/s	$S_{\Delta\omega_{max}}(t_0)$ - new tool			$S_{\Delta\omega_{max}}(t_d)$ - worn tool		
	Mode 1	Mode 2	Mode 3	Mode 1	Mode 2	Mode 3
550-750	0,06213	0,0823	0,13359	0,17766	0,62313	0,44226
1200-1500	0,44037	0,31129	0,54264	0,32205	0,09416	0,44105
1950-2100	0,10634	0,09448	0,12908	0,31057	0,09919	0,351
Total area of scalogramms $S_{\Sigma}$	0,62765	0,65945	0,77467	0,88101	0,80046	1,32121

The wear coefficient  $k_{\Delta\omega_{max}}$  for 3 modes are given in table 3.

Table 3. Wear coefficient values.

Frequency bands of local maximum, rad/s	Mode 1	Mode 2	Mode 3
$\Delta\omega_{low}$ 550-750	0,491	0,160	0,515
$\Delta\omega_{mid}$ 1200-1500	1,919	4,013	1,626
$\Delta\omega_{hi}$ 1950-2100	0,481	1,156	0,486

The results of analysis are presented in Table 3. These results make it possible to see the characteristic feature: in the low-frequency region (550-750 rad/s), as the tool wear,  $k_{\Delta\omega_{max}}$  decreases, and in the area of conditionally medium frequencies region (1200-1500 rad/s) – increases.

The revealed regularity helps to formulate the condition for the appearance of a critical wear value when machining with a multi-tooth tool:

$$\begin{cases} k_{\Delta\omega_{max}}(t) \leq k_{low}, & \Delta V_{max} = \Delta V_{low}, \\ k_{\Delta\omega_{max}}(t) \geq k_{mid}, & \Delta V_{max} = \Delta V_{mid}, \end{cases} \quad t < t_d, \quad (15)$$

where  $k_{low}, k_{mid}$  are the limit values of the wear identification coefficient for the low and medium frequency range, respectively.

In other words, as the cutting tool wear, the spectral density of the energy of the Morlet wavelet image in the low-frequency region  $\Delta\omega_{low}$  increases ( $k_{\Delta\omega_{max}}$  decreases), and in the medium frequencies region  $\Delta\omega_{mid}$  decreases ( $k_{\Delta\omega_{max}}$  increases).

## 5. Conclusion

A model of milling process based on Morlet decomposition of vibroacoustic signals were proposed. Analyzing of the wavelet scalogramms of the signal at various processing modes, we received stable frequency bands of local maxima: 550-750 rad/s,

1200-1500 rad/s and 1950-2100 rad/s. Authors obtained trends to change the spectral energy density at the tool wear for the first and second frequency bands. The cross-factor  $CF_{\Delta\omega_{max}}$  can serve a numerical characteristic of change of this trend. The cross-factor determined by the dependence (10) and equal to the ratio of the average spectral density of the signal energy in the frequency bands of the local maximum of the scalogram to the average spectral energy density throughout the frequency region of the scalogram resolution. To identify the wear we proposed a new coefficient  $k_{\Delta\omega_{max}}$  that equal to the ratio of the cross-factors of acoustic emission signals for a new and wear tool, respectively. The coefficient of the wear identification increases where the dimensional wear increases in low-frequency region. These coefficient decreases in medium frequencies region. The experimentally determined regularity of the change a new condition that helps to improve early wear detection of the cutting tool made it possible to formalize the tool wear model with criterial constraints on the dependence.

## References

- [1] Machining Data Handbook. Machinability Data Center, Cincinnati, OH, 1980.
- [2] Armarego EJA, Brown RH. The Machining of Metals. New Jersey: Prentice-Hall Inc., 1969.
- [3] Rahman M, Seah WKH, Teo TT. The machinability of Inconel 718. *Journal of Materials Processing Technology* 1997; 63: 199–204.
- [4] Shaw MC. *Metal Cutting Principles*. Oxford University Press, 2005.
- [5] Trent EM. *Metal Cutting*, second ed. London: Butterworths, 1984.
- [6] Astafyeva NM. Wavelet analysis: the basics of theory and examples of its application. *Success of physical sciences* 1996; 166(11): 1145–1170.
- [7] Vityazev VV. *Wavelet analysis of temporal series: manual*. Saint-Petersburg : Publishing house of Saint-Petersburg University, 2001; 58 p.
- [8] Dobeshi I. *Ten Lectures on Wavelets*. Izhevsk : SRC Regular and chaotic dynamics, 2001; 464 p.
- [9] Koronovsky AA, Khramov AE. *Continuous wavelet analysis and its applications*. M.: Fizmatlit, 2003; 176 p.
- [10] Mallat S. *Wavelets in the signal processing: translated from English*. M.: Mir, 2005; 671 p.
- [11] Axinte DA, Andrews P. Some considerations on tool wear and workpiece surface quality of holes finished by reaming or milling in a nickel base superalloy. *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture* 2007; 221: 591–603.
- [12] Axinte DA, Dewes RC. Surface integrity of hot work tool steel after highspeed milling-experimental data and empirical models. *Journal of Materials Processing Technology* 2002; 127: 325–335.
- [13] Axinte DA, Gindy N, Fox K, Unanue I. Process monitoring to assist the workpiece surface quality in machining. *International Journal of Machine Tools and Manufacture* 2004; 44: 1091–1108.
- [14] Beggan C, Woulfe M, Young P, Byrne G. Using acoustic emission to predict surface quality, *International Journal of Advanced Manufacturing Technology* 1999; 15: 737–742.
- [15] Mantle AL, Aspinwall DK. Surface integrity of a high speed milled gamma titanium aluminide, *Journal of Materials Processing Technology* 2001; 143–150.
- [16] Sharman ARC, Aspinwall DK, Dewes RC, Bowen P. Workpiece surface integrity considerations when finish turning gamma titanium aluminide. *Wear* 2001; 249: 473–481.
- [17] Choudhury IA, El-Baradie MA. Machinability assessment of Inconel 718 by factorial design of experiment coupled with response surface methodology. *Journal of Materials Processing Technology* 1999; 95: 30–39.
- [18] Everson CE, Cheraghi SH. Application of acoustic emission for precision drilling process monitoring. *International Journal of Machine Tools and Manufacture* 1999; 39: 371–387.
- [19] Axinte D, Axinte M, Tannock JD. A multicriteria model for cutting fluid evaluation, *Proceedings of the Institution of Mechanical Engineers. Journal of Engineering Manufacture* 2003; 217: 1341–1353.
- [20] Axinte D, Dewes R, Ng E, Sage C, Soo S. The influence of cutter orientation and workpiece angle on machinability when high-speed milling Inconel 718 under finishing conditions. *International Journal of Machine Tools and Manufacture* 2007; 47: 1839–1846.
- [21] Toenshoff HK, Ianasaki I. *Sensors in Manufacturing*. Wiley-VCH Verlag GmbH, Weinheim, 2001.
- [22] Menon AK, Boutaghou Z-E. Time–frequency analysis of tribological systems. Part II: tribology of head-disk interactions. *Tribology International* 1998; 31: 511–518.
- [23] Menon AK, Boutaghou Z-E. Time–frequency analysis of tribological systems. Part I: implementation and interpretation. *Tribology International* 1998; 31: 501–510.
- [24] Cohen L. *Time–Frequency Analysis*. New Jersey: Prentice-Hall, 1995.
- [25] Lee SU, Robb D, Besant C. The directional Choi–Williams distribution for the analysis of rotor-vibration signals. *Mechanical Systems and Signal Processing* 2001; 15: 789–811.
- [26] Wigner EP. On the quantum correlation for thermodynamic equilibrium. *Physics Review* 1932; 40: 749–759.
- [27] W PR, Hammond JK. The analysis of non-stationary signals using time–frequency methods. *Journal of Sound and Vibration*, 1996.
- [28] Choi H-I, Williams J. Improved time–frequency representation of multicomponent signals using exponential kernels. *IEEE/ASME Transactions on Acoustics, Speech and Signal processing* 1989; 37: 862–871.
- [29] Khvostikov AS, Schetinina VS. Diagnosis of cutting process by applying Wavelet – analysis of acoustic emission signal. *Digital signal processing* 2007; 4: 40–43.
- [30] Yunxin Zhao LEA, Robert J, Marks II. The use of cone-shaped kernels for generalized time–frequency representations of nonstationary signals. *IEEE/ASME Transactions on Acoustics, Speech and Signal processing* 1990; 38: 1084–1091.
- [31] Weston RH. A formant detection system in which signal coding properties of a neuron network are used. *Journal of Sound and Vibration* 1975; 40:191–217.
- [32] Sidorov AS. *Monitoring and forecasting tool wear in mechatronic machine systems*. Abstract of dissertation for the degree of Candidate of Technical Sciences. Ufa, 2007.
- [33] Pechenin VA et al. Method of controlling cutting tool wear based on signal analysis of acoustic emission for milling. *Dynamics and Vibroacoustics of Machines (DVM2016)*.
- [34] Ramakrishna RPK, Prasad P, Srinivasa PP, Shantha V. Acoustic emission technique as a means for monitoring single point cutting tool wear, 2000.
- [35] Marinescu D. Axinte A time–frequency acoustic emission-based monitoring technique to identify workpiece surface malfunctions in milling with multiple teeth cutting simultaneously. *International Journal of Machine Tools & Manufacture* 2009; 49: 53–65.
- [36] Richard Y, Chiou A, Steven Y. Liang b Analysis of acoustic emission in chatter vibration with tool wear effect in turning. *International Journal of Machine Tools & Manufacture* 2000; 40.
- [37] Postnikov EB. Wavelet decomposition with Morlet wavelet: calculation method, based on solution of diffusive differential equations. *Computer – eaided research and modelling* 2009; 1(1): 5–12.
- [38] Yakovlev AN. *Introduction to wavelet decomposition: Manual*. Novosibirsk: Publishing House NSTU, 2003; 104 p.

# Intermediate asymptotic behavior of the stress and damage fields in the vicinity of the mixed-mode crack tip under creep regime

L. Stepanova<sup>1</sup>, E. Mironova<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The creep crack problems in damaged materials under mixed mode loading (Mode I and Mode II loading) in the framework of creep-damage coupled formulation are considered. The class of the self-similar solutions to the plane creep crack problems in a damaged medium under mixed-mode loading is given. With the similarity variable and the self-similar representation of the solution for a power-law creeping material and the Kachanov-Rabotnov power-law damage evolution equation the near crack-tip stresses, creep strain rates and continuity distributions for plane stress and plane strain conditions are obtained. The similarity solutions are based on the hypothesis of the existence of the completely damaged zone near the crack tip. It is shown that the asymptotic analysis of the near crack-tip fields gives rise to the nonlinear eigenvalue problems. The technique permitting to find all the eigenvalues numerically is proposed and numerical solutions of the nonlinear eigenvalue problems arising from the mixed-mode crack problems in a power-law medium under plane stress conditions are obtained. Using the approach developed the eigenvalues different from the eigenvalues corresponding to the Hutchinson-Rice-Rosengren (HRR) problem are found. The angular distributions of the stress and the continuity fields are selected as the crack tip fields of interest. Having obtained the eigenspectra and eigensolutions the geometry of the completely damaged zone in the vicinity of the crack tip is found for all values of the mixity parameter.

*Keywords:* damage parameter; continuity parameter; stress-strain fields near the crack tip; mixed-mode loading; asymptotic solution; similarity variable, self-similar solution; creep-damage coupling; nonlinear eigenvalue problems; eigenvalue spectrum

---

## 1. Introduction

Important advances in creep damage models for crack growth analysis have been made in the last two decades as scientists and engineers strive to imbue continuum-based models with more realistic details at microstructure damage mechanisms in the creep process [1-16]. Such damage models can be also found in [17-20]. Knowledge of stress, strain and displacement fields in the vicinity of the crack tip under mixed-mode loading conditions is important for the justification of fracture mechanics criteria and has attracted considerable attention nowadays [2-16]. Asymptotic analysis of the mechanical fields in front of stationary and propagating cracks facilitate the understanding of the mechanical and physical state in front of crack tips and they enable prediction of crack growth and failure. Furthermore, together with the stress, strain and displacement fields in the vicinity of the crack tip the damage distribution around the crack tip is a question of special attention [4,6,9,19,20]. Damage field around a crack tip essentially affects the surrounding stress field, and hence governs the crack extension behavior in the material. This effect of the damage field is an important problem either in the discussion of stability and convergence in crack extension analysis. So far mainly crack problems for the pure opening mode I at symmetrical loading have been thoroughly treated [6], [14]. The corresponding fracture criteria have been obtained on the assumption that the crack continues to extend along its original line (two-dimensional case) or plane (three-dimensional case) in a straightforward manner on the ligament. Nowadays the analysis of mixed-mode loading of cracked structures in nonlinear materials is of particular interest. In engineering practice, there are plenty of examples and reasons leading to mixed-mode loading of cracked structures when mode I is superimposed by mode II and/or III, the symmetry (or antisymmetry) is violated and the situation is related to mixed-mode loading [7]. The type of loading on a structure (tension, shear, bending, torsion) can also change during service. For a crack this results in an alteration of opening mode I, II and III which is why the study of mixed-mode loads is of particular importance [7-9,14,24]. In linear fracture mechanics the principle of superposition allows to obtain solutions for mixed mode I/II crack problems whereas in nonlinear fracture mechanics many questions are still open [9-22]. Analysis of the near crack-tip fields in power-law hardening (or power-law creeping) damaged materials under mixed-mode loading results in new nonlinear eigenvalue problems in which the whole spectrum of the eigenvalues and orders of stress singularity have to be determined [20-23]. For instance, in [21] asymptotic stress, strain, and displacement distributions in the vicinity of the mixed-mode crack for the stress-state sensitive elastic materials are considered and a nontrivial solution with the eigenvalue  $s = 1$  in the displacement series is found explicitly. Investigation of the asymptotic behaviour of the stress, strain, and displacements in the vicinity of a mixed-mode crack in the stress-state sensitive materials leads to conclusion that the traditional approaches, such as the superposition of the solutions as well as the assumptions for the symmetrical or antisymmetrical stress distributions can not be used. Therefore, in order to obtain the crack tip fields it is necessary to solve a new eigenvalue problem which in general case can be nonlinear. Nevertheless, in this work, a nontrivial solution with  $s = 1$  in the displacement series is found explicitly that demonstrates such specific features as the volume change under the condition of the remote shear loading and so on. It is shown [21] that the stronger the material stress-state-sensitivity is the stronger the resulting stress, strain and displacement fields deviate from the linear elastic ones. In [22] the stress field in the closest vicinity of a sharp material inclusion tip is characterized by 1 or 2 singular exponents. The exponents are calculated as an eigenvalue problem. The stress description by only one or two terms is not sufficient and leads to misleading results.

The objective of this study is to analyze the crack-tip fields in a damaged material under mixed-mode loading conditions and to consider the meso-mechanical effect of damage on the stress-strain state near the crack tip. The method proposed has been applied to nonlinear eigenvalue problems arising from the problem of the determining the near crack-tip fields in the damaged



materials. In continuum damage mechanics [2,6,7,10,12,13,16], the damage state at an arbitrary point in the material is represented by a properly defined integrity (continuity) variable  $\psi(r, \theta)$ . The integrity parameter reaches its critical value at fracture. According to this notion, a crack in a fracture process can be modeled with the concept of a completely damaged zone in the vicinity of the crack tip. Namely a crack can be represented by a region where the integrity state has attained to its critical state  $\psi = \psi_{cr}$ , i.e., by the completely damaged zone (CDZ) [6]. Then the development of the crack and its preceding damage can be elucidated by analyzing the local states of stress, strain and damage. The CDZ may be interpreted as the zone of critical decrease in the effective area due to damage development. Inside the completely damaged zone the damage involved reaches its critical value (for instance, the damage parameter reaches unity) and a complete fracture failure occurs. In view of material damage stresses are relaxed to vanishing [6,9,24,25]. Therefore, one can assume that the stress components in the CDZ equal zero. Outside the zone damage alters the stress distribution substantially compared to the corresponding non-damaging material. Well outside the CDZ the continuity parameter is equal to 1. Asymptotic remote boundary conditions are the asymptotic approaching the HRR solution. Dimensional analysis of the system formulated shows that the damage mechanics equations must have similarity solutions [9,24]. The present paper extends works [19,20,23,24,26] and constructs the asymptotic stress and continuity fields for stationary mixed-mode crack in damaged media under creep conditions.

## 2. Mixed mode crack problem: mathematical statement of the problem and fundamental equations

Static mixed mode crack problems under plane stress and plane strain conditions under creep regime are considered. The equilibrium equations and compatibility condition in the polar coordinate system can, respectively, be written as

$$r\sigma_{rr,r} + \sigma_{r\theta,\theta} + \sigma_{rr} - \sigma_{\theta\theta} = 0, \quad \sigma_{\theta\theta} + r\sigma_{r\theta,r} + 2\sigma_{r\theta} = 0, \quad (1)$$

$$2(r\varepsilon_{r\theta,\theta})_{,r} = \varepsilon_{rr,\theta\theta} - r\varepsilon_{rr,r} + r(r\varepsilon_{\theta\theta})_{,rr}. \quad (2)$$

The constitutive equations are described by the power law stress-strain rate relations incorporating the damage state parameter  $\omega = 1 - \psi$  and the creep strain rate is defined as follows

$$\varepsilon_{ij} = 3(\sigma_e / \psi)^{n-1} s_{ij} / (2\psi) \quad (3)$$

where  $s_{ij} = \sigma_{ij} - \sigma_{kk}\delta_{ij}/3$  are the deviatoric stress tensor components;  $B, n$  are material constants which control secondary creep behavior and can be determined from a log-log plot of the creep strain rate vs the applied stress;  $\psi$  is an integrity (continuity) parameter ( $\psi = 1$  indicates no damage and  $\psi = 0$  corresponds to complete damage);  $\varepsilon_{ij}$  are the strain components which for the plane stress and plane strain conditions take the form:

$$\varepsilon_{rr} = B\sigma_e^{n-1}(2\sigma_{rr} - \sigma_{\theta\theta})/(2\psi^n), \quad \varepsilon_{\theta\theta} = B\sigma_e^{n-1}(2\sigma_{\theta\theta} - \sigma_{rr})/(2\psi^n), \quad \varepsilon_{r\theta} = 3B\sigma_e^{n-1}\sigma_{r\theta}/(2\psi^n), \quad (4)$$

$$\varepsilon_{rr} = 3B\sigma_e^{n-1}(\sigma_{rr} - \sigma_{\theta\theta})/(4\psi^n), \quad \varepsilon_{\theta\theta} = 3B\sigma_e^{n-1}(\sigma_{\theta\theta} - \sigma_{rr})/(4\psi^n), \quad \varepsilon_{r\theta} = 3B\sigma_e^{n-1}\sigma_{r\theta}/(2\psi^n). \quad (5)$$

The von-Mises equivalent stress is expressed by  $\sigma_e = \sqrt{\sigma_{rr}^2 + \sigma_{\theta\theta}^2 - \sigma_{rr}\sigma_{\theta\theta} + 3\sigma_{r\theta}^2}$  and  $\sigma_e = (\sqrt{3}/2)\sqrt{(\sigma_{rr} - \sigma_{\theta\theta})^2 + 4\sigma_{r\theta}^2}$  for plane stress and plane strain conditions respectively. The constitutive model (3) is the phenomenological model of Kachanov and Rabotnov widely employed in creep damage theory and in damage analysis of high temperature structures [1, 6, 18, 26, 27]. The material parameters pertinent to equations (2) for copper, the aluminium alloy, ferritic steels obtained from creep curves are given by Riedel in [18]. By noting that the creep damage is brought about by the development of microscopic voids in creep process, L.M. Kachanov represented the damage state by a scalar integrity variable  $\psi$  ( $0 \leq \psi \leq 1$ ) where  $\psi = 1$  and  $\psi = 0$  signify the initial undamaged state and the final completely damaged state (or final fractured state), respectively [6,12,13,26]. L.M. Kachanov [27] described the damage development by means of an evolution equation

$$d\psi / dt = -A(\sigma_{eqv} / \psi)^m, \quad (6)$$

where  $A$  and  $m$  are material constants which control tertiary creep behavior,  $\sigma_{eqv} = \alpha\sigma_1 + \beta\sigma_e + \gamma\sigma_{kk}$ ,  $\sigma_1$  is the maximum principal stress,  $\sigma_{kk}$  is the hydrostatic stress;  $\alpha$  is the material constant, which describes the effect of the multi-axial stress state behavior of material and ranges from  $\alpha = 0$  (equivalent stress dominant) to  $\alpha = 1$  (maximum principal stress dominant). The accurate prediction of  $\alpha$  value plays an important role in the application of the multi-axial Kachanov-Rabotnov damage model. The solution of Eqs. (1) – (5) should satisfy the traditional traction free boundary conditions on the crack surfaces

$$\sigma_{r\theta}(r, \theta = \pm\pi) = 0, \quad \sigma_{\theta\theta}(r, \theta = \pm\pi) = 0. \quad (7)$$

The mixed-mode loading can be characterized in terms of the mixity parameter  $M^P$  [5,7,24] which is defined as

$$M^P = \frac{2}{\pi} \arctan \left( \lim_{r \rightarrow 0} \left| \frac{\sigma_{\theta\theta}(r, \theta = 0)}{\sigma_{r\theta}(r, \theta = 0)} \right| \right). \quad (8)$$

The mixity parameter  $M^P$  equals 0 for pure mode II; 1 for pure mode I, and  $0 < M^P < 1$  for different mixities of modes I and II. Thus, for combine-mode fracture the mixity parameter  $M^P$  completely specifies the near-crack-tip fields for a given value of the creep exponent.

### 3. Asymptotic solution

One can assume that the completely damaged zone in the vicinity of the crack tip evolves during the deformation of the cracked body. The asymptotic solution for the stress-strain fields and the continuity field can be found outside the completely damaged zone. As the continuity parameter reaches at infinity the value corresponding to the undamaged materials it is possible to realize the approach described below.

#### 3.1. Similarity solution

Dimensional analysis of the system formulated shows that the damage mechanics equations must have similarity solutions of the form [18]

$$\sigma_{ij}(r, \theta, t) = (At)^{-1/m} \hat{\sigma}_{ij}(R, \theta), \quad \psi(r, \theta, t) = \hat{\psi}(R, \theta) \quad (9)$$

where  $R = r(At)^{-(n+1)/m} BI_n / C^*$  is the similarity variable. It should be noted that the remote boundary conditions can be formulated in a more general form  $\sigma_{ij}(r \rightarrow \infty, \theta, t) = \tilde{C} r^s \bar{\sigma}_{ij}(\theta, n)$ , where the stress singularity exponent  $s$  is unknown and has to be determined as a part of solution,  $C$  is the amplitude of the stress field at infinity defined by the specimen configuration and loading conditions. For the power-law creep constitutive relations, the power-law damage evolution equation and the more general remote boundary conditions the self-similar variable  $R = r(At)^{1/(sm)} \tilde{C}$  can be introduced. After introducing the self-similar variable the equilibrium equations, the constitutive equations, the compatibility condition retain their forms, whereas the damage evolution equation becomes

$$R\hat{\psi}_{,R} = -sm(\hat{\sigma}_e / \hat{\psi})^m \quad (10)$$

(the superscript  $\hat{\cdot}$  is further omitted). By postulating the Airy stress function  $\chi(R, \theta)$  expressed in the polar coordinate system, the stress components state are expressed as:  $\sigma_{\theta\theta} = \chi_{,RR}$ ,  $\sigma_{RR} = \chi_{,R} / R + \chi_{,\theta\theta} / R^2$ ,  $\sigma_{R\theta} = -(\chi_{,\theta} / R)_{,R}$ .

The asymptotic solution outside the completely damaged zone ( $R \rightarrow \infty$ ) is sought in the form

$$\chi(R, \theta) = \sum_{j=0}^{\infty} R^{\lambda_j+1} f_j(\theta), \quad \psi(r, \theta) = 1 - \sum_{j=0}^{\infty} R^{\gamma_j} g_j(\theta). \quad (11)$$

In view of (11) the asymptotic presentation of the stress tensor components has the form

$$\sigma_{ij}(R, \theta) = \sum_{k=0}^{\infty} R^{\lambda_k-1} \sigma_{ij}^{(k)}(\theta), \quad \sigma_{RR}^{(k)}(\theta) = (\lambda_k + 1)f_k(\theta) + f_k''(\theta), \quad \sigma_{R\theta}^{(k)}(\theta) = -(\lambda_k + 1)f_k'(\theta), \quad \sigma_{\theta\theta}^{(k)}(\theta) = \lambda_k(\lambda_k + 1)f_k(\theta). \quad (12)$$

It can be shown that the asymptotic series expansion of the strain rate tensor components can be written as:

$$\varepsilon_{ij}(R, \theta) = \sum_{k=0}^{\infty} R^{(\lambda_0-1)(n+km)} \varepsilon_{ij}^{(k)}(\theta). \quad (13)$$

#### 3.2. Structure of the asymptotic solution. The leading term of the asymptotic governing equations

First consider the leading terms of the asymptotic expansions (11):  $\chi(R, \theta) = r^{\lambda_0+1} f_0(\theta)$ ,  $\psi = 1$  where  $\lambda_0$  is indeterminate exponent and  $f_0(\theta)$  is an indeterminate function of the polar angle, respectively. In view of the asymptotic presentation for the Airy stress potential (11) the asymptotic stress field at the crack tip is derived as follows  $\sigma_{ij}(R, \theta) = R^{\lambda_0-1} \sigma_{ij}(\theta)$ , where  $\lambda_0 - 1$  denotes the exponent representing the singularity of the stress field, and will be called the stress singularity exponent hereafter. According to Eq. 13 the asymptotic strain field as  $R \rightarrow \infty$  takes the form  $\varepsilon_{ij}(R, \theta) = BR^{(\lambda_0-1)n} \varepsilon_{ij}^{(0)}(\theta)$ . The compatibility condition (Eq. 2) results in the nonlinear forth-order ordinary differential equation (ODE) for the function  $f_0(\theta)$ :

$$\begin{aligned} & f_e^2 f_0'' \left\{ (n-1) \left[ (1-\lambda_0^2) f_0 + f_0'' \right]^2 + f_e^2 \right\} + (n-1)(n-3) \left\{ \left[ (1-\lambda_0^2) f_0 + f_0'' \right] \left[ (1-\lambda_0^2) f_0' + f_0''' \right] + 4\lambda_0^2 f_0' f_0'' \right\} \left[ (1-\lambda_0^2) f_0 + f_0'' \right] + \\ & + (n-1) f_e^2 \left\{ \left[ (1-\lambda_0^2) f_0' + f_0''' \right]^2 + \left[ (1-\lambda_0^2) f_0 + f_0'' \right] (1-\lambda_0^2) f_0'' + 4\lambda_0^2 (f_0''^2 + f_0' f_0''') \right\} \left[ (1-\lambda_0^2) f_0 + f_0'' \right] + \\ & + 2(n-1) f_e^2 \left\{ \left[ (1-\lambda_0^2) f_0 + f_0'' \right] \left[ (1-\lambda_0^2) f_0' + f_0''' \right] + 4\lambda_0^2 f_0' f_0'' \right\} \left[ (1-\lambda_0^2) f_0 + f_0'' \right] + \\ & + \tilde{N}_1 (n-1) f_e^2 \left\{ \left[ (1-\lambda_0^2) f_0 + f_0'' \right] \left[ (1-\lambda_0^2) f_0' + f_0''' \right] + 4\lambda_0^2 f_0' f_0'' \right\} f_0' + \\ & + C_1 f_e^4 f_0'' - C_2 f_e^4 \left[ (1-\lambda_0^2) f_0 + f_0'' \right] + f_e^4 (1-\lambda_0^2) f_0'' = 0, \end{aligned} \quad (14)$$

where the following notations are adopted  $f_e^2 = \left[ (1-\lambda_0^2) f_0 + f_0'' \right]^2 + 4\lambda_0^2 f_0'^2$ ,  $C_1 = 4\lambda_0 [(\lambda_0 - 1)n + 1]$ ,  $C_2 = (\lambda_0 - 1)n [(\lambda_0 - 1)n + 1]$  for plane strain conditions;



$$\begin{aligned}
 & f_e'' f_e^2 \left\{ (n-1)[(\lambda_0+1)(2-\lambda_0)f_0+2f_0''^2] / 2+2f_e^2 \right\} + 6[(\lambda_0-1)n+1] \lambda_0 \left\{ (n-1)f_e^2 h f_e' + f_e^4 f_e'' \right\} + (n-1)(n-3)h^2 \times \\
 & \times [(\lambda_0+1)(2-\lambda_0)f_0+2f_0''] + (n-1)f_e^2 [(\lambda_0+1)(\lambda_0+2)f_0+2f_0''] \left\{ [(\lambda_0+1)f_0'+f_0''^2] + [(\lambda+1)f+f''](\lambda+1)f'' + \right. \\
 & \left. + (\lambda_0+1)^2 \lambda_0^2 (f_0'^2 + f_0''^2) - (\lambda_0+1)^2 \lambda_0 f_0 f_0'' / 2 - [(\lambda_0+1)f_0'+f_0''](\lambda_0+1)\lambda_0 f_0'+f_e^4 (\lambda_0+1)(2-\lambda_0)f_0'' - \right. \\
 & \left. - [(\lambda+1)f+f''](\lambda+1)\lambda f'' + 3\lambda_0^2 (f_0''^2 + f_0' f_0'') \right\} + 2(n-1)f_e^2 h [(\lambda_0+1)(2-\lambda_0)f_0'+2f_0''] / 2 - \\
 & - (\lambda-1)n f_e^4 [(\lambda+1)(2-\lambda)f_0+2f_0''] + [(\lambda_0-1)n+1](\lambda_0-1)n f_e^4 [(\lambda_0+1)(2\lambda_0-1)f_0-f_0''] = 0,
 \end{aligned} \tag{15}$$

where the following notations are adopted

$$\begin{aligned}
 f_e &= \sqrt{[(\lambda_0+1)f_0+f_0'']^2 + (\lambda_0+1)^2 \lambda_0^2 f_0^2 - [(\lambda_0+1)f_0+f_0''](\lambda_0+1)\lambda_0 f_0 + 3\lambda_0^2 (f_0')^2}, \quad h = [(\lambda_0+1)f_0+f_0''] \times \\
 & \times [(\lambda_0+1)f_0'+f_0''] + (\lambda_0+1)^2 \lambda_0^2 f_0 f_0' - [(\lambda_0+1)f_0'+f_0''](\lambda_0+1)\lambda_0 f_0' / 2 - [(\lambda_0+1)f_0+f_0''](\lambda_0+1)\lambda_0 f_0' / 2 + 3\lambda_0^2 f_0' f_0''
 \end{aligned}$$

for plane stress conditions respectively. The boundary conditions follow from the traction-free boundary conditions:

$$f_0(\theta = \pm\pi) = 0, \quad f_0'(\theta = \pm\pi) = 0. \tag{16}$$

Hence the governing equations (1) – (6) are transformed into nonlinear eigenvalue problems of ordinary differential equations (ODE) with respect to the circumferential coordinate around the crack tip. Thus it is necessary to find the nontrivial solutions of (14) and (15) satisfying the boundary conditions (16).

#### 4. Nonlinear eigenvalue problems. Numerical results. Eigenspectra of nonlinear eigenvalue problems

The method is further developed in the present work to analyze nonlinear eigenvalue problems of ODE for equations (14) and (15). The stress singularity orders  $\lambda_0$  and the associated eigenfunctions  $f_0(\theta)$  for plane strain and plane stress conditions are obtained. In general, most of the existing methods have their own merits and are complementary, depending on the nature of the problems to be solved. To solve the two-point boundary value problems an efficient method to deal with eigenvalue problems of ODE is apparently needed. In this paper the eigenvalues of the nonlinear eigenvalue problems are obtained by the technique developed in [24]. The algorithm is intended for mixed mode loading when the eigenvalue is not known and should be determined as a part of the solution. The main idea is to find eigenvalues different from the classical eigenvalue corresponding to the HRR problem  $\lambda_0 = n / (n + 1)$ . To find new eigenvalues an additional requirement following from physical or mathematical considerations is needed. The new condition of continuity of the radial stress on the line extending the crack is used. The numerical results are given in Tables 1-4 for plane strain conditions and in Tables 5-8 for plane stress conditions respectively.

Table 1. Eigenvalues of the nonlinear eigenvalue problem for  $n = 2$  (plane strain conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.309869	-0.006883	-0.409577	0.746037	-0.292556
0.9	-0.309310	-0.007461	-0.344635	0.777310	-0.313217
0.8	-0.306884	-0.009687	-0.210585	0.786405	-0.327997
0.7	-0.302045	-0.015423	-0.089789	0.757415	-0.321367
0.6	-0.290875	-0.033625	0.019798	0.702499	-0.298084
0.5	-0.282566	-0.036275	0.108960	0.650499	-0.275722
0.4	-0.274875	-0.024113	0.186358	0.590347	-0.249031
0.3	-0.271672	-0.007071	0.260397	0.543846	-0.230473
0.2	-0.272741	0.007771	0.337598	0.500515	-0.215603
0.1	-0.275936	0.012592	0.442771	0.459684	-0.203360
0.05	-0.277383	0.008378	0.518376	0.437743	-0.196991

Table 2. Eigenvalues of the nonlinear eigenvalue problem for  $n = 3$  (plane strain conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.261580	0.165263	-0.783132	0.911946	-0.694200
0.9	-0.260520	0.160476	-0.817484	0.813416	-0.617651
0.8	-0.256052	0.144177	-0.838728	0.589120	-0.442762
0.7	-0.248870	0.126642	-0.831073	0.372380	-0.272113
0.6	-0.240700	0.109252	-0.814365	0.161493	-0.090100
0.5	-0.230512	0.082163	-0.787752	-0.101438	0.113079
0.4	-0.226700	0.083062	-0.782363	-0.155373	0.125205
0.3	-0.227669	0.095646	-0.801477	-0.213141	0.161948
0.2	-0.235040	0.104379	-0.822132	-0.280236	0.210731
0.1	-0.250783	0.093861	-0.837159	-0.369869	0.280496
0.05	-0.262025	0.067491	-0.805601	-0.431242	0.329927

Table 3. Eigenvalues of the nonlinear eigenvalue problem for  $n = 4$  (plane strain conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.235100	0.236297	-0.928023	1.014212	-0.898791
0.9	-0.234510	0.234943	-0.965785	0.912209	-0.808333
0.8	-0.227530	0.212686	-0.974785	0.667675	-0.587889
0.7	-0.220460	0.191738	-0.955971	0.440489	-0.384975
0.6	-0.214380	0.171747	-0.928916	0.232746	-0.200054
0.5	-0.206813	0.139765	-0.888099	-0.072201	0.0926523
0.4	-0.204496	0.132479	-0.867790	-0.132744	0.116858
0.3	-0.206710	0.142094	-0.879023	-0.198359	0.172059
0.2	-0.215000	0.149162	-0.898657	-0.275789	0.240456
0.1	-0.233547	0.133510	-0.937100	-0.380728	0.337187
0.05	-0.249140	0.101767	-0.947125	-0.457725	0.410773

Table 4. Eigenvalues of the nonlinear eigenvalue problem for  $n = 5$  (plane strain conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.221500	0.285827	-1.027835	1.074981	-1.024605
0.9	-0.218248	0.276260	-1.055626	0.958976	-0.911524
0.8	-0.210625	0.252063	-1.058479	0.706036	-0.6667539
0.7	-0.204759	0.230065	-1.034373	0.474972	-0.446257
0.6	-0.200055	0.208164	-1.00100	0.266302	-0.248949
0.5	-0.194290	0.175457	-0.954079	-0.053984	0.0932645
0.4	-0.192320	0.160817	-0.921897	-0.120061	0.112156
0.3	-0.194812	0.168460	-0.927342	-0.190532	0.177562
0.2	-0.202342	0.173107	-0.940975	-0.274124	0.257060
0.1	-0.221158	0.155956	-0.989844	-0.387606	0.369339
0.05	-0.237872	0.121203	-1.040272	-0.473791	0.457832

Table 5. Eigenvalues of the nonlinear eigenvalue problem for  $n = 2$  (plane stress conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.302240	-0.178386	-0.327269	0.3355941	0.227941
0.9	-0.300320	-0.178070	-0.366381	0.2575700	0.292668
0.8	-0.286090	-0.192771	-0.407730	-0.088471	0.769239
0.7	-0.267890	-0.215107	-0.429407	-0.202419	0.293261
0.6	-0.260930	-0.209033	-0.466896	-0.244312	0.207239
0.5	-0.252332	-0.193741	-0.519713	-0.270032	0.157534
0.4	-0.243698	-0.168133	-0.584743	-0.285889	0.120618
0.3	-0.237019	-0.136613	-0.661107	-0.302160	0.092347
0.2	-0.232479	-0.096929	-0.721807	-0.315939	0.067278
0.1	-0.229872	-0.051085	-0.763118	-0.331096	0.044413
0.05	-0.229234	-0.026058	-0.773639	-0.338664	0.033298

Table 6. Eigenvalues of the nonlinear eigenvalue problem for  $n = 4$  (plane stress conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.25900	0.046297	-0.554192	0.728468	-0.356792
0.9	-0.25560	0.044068	-0.588519	0.643051	-0.313277
0.8	-0.24450	0.038444	-0.620481	0.448631	-0.213325
0.7	-0.23350	0.033607	-0.630460	0.251112	-0.103504
0.6	-0.22020	0.012148	-0.623434	-0.120807	0.152059
0.5	-0.21079	0.009305	-0.626547	-0.170003	0.103454
0.4	-0.20527	0.023364	-0.636587	-0.207179	0.106370
0.3	-0.20303	0.043891	-0.663582	-0.248072	0.119894
0.2	-0.20573	0.063168	-0.697544	-0.292794	0.138533
0.1	-0.21570	0.069677	-0.750661	-0.349948	0.165542
0.05	-0.22414	0.056261	-0.777450	-0.387722	0.184522

Table 7. Eigenvalues of the nonlinear eigenvalue problem for  $n = 6$  (plane stress conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.23620	0.092284	-0.640557	0.802232	-0.504505
0.9	-0.23000	0.088445	-0.672059	0.707014	-0.442238
0.8	-0.21800	0.084838	-0.700964	0.505557	-0.312884
0.7	-0.20880	0.082946	-0.707758	0.314326	-0.192580
0.6	-0.19930	0.069616	-0.698358	-0.075083	0.311685
0.5	-0.19078	0.059067	-0.688624	-0.137724	0.088869
0.4	-0.18870	0.071683	-0.691122	-0.179959	0.109436
0.3	-0.19332	0.091028	-0.713498	-0.227876	0.138206
0.2	-0.20619	0.106166	-0.748785	-0.282897	0.173351
0.1	-0.23041	0.104842	-0.833112	-0.356879	0.223327
0.05	-0.24985	0.084782	-0.912969	-0.410296	0.261073

Table 8. Eigenvalues of the nonlinear eigenvalue problem for  $n = 8$  (plane stress conditions).

Mixity parameter	$\lambda_0$	$f_0''(0)$	$f_0'''(0)$	$f_0''(-\pi)$	$f_0'''(-\pi)$
0.95	-0.224000	0.113333	-0.684626	0.831308	-0.572658
0.9	-0.216700	0.109593	-0.715353	0.732220	-0.501304
0.8	-0.205400	0.107456	-0.743801	0.528789	-0.358568
0.7	-0.197400	0.107175	-0.749584	0.337838	-0.227508
0.6	-0.189810	0.098453	-0.740423	0.126048	-0.079175
0.5	-0.181760	0.082367	-0.723195	-0.123024	0.083357
0.4	-0.182300	0.093150	-0.723228	-0.167982	0.111747
0.3	-0.191700	0.110578	-0.746282	-0.219157	0.146879
0.2	-0.209870	0.122860	-0.785793	-0.278818	0.189787
0.1	-0.238320	0.118773	-0.892595	-0.359530	0.250648
0.05	-0.261703	0.97069	-1.035152	-0.419102	0.297851

### 5. Results and Discussion. Shape of the completely damaged zone in the vicinity of the crack tip

Having obtained the angular distributions of the stress, strain and continuity fields one can determine the geometry of the completely damaged zone modeled in the vicinity of the crack tip and given by multi-parameter expansion for the continuity:

$$\begin{aligned}
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) = 0, \\
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) - R^{\gamma_1} g_1(\theta) = 0, \\
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) - R^{\gamma_1} g_1(\theta) - R^{\gamma_2} g_2(\theta) = 0, \\
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) - R^{\gamma_1} g_1(\theta) - R^{\gamma_2} g_2(\theta) - R^{\gamma_3} g_3(\theta) = 0, \\
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) - R^{\gamma_1} g_1(\theta) - R^{\gamma_2} g_2(\theta) - R^{\gamma_3} g_3(\theta) - R^{\gamma_4} g_4(\theta) = 0, \\
 \psi(R, \theta) &= 1 - R^{\gamma_0} g_0(\theta) - R^{\gamma_1} g_1(\theta) - R^{\gamma_2} g_2(\theta) - R^{\gamma_3} g_3(\theta) - R^{\gamma_4} g_4(\theta) - R^{\gamma_5} g_5(\theta) = 0.
 \end{aligned}
 \tag{17}$$

One can compare the boundaries of the CDZ given by the two-term, three-term, four-term, five-term and six-term asymptotic expansions of the integrity (continuity) parameter (Eq. 17). It is turned out that if the asymptotic remote boundary condition is postulated as the condition of the asymptotic approaching the HRR-field then the shapes of the CDZ given by the two-term asymptotic expansion and three-term asymptotic expansions differ essentially from each other. The new stress asymptotic behavior results in the contours of the CDZ which converge to the limit contour shown in Figs. 1-7. The new far field stress asymptotic can be interpreted as the intermediate stress asymptotics valid for times and distances at which effects of the initial and boundary conditions on the stress and damage distributions are lost. The geometry of the completely damage zone for different values of the mixity parameter is shown in Figs. 1-5 where  $k = 1, 2, 3, 4, 5$  is designed the boundary of the CDZ built by the use of the  $k + 1$  - term asymptotic expansion of continuity. The red line shows the boundary of the CDZ obtained by the two-term asymptotic expansion of the integrity parameter whereas the blue line shows the boundary of the CDZ obtained by the three-term asymptotic expansion of the integrity parameter. The green line shows the boundary of the CDZ obtained by the four-term asymptotic expansion of the integrity parameter. From Figs. 1-7 it can be seen that the boundary of the CDZ determined by the use of the  $k + 1$  - term asymptotic expansion of continuity is very close to the boundary built by the  $k$  - term asymptotic expansion of the continuity parameter whereas the HRR stress field results in the boundary of the CDZ given by the two-term expansion which differs substantially from the boundary of the CDZ given by the three-term asymptotic expansion of the integrity parameter by the form and dimensions.

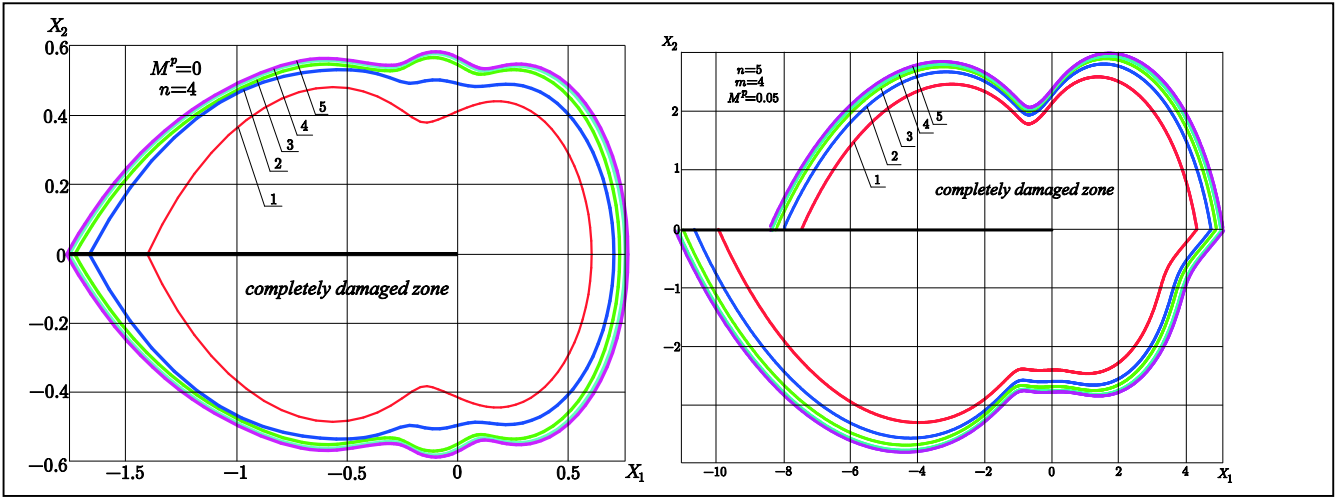


Fig.1. Geometry of the completely damaged zone in the vicinity of the crack tip for  $M^p = 0$  and  $M^p = 0.05$ .

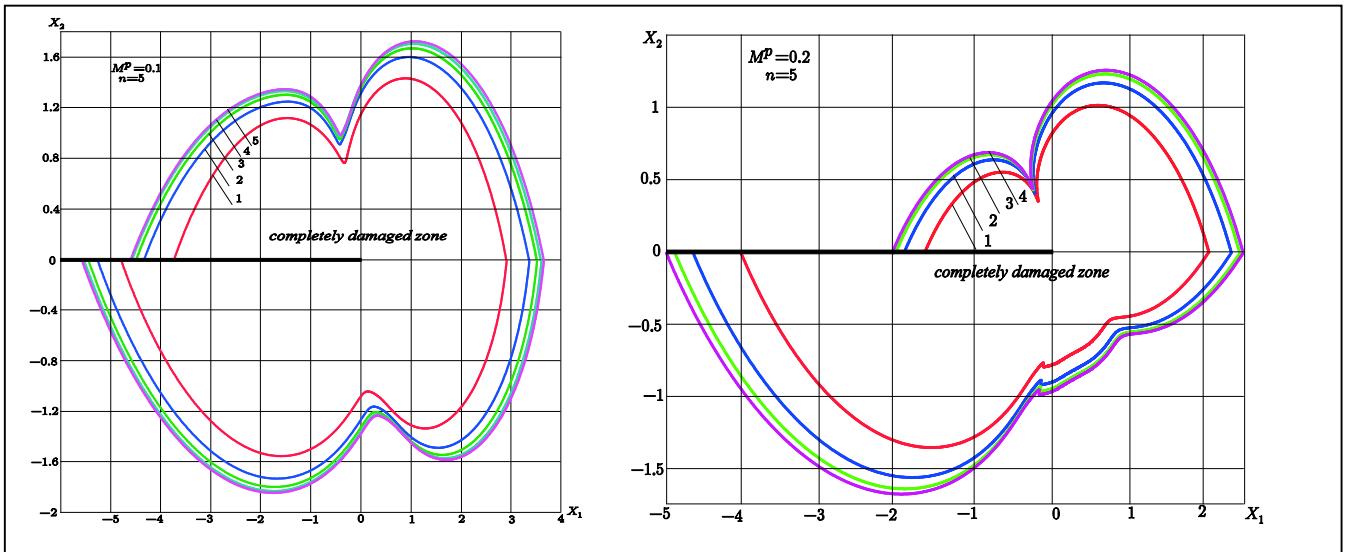


Fig. 2. The geometry of the totally damaged zone near the crack tip under mixed mode loading for  $M^p = 0.1$  and  $M^p = 0.2$ .

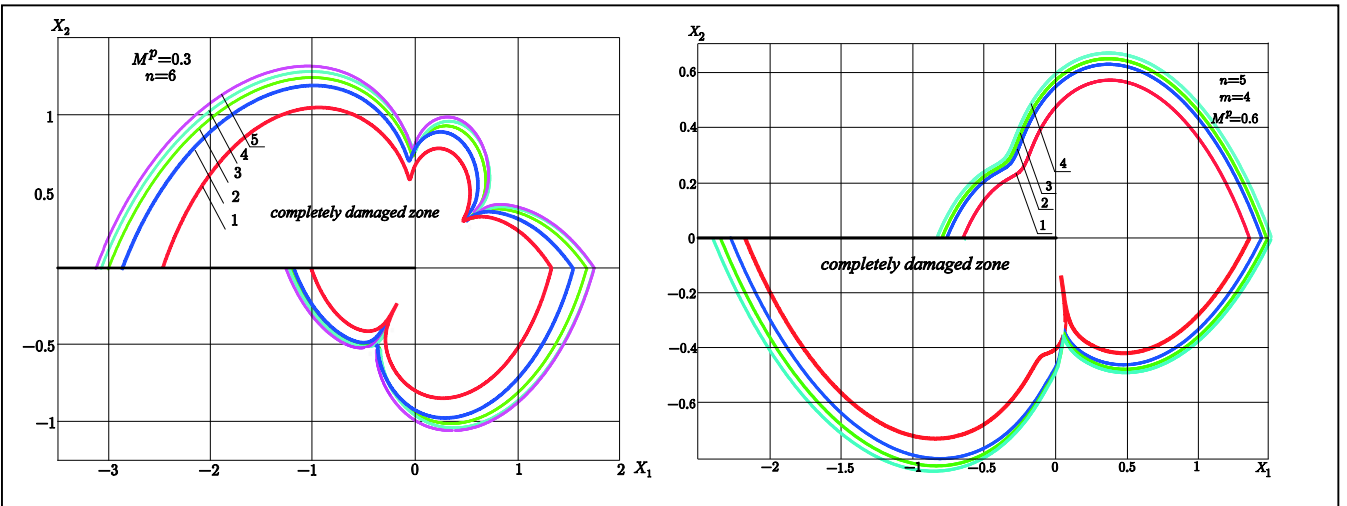


Fig.3. Geometry of the completely damaged zone in the vicinity of the crack tip under mixed mode loading for  $M^p = 0$  and  $M^p = 0.05$  (plane strain conditions).

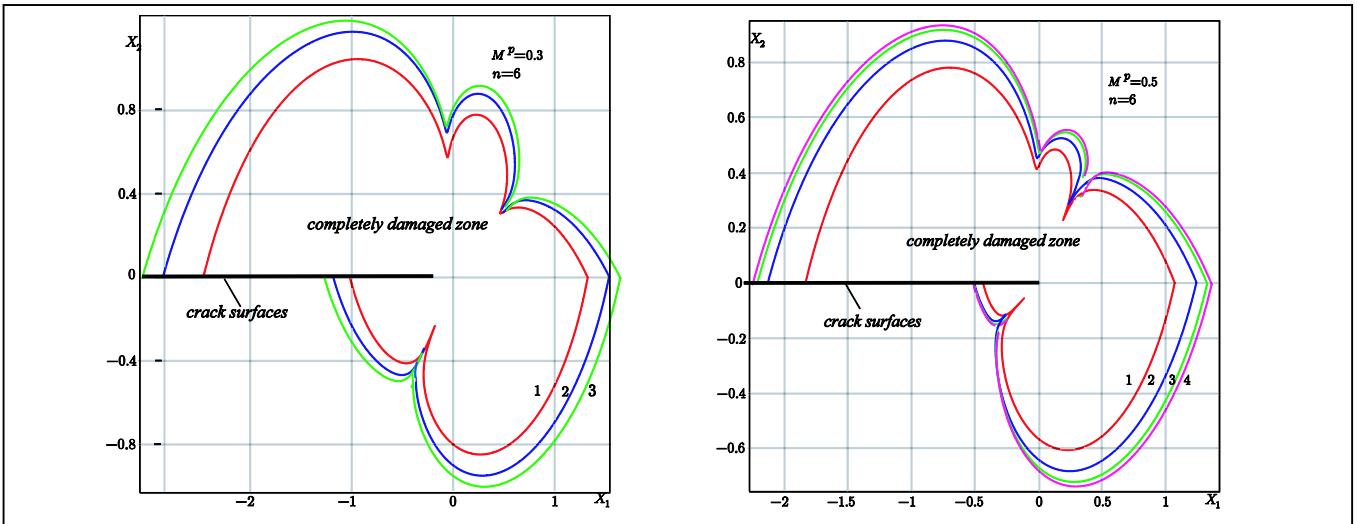
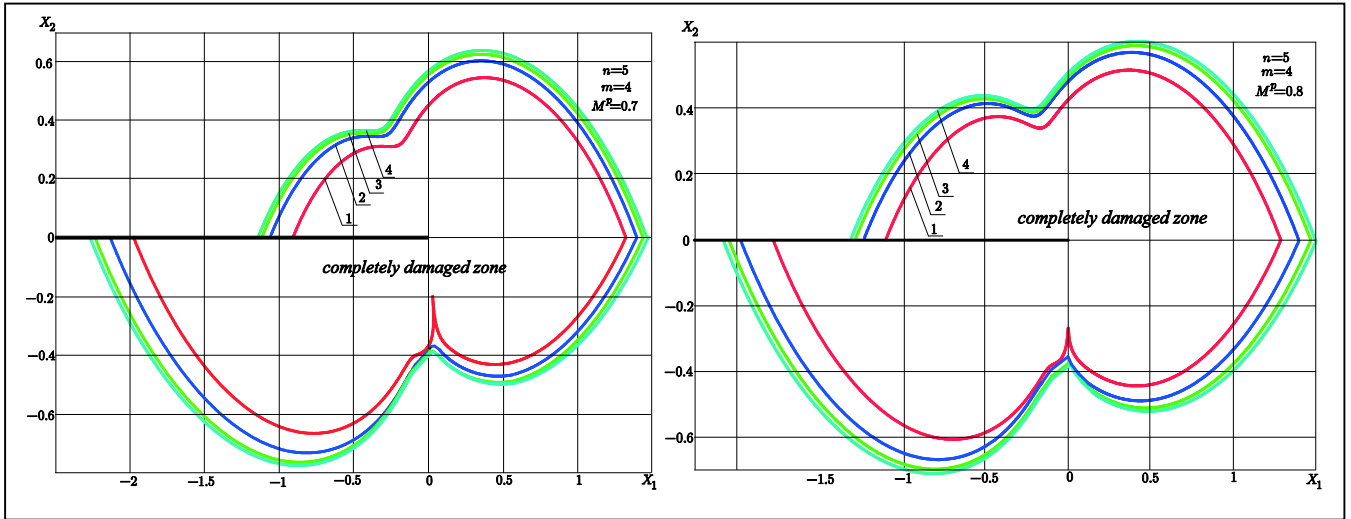


Fig. 4. Geometry of the completely damaged zone in the vicinity of the crack tip.

Fig. 5. Geometry of the completely damaged zone in the vicinity of the crack tip under mixed mode loading for or  $M^p = 0.9$  and  $M^p = 1$  (plane strain conditions).

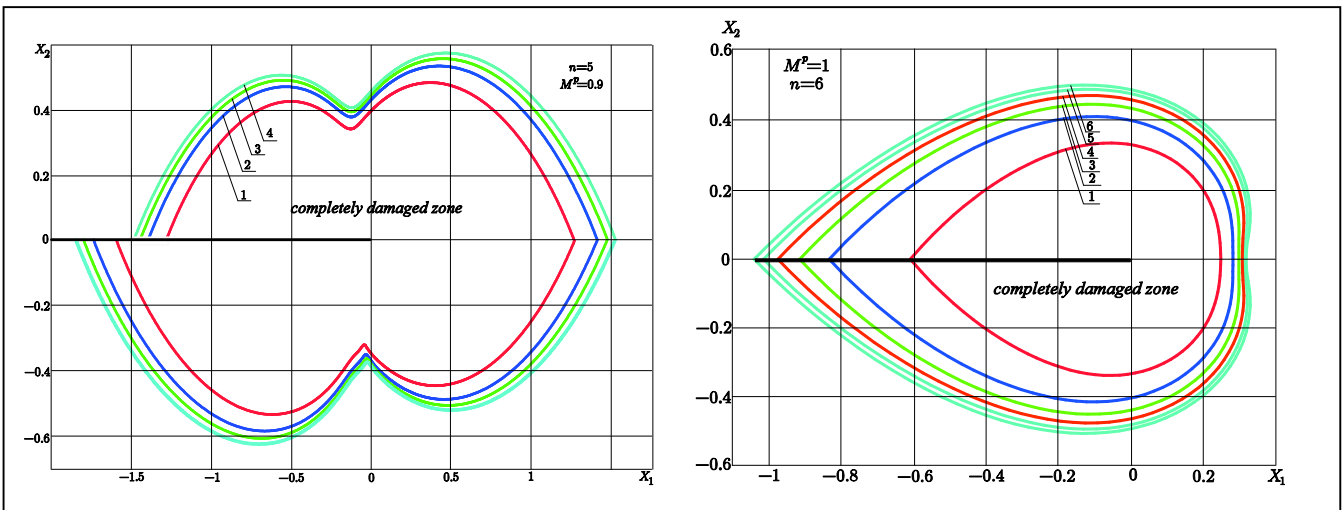
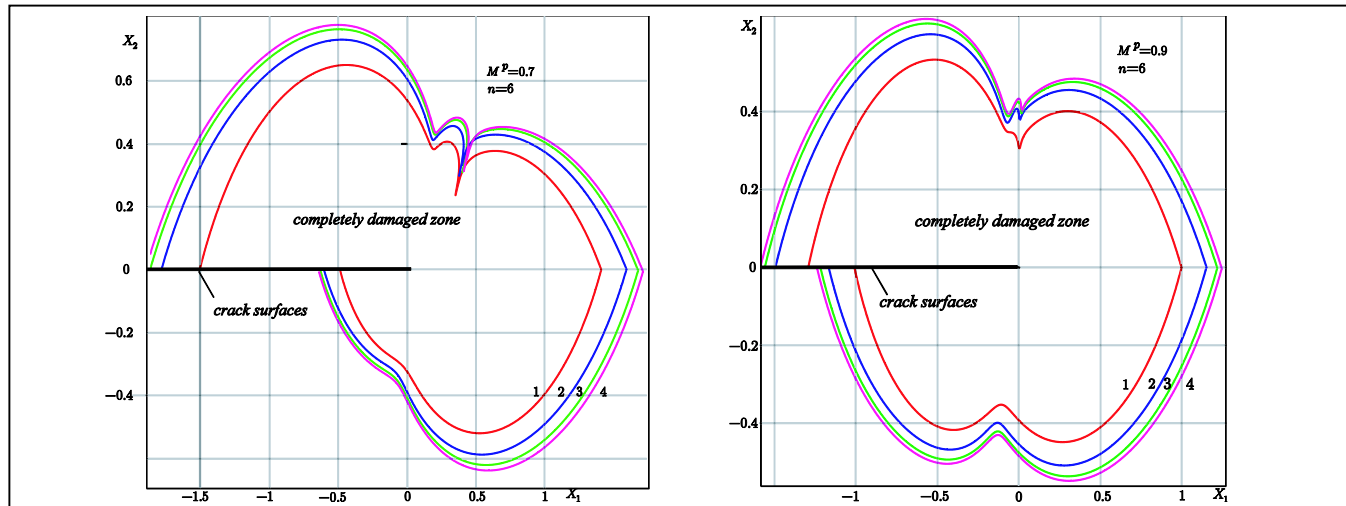


Fig. 6. Geometry of the completely damaged zone in the vicinity of the crack tip under mixed mode loading for  $M^p = 0.3$  and  $M^p = 0.5$ .

## 6. Conclusion

Asymptotic crack-tip fields in damaged materials are developed for a stationary plane stress and plane strain crack under mixed mode loading conditions in a full range of the mixity parameter varying from the value corresponding to pure Mode I loading to pure Mode II loading. The asymptotic solutions are obtained by the use of the similarity variable and the similarity

presentation of the solution. On the basis of the self-similar representation of the solution the near crack-tip stress, creep strain rate and continuity distributions are given. It is shown that meso-mechanical effect of damage accumulation near the crack tip results in new intermediate stress field asymptotic behavior and requires the solution of nonlinear eigenvalue problems. To attain eigensolutions a numerical scheme is worked out and the results obtained provide the additional eigenvalues of the HRR problem. By the use of the method proposed the whole set of eigenvalues for the mode crack in a power law material under mixed mode loading can be determined. The self-similar solutions are based on the idea of the existence of the completely damaged zone near the crack tip. The stress and creep strain rate angular functions are constructed. The higher order terms of the asymptotic expansions of stresses, creep strain rates and continuity parameter allowing to obtain the contours of the completely damaged zone in the vicinity of the crack tip are derived and investigated. The extent of the area in the vicinity of the crack tip where the material undergoes damage for the specimen under tensile loading is studied in [29–33]. The results obtained in the



present paper are in good agreement with the results of [29–33].

Fig. 7. Geometry of the completely damaged zone near the crack tip under mixed mode loading for  $M^p = 0.7$  and  $M^p = 0.9$ .

## Acknowledgements

Financial support from the Russian Foundation of Basic Research (project No. 16-08-00571) is gratefully acknowledged.

## References

- [1] Meng Q, Weng Z. Creep damage models and their applications for crack growth analysis in pipes: A review. *Engineering Fracture Mechanics* 2016; 1–30.
- [2] Altenbach H, Sadowski T. *Failure and Damage Analyses of Advanced Materials*. Berlin: Springer, 2015.
- [3] Barenblatt GI. *Deformation and Fracture: Lectures on Fluid Mechanics and the Mechanics of Deformable Solids for Mathematicians and Physicists*. Cambridge University Press, 2014.
- [4] Bui HD. *Fracture Mechanics. Inverse Problems and Solutions*. Dordrecht: Springer, 2006.
- [5] Stepanova LV, Roslyakov PS. Multi-parameter description of the crack-tip stress field: Analytic determination of coefficients of crack-tip stress expansions in the vicinity of the crack tips of two finite cracks in an infinite plane medium. *International Journal of Solids and Structures* 2016; 100–101: 10–28.
- [6] Murakami S. *Continuum Damage Mechanics. A Continuum Mechanics Approach to the Analysis of Damage and Fracture*. Dordrecht: Springer, 2012.
- [7] Kuna M. *Finite Elements in Fracture Mechanics. Theory-Numerics-Applications*. Dordrecht: Springer, 2013.
- [8] Soyarslan C, Richter H, Bargmann S. Variants of Lemaitre damage model and their use in formability prediction of metallic materials. *Mechanics of Materials* 2016; 92: 58–79.
- [9] Stepanova LV, Adylina EM. Stress-strain state in the vicinity of a crack under mixed loading. *Journal of Applied Mechanics and Technical Physics* 2014; 55(5): 885–895.
- [10] Ochsner A. *Continuum Damage and Fracture Mechanics*. Springer Science + Business Media, Singapore, 2016.
- [11] Richard HA, Schramm B, Schirmeisen N-H. Cracks on Mixed Mode loading – Theories, experiments, simulations. *International Journal of Fatigue* 2014; 62: 93–103.
- [12] Voyiadis GZ, Kattan PI. *Damage Mechanics with Finite Elements: Practical Applications with Computer Tools*. Berlin: Springer, 2012.
- [13] Voyiadis GZ. *Handbook of Damage Mechanics*. New York: Springer – Verlag, 2015.
- [14] Tumanov AV, Shlyannikov VN, Kishen CJM. An automatic algorithm for mixed mode crack growth rate based on drop potential method. *International Journal of Fatigue* 2015; 81: 227–237.
- [15] Wei RP. *Fracture Mechanics. Integration of Mechanics, Materials Science and Chemistry*, Cambridge University Press, 2014.
- [16] Zhang W, Cai Y. *Continuum Damage Mechanics and Numerical Applications*. Heidelberg: Springer Science & Business Media, 2010.
- [17] Chousal JAG, M.F.S.F. de Moura, Mixed mode I+II continuum damage model applied to fracture characterization of bonded joints. *Int. J. of Adhesion and Adhesives* 201 ; 41: 92–97.
- [18] Riedel H. *Fracture at High Temperature*. Berlin: Springer–Verlag, 1987.
- [19] Stepanova LV. Eigenvalue of the antiplane-shear crack problem for a power-law material. *Journal of Applied Mechanics and Technical Physics* 2008; 49(1): 142–147.
- [20] Stepanova LV. Eigenspectra and orders of stress singularity at a mode I crack tip for a power-law medium. *Comptes Rendus. Mecanique* 2008; 336(1-2): 232–237.
- [21] Beliakova TA, Kulagin VA. The eigenspectrum approach and T-stress at the mixed-mode crack tip for a stress-state-dependent material. *Procedia Materials Science* 2014; 3: 147–152.
- [22] Krepl O, Klusak J. Reconstruction of a 2D stress field around the tip of a sharp material inclusion. *Procedia Structural Integrity* 2016; 2: 1920–1927.
- [23] Stepanova LV. Eigenvalue analysis for a crack in a power-law material. *Computational Mathematics and Mathematical Physics* 2009; 49(8): 1332–1347.

- [24] Stepanova LV, Yakovleva EM. Mixed-mode loading of the cracked plate under plane stress conditions. *PNRPU Mechanics Bulletin* 2014; 3: 129–162.
- [25] Torabi AR, Abedinasab SM. Brittle fracture in key-hole notches under mixed mode loading. Experimental study and theoretical predictions. *Engineering Fracture Mechanics* 2015; 134: 35–53.
- [26] Stepanova LV, Igonin SA. Asymptotics of the near-crack-tip stress field of a growing fatigue crack in damaged materials: Numerical experiment and analytical solution. *Numerical Analysis and Applications* 2015; 8(2): 168–181.
- [27] Kachanov LM. On rupture time under condition of creep. *Izvestia Akademi Nauk SSSR Otd Tekhn Nauk* 1958; 8: 26–31.
- [28] Rabotnov YN. Creep problems in structure members. Amsterdam: North-Holland, 1969.
- [29] Vesely V, Frantik P. Reconstruction of a fracture process zone during tensile failure of quasi-brittle materials. *Applied and Computational Mechanics* 2010; 4: 237–250.
- [30] Galouei M, Fakhimi A. Size effect, material ductility and shape of fracture process zone in quasi-brittle materials. *Computers and Geotechnics* 2015; 65: 126–135.
- [31] Frantik P, Vesely V, Kersner Z. Parallelization of lattice modelling for estimation of fracture process zone extent in cementitious composites. *Advances in Engineering Software* 2013; 60-61: 48–57.
- [32] Fakhimi A, Wan F. Discrete element modeling of the process zone shape in mode I fracture at peak load and in post-peak regime. *International Journal of Rock Mechanics & Mining Sciences* 2016; 85: 119–128.
- [33] Wei MD, Dai F, Xu NW, Zhao T, Xia KW. Experimental and numerical study on the fracture process zone and fracture toughness determination for ISRM-suggested semi-circular bend rock specimen. *Engineering Fracture Mechanics* 2016; 154: 43–56.

# Calculation of critical conditions for the filtration combustion model

O. Vidilina<sup>1</sup>, E. Shchepakina<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The paper is devoted to the study of the dynamic model of the autocatalytic combustion reaction in an inert medium with partial heat removal from the reaction phase to the environment. We pay particular attention to modelling of the critical regime, which is a kind of a watershed between the slow burning regimes and explosion modes. New algorithm for computing a critical value of the control parameter is presented.

*Keywords:* filtration combustion; thermal explosion; critical phenomena; singular perturbations; integral manifolds; canards

## 1. Introduction

In last few years there was an increase in researches concerning multiphase combustions systems. The results of the studies are widely used in the problems of safety of gas emissions, explosive dust clouds, mixture detonations, transportation and use of combustible and explosive substances.

In the present paper we consider a mathematical model of autocatalytic combustion reaction in a multiphase medium. The multiphase nature of the process arises from the existence of inert phase alongside with the reactant phase. The inert medium could correspond, for example, to a dusty medium or a porous matrix. We paid particular attention to the modelling of the critical regime that is kind of a watershed between the slow burning regimes and thermal explosions.

The main goal of the mathematical theory of thermal explosion [1-4] is to study the dynamics of the combustion process for a given dimensions of the reactor, thermophysical and kinetic characteristics, heat transfer coefficient. These characteristics correspond to parameters of the differential system, which is a mathematical model of the process. Under certain conditions of values of these parameters, the reaction proceeds for as long as possible without transition into the explosion or the slow burning mode. We call such regime a critical one.

The goal of the present work is to determine the values of the parameters that correspond to the critical regime. In order to find the values we consistently applied analytical and numerical methods. The main result of the paper is the derivation of the algorithm used in calculation of a value of the control parameter of the system that corresponds to the critical regime. The critical value of the control parameter is a solution of an algebra-differential system.

## 2. Model

We consider combustion model of a rarefied gas mixture in an inert porous, or in a dusty, medium. We assume that the temperature distribution and phase-to-phase heat exchange are uniform. The chemical conversion kinetics are represented by a one-stage, irreversible reaction. The dimensionless model in this case has the form [5-7]:

$$\gamma \frac{d\theta}{d\tau} = \eta (1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) - \alpha(\theta - \theta_c) - \delta\theta, \quad (1)$$

$$\gamma_c \frac{d\theta_c}{d\tau} = \alpha(\theta - \theta_c), \quad (2)$$

$$\frac{d\eta}{d\tau} = \eta (1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right), \quad (3)$$

with initial condotions

$$\eta(0) = \eta_0 / (1 + \eta_0) = \bar{\eta}_0, \quad \theta(0) = \theta_c(0) = 0. \quad (4)$$

Here  $\theta$  and  $\theta_c$  are the dimensionless temperatures of the reactant phase and of the inert phase, respectively;  $\eta$  is the depth of conversion;  $\tau$  denotes the dimensionless time;  $\eta_0$  is the parameter for autocatalyticity (this kinetic parameter characterizes the degree of self-acceleration of the reaction: the lower the value, the more marked the autocatalytic reaction will be). The terms  $-\delta\theta$  and  $-\alpha(\theta - \theta_c)$  reflect the external heat dissipation and phase-to-phase heat exchange. The parameters  $\gamma$  and  $\gamma_c$  characterize the physical features of the reactant phase and of the inert phase, respectively. System (1)-(3) is singularly perturbed since  $\beta$  and  $\gamma$  are the small for typical combustible gas mixture [1-3].

Depending on the relation between values of the parameters, the chemical reaction either moves to a slow regime with decay of the reaction, or into a regime of self-acceleration which leads to an explosion. So, if we change the value of one parameter, with fixed values of the other parameters, we can change the type of chemical reaction. Let us consider  $\alpha$  as a control parameter. For some value of  $\alpha$  (we call it critical) the reaction is maintained and gives rise to a rather sharp transition from slow motions to explosive ones. The transition region from slow regimes to explosive ones exists due to the continuous dependence of the system



(1)-(3) on the parameter  $\alpha$ . To find the critical value of the parameter  $\alpha$ , it is possible to use special asymptotic formulae [4, 7, 8]. That approach was used in [5-7, 9] for system (1)-(3), in [10-19] for other laser and chemical systems, and in [20-24] for some biological problems. In the next section the main results concerning this approach obtained for system (1)-(3) are given. The realizability conditions for the critical regime were obtained in the form of a system of non-linear algebra-differential equations, but the problem of calculating the critical value of the control parameter with the help of this system had not been solved. The paper is devoted to develop an algorithm for calculating the critical parameter value. The readers can find details of this algorithm in Sections 4 and 5.

### 3. Modelling of the critical regime

We will consider the case  $\eta_0 = 0$  for the sake of simplicity, taking into account that for case  $\eta_0 \neq 0$  the correction to the initial conditions can be found with help of fast integral manifolds [4].

The slow surface  $S$  of system (1)-(3) is described by the equation (see Figure 1):

$$\eta(1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) - \alpha(\theta - \theta_c) - \delta\theta = 0.$$

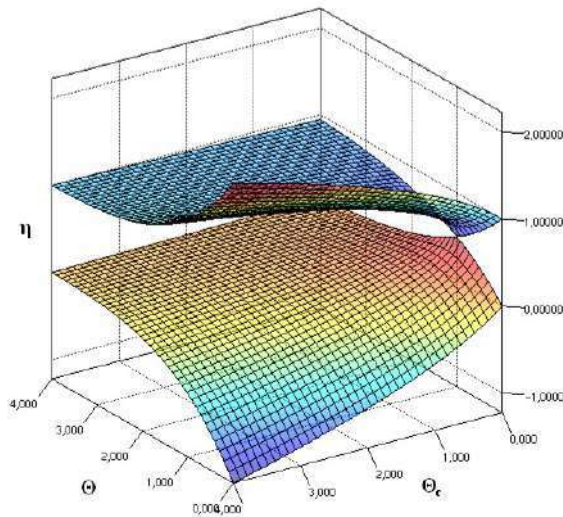


Fig. 1. The slow surface of system (1)-(3).

This surface is a zero-order ( $\gamma = 0$ ) approximation of a slow integral manifold of the system [4, 7, 8]. Recall, that the slow integral manifold of a singularly perturbed system is defined as an invariant surface of slow motions, i.e., the flow on it has the order  $O(1)$  as  $\gamma \rightarrow 0$ . Far from the slow surface, the fast variables of the system vary very rapidly, with a speed of order  $O(1/\gamma)$  as  $\gamma \rightarrow 0$ .

The intersection of the slow surface with the surface of irregular points (see Figure 2), given by the expression

$$\eta(1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) \frac{1}{(1 + \beta\theta)^2} - \alpha - \delta = 0$$

determines a breakdown curve. The breakdown curve separates the stable ( $S^s$ ) and unstable ( $S^u$ ) subsets of the slow surface  $S$ , see Figure 3. System (1)-(3) has a stable integral manifold ( $S_\varepsilon^s$ ) and an unstable integral manifold ( $S_\varepsilon^u$ ) near  $S^s$  and  $S^u$ , respectively.

When  $\alpha > \alpha^*$  the trajectories of the system starting at the initial point move along the stable branch  $S^s$  and the temperature  $\theta$  does not reach relatively large values (see Figure 4). These trajectories correspond to the slow burning regimes.

When  $\alpha < \alpha^*$  the system's trajectories, having reached the breakdown curve along  $S^s$  at the tempo of the slow variable, jump into the explosive regime (see Figure 5).

Due to the continuous dependence of the right-hand side of (1)-(3) on the parameter  $\alpha$  there are some intermediate trajectories in the region between those shown above. For some value  $\alpha = \alpha^*$  we can glue the stable and unstable slow integral manifolds at a point of the breakdown curve to get a canard [7, 8, 25, 26], i.e., the system's trajectory which at first move along the stable slow integral manifold and then continue for a while along the unstable slow integral manifold, see Figure 6.

The canard describes the critical regime that separates the domain of slow burning modes and the domain of thermal explosion. A deviation from the value of  $\alpha^*$  leads to the destruction of the gluing of stable and unstable slow integral manifolds with a subsequent reaction's transition either to the slow regime (when the trajectory of the system unfolds along a stable manifold from the breakdown curve) or into the thermal explosion mode (when the trajectory, reaching the breakdown curve, jumps from the slow manifold and rapidly runs away from it).

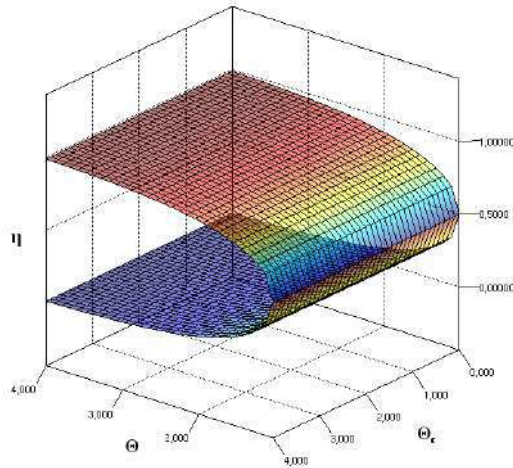


Fig. 2. The surface of irregular points.

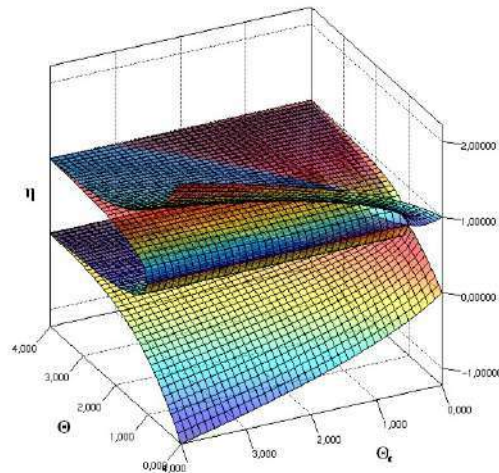


Fig. 3. The intersection of the slow surface of system (1)-(3) with the surface of irregular points. On the intersection (the breakdown line) the stability of the slow manifold changes. The upper and lower sheets of the slow surface are stable, the part enclosed between the surface of irregular points is unstable.

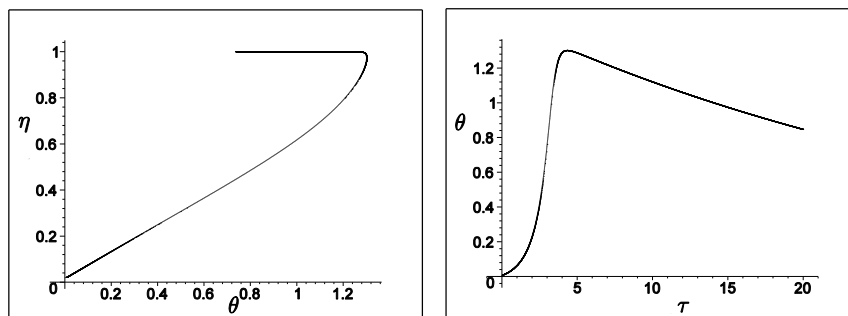


Fig. 4. The trajectory (left) and the  $\theta$ -component (right) in the case of a slow birning regime:  
 $\alpha = 3, \beta = 0.1, \gamma = 0.001, \gamma_c = 0.7, \bar{\eta}_0 = 0.02, \delta = 0.02.$

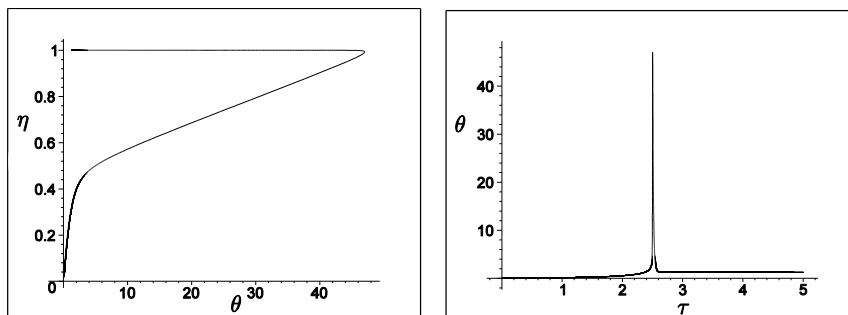


Fig. 5. trajectory (left) and the  $\theta$ -component (right) in the case of thermal explosion:  
 $\alpha = 0.7, \beta = 0.1, \gamma = 0.001, \gamma_c = 0.7, \bar{\eta}_0 = 0.02, \delta = 0.02.$

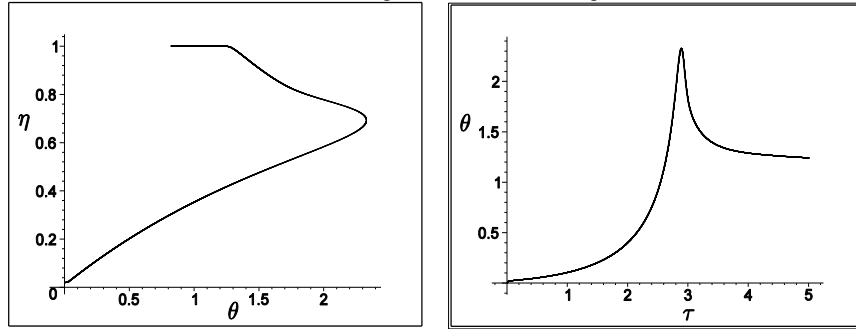


Fig. 6. The canard and the  $\theta$ -component (right) in the case of critical regime:  $\alpha = 0.949$ ,  $\beta = 0.1$ ,  $\gamma = 0.001$ ,  $\gamma_c = 0.7$ ,  $\bar{\eta}_0 = 0.02$ ,  $\delta = 0.02$ .

To calculate the critical value of the parameter  $\alpha = \alpha^*$  and the asymptotic expressions for the corresponding canard

$$\alpha^* = \alpha_0 + \gamma\alpha_1 + o(\gamma),$$

$$\theta(\eta, \gamma) = \varphi_0(\eta) + \gamma\varphi_1(\eta) + o(\gamma), \quad (5)$$

$$\theta_c(\eta, \gamma) = \psi_0(\eta) + \gamma\psi_1(\eta) + o(\gamma),$$

we use the usual method of eliminating an independent variable. In this case, the system (1)-(3) takes the form

$$\gamma \frac{d\theta}{d\eta} \eta (1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) = \eta (1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) - \alpha(\theta - \theta_c) - \delta\theta,$$

$$\gamma_c \frac{d\theta_c}{d\eta} \eta (1 - \eta) \exp\left(\frac{\theta}{1 + \beta\theta}\right) = \alpha(\theta - \theta_c).$$

We substitute (5) into these equations to get

$$\begin{aligned} \gamma(\varphi_0' + \gamma\varphi_1') \eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) \left[1 + \gamma \frac{\varphi_1}{(1 + \beta\varphi_0)^2}\right] &= \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) \left[1 + \gamma \frac{\varphi_1}{(1 + \beta\varphi_0)^2}\right] \\ &\quad - (\alpha_0 + \gamma\alpha_1)(\varphi_0 - \psi_0 + \gamma(\varphi_1 - \psi_1)) - \delta(\varphi_0 + \gamma\psi_1) + o(\gamma), \end{aligned} \quad (6)$$

$$\begin{aligned} \gamma_c(\varphi_0' + \gamma\varphi_1') \eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) \left[1 + \gamma \frac{\varphi_1}{(1 + \beta\varphi_0)^2}\right] \\ = (\alpha_0 + \gamma\alpha_1)(\varphi_0 - \psi_0 + \gamma(\varphi_1 - \psi_1)) + o(\gamma). \end{aligned} \quad (7)$$

Setting  $\gamma = 0$  in (6) and (7), we obtain

$$\eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) - \alpha_0(\varphi_0 - \psi_0) - \delta\varphi_0 = 0, \quad (8)$$

$$\gamma_c \varphi_0' \eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) = \alpha_0(\varphi_0 - \psi_0). \quad (9)$$

From the equation of the breakdown curve, taking into account (5), we have

$$\eta^* (1 - \eta^*) \exp\left(\frac{\varphi_0^*}{1 + \beta\varphi_0^*}\right) \frac{1}{(1 + \beta\varphi_0^*)^2} - (\alpha_0 + \delta) = 0, \quad (10)$$

where  $(\eta^*, \varphi_0^*, \psi_0^*)$  is the gluing point of the slow integral manifolds.

After double differentiation (8) with respect to  $\eta$ , with taking into account (10), we get one more condition at the gluing point:

$$(1 - 2\eta^*) \exp\left(\frac{\varphi_0^*}{1 + \beta\varphi_0^*}\right) + \alpha_0\psi_0'^* = 0, \quad \psi_0'^* = \psi_0'(\eta^*). \quad (11)$$

Thus, the expressions (8)-(11) give us the zeroth order approximations of the critical value of the control parameter and the canard.

Further, in order to find the first order approximations, we equate the coefficients of  $\gamma$  in the first degree in the system (6), (7). As a result, we get:

$$\varphi_0 \eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) = \left[\eta (1 - \eta) \exp\left(\frac{\varphi_0}{1 + \beta\varphi_0}\right) \frac{\varphi_1}{(1 + \beta\varphi_0)^2} - (\alpha_0 + \delta)\right] \varphi_1 + \alpha_0\psi_1 - \alpha_1(\varphi_0 - \psi_0), \quad (12)$$

$$\eta(1-\eta)\exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right)\left[\varphi_0' + \gamma_c\psi_1' + \frac{\varphi_1(\gamma_c\psi_0'-1)}{(1+\beta\varphi_0)^2}\right] = -\delta\varphi_1. \quad (13)$$

$$\alpha_1 = \frac{1}{\varphi_0^* - \psi_0^*} \left[ \alpha_0 \psi_1'^* - \varphi_0'^* \eta^* (1 - \eta^*) \exp\left(\frac{\varphi_0^*}{1+\beta\varphi_0^*}\right) \right]. \quad (14)$$

Here  $\varphi_0'^* = \varphi_0'(\eta^*)$ .

The expressions (12)-(14) determine the first order approximations of the critical value of the control parameter and the canard. It should be noted that to calculate the values of  $\alpha_0$  and  $\alpha_1$  from (8)-(14) it is necessary to apply numerical methods. The development of the algorithm for finding the critical value of the control parameter is our next goal.

#### 4. The gluing point

In order to verify the correctness of the algorithm, developed in the present paper, we can use some specific case when an analytical solution of (8)-(14) is available and compare it to the one yielded by our method. For this goal we now consider the case  $\delta = 0$  which corresponds the absence of external heat dissipation.

In the case  $\delta = 0$  system (1)-(3) possesses a first integral

$$\eta - \gamma\theta - \gamma_c \theta_c = 0.$$

With the help of this first integral, we can reduce the order of system (1)-(3) by eliminating the variable  $\theta_c$ . As a result we obtain a plane system:

$$\begin{aligned} \gamma \frac{d\theta}{d\tau} &= \eta(1-\eta)\exp\left(\frac{\theta}{1+\beta\theta}\right) - \alpha(1 + \gamma/\gamma_c)\theta + \alpha/\gamma_c(\eta - \bar{\eta}_0), \\ \frac{d\eta}{d\tau} &= \eta(1-\eta)\exp\left(\frac{\theta}{1+\beta\theta}\right). \end{aligned}$$

Here we have to deal with the slow curve rather than then slow surface [5-7, 9]. The coordinates of the gluing point of the integral manifolds for some value  $\alpha = \alpha_0$  can be found from the self-intersection conditions of the slow curve, which in our case have the form

$$\eta^*(1-\eta^*)\exp\left(\frac{\theta^*}{1+\beta\theta^*}\right) - \alpha\theta^* + \frac{\alpha}{\gamma_c}\eta^* = 0, \quad (15)$$

$$(1-2\eta^*)\exp\left(\frac{\theta^*}{1+\beta\theta^*}\right) + \frac{\alpha}{\gamma_c} = 0, \quad (16)$$

$$\eta^*(1-\eta^*)\exp\left(\frac{\theta^*}{1+\beta\theta^*}\right)\frac{1}{(1+\beta\theta^*)^2} - \alpha = 0. \quad (17)$$

From (15) and (17) we get

$$\eta^* = \gamma_c\theta^* - (1 + \beta\theta^*)^2\gamma_c. \quad (18)$$

Using (15) and (16), we obtain

$$(1-2\eta^*)(\theta^* - \gamma_c^{-1}\eta^*) + \gamma_c^{-1}\eta^*(1-\eta^*) = 0.$$

From here it follows that

$$\gamma_c(1 + \beta\theta^*)^4 = \gamma_c\theta^{*2} - \theta^*. \quad (19)$$

Equations (18) and (19) give us the values  $\theta^*$  and  $\eta^*$ . Further, from equation (16) we obtain

$$\alpha_0 = \gamma_c(2\eta^* - 1)\exp\left(\frac{\theta^*}{1+\beta\theta^*}\right).$$

Using the smallness of the parameter  $\beta$ , from (19) we can analytically find the value  $\theta^*$  in the form of an asymptotic expansion with respect to the parameter  $\beta$ :  $\theta^* = \theta_{00}^* + \beta\theta_{01}^* + o(\beta)$ . Similar expansions can also be written for  $\alpha_0$  and  $\eta^*$ . In the case  $\beta = 0$  we have:

$$\theta_{00}^* = \frac{1}{2}(\gamma_c^{-1} + \sqrt{4 + \gamma_c^{-2}}),$$

$$\eta_{00}^* = \gamma_c(\theta_{00}^* - 1),$$

$$\alpha_{00} = \frac{\exp(\theta_{00}^*)}{2 + \sqrt{4 + \gamma_c^{-2}}}.$$

**5. Algorithm for calculating the critical value of the control parameter**

Let us describe an algorithm for solving the system of nonlinear algebra-differential equations (8)–(11) with initial conditions that we obtained from (4)–(5):

$$\eta(0) = \frac{\eta_0}{1+\eta_0} = \bar{\eta}_0, \varphi_0(0) = \varphi_0.$$

We consider numerical solution of the system. Using (8) we get:

$$\psi_0 = \varphi_0 \left( 1 + \frac{\delta}{\alpha_0} \right) - \frac{\eta(1-\eta)}{\alpha_0} \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right).$$

Taking the derivative yields:

$$\psi'_0 = \frac{1}{\alpha_0} \left[ \varphi'_0(\alpha_0 + \delta) + (1 - 2\eta) \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right) - \eta(1 - \eta) \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right) \frac{\varphi'_0}{(1+\beta\varphi_0)^2} \right].$$

Now, let us substitute  $\psi_0$  and  $\psi'_0$  into (9):

$$\varphi'_0 = \left( \frac{\alpha_0}{\gamma_c} - \frac{\alpha_0 \delta \varphi_0}{\gamma_c \eta(1-\eta) \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right)} + \frac{(1-2\eta)}{\eta(1-\eta)} \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right) \right) / \left( \alpha_0 + \delta - \frac{\eta(1-\eta) \exp\left(\frac{\varphi_0}{1+\beta\varphi_0}\right)}{(1+\beta\varphi_0)^2} \right).$$

We get a first order ordinary differential equation with respect to  $\varphi_0 = \varphi_0(\eta)$ , where  $\eta$  is a variable and  $\alpha_0$  is a constant.

For any given value of  $\alpha_0$  we can solve the equation using Runge-Kutta fourth-order method. This method has a good precision and, in spite of its laboriousness, is widely used to obtain a numerical solution for ordinary differential equations. Moreover, we can further benefit from this method by using an adaptive stepsize.

Next, in order to the coordinates of the gluing point we find  $\eta^*$  and  $\varphi_0^*$  from (10) and (11):

$$\eta^*(1 - \eta^*) \exp\left(\frac{\varphi_0^*}{1+\beta\varphi_0^*}\right) \frac{1}{(1+\beta\varphi_0^*)^2} - (\alpha_0 + \delta) = 0,$$

$$(1 - 2\eta^*) \exp\left(\frac{\varphi_0^*}{1+\beta\varphi_0^*}\right) + \alpha_0 \frac{\eta^*(1-\eta^*) \exp\left(\frac{\varphi_0^*}{1+\beta\varphi_0^*}\right) - \delta \varphi_0^*}{\gamma_c \eta^*(1-\eta^*) \exp\left(\frac{\varphi_0^*}{1+\beta\varphi_0^*}\right)} = 0.$$

We can also solve this nonlinear system numerically using any of existing iterative methods. Let us note that sometimes it is possible to solve such systems by applying the elimination and back substitution method. However, in vast majority of cases iterative methods are used. Next, we substitute the solution  $\eta^*$  into (8) to obtain  $\varphi_0^{**}$ . Minimizing the difference between  $\varphi_0^*$  and  $\varphi_0^{**}$ , we can find  $\alpha_0$  with arbitrary precision.

Analogously, we get  $\alpha_1$  from (12)-(14) as we already computed  $\alpha_0$ .

The algorithm was implemented using Java. To verify the correctness of the algorithm we considered a special case ( $\beta = 0, \gamma = 0, \delta = 0$ ), that allows us to solve the system analytically. Then we compared the analytical solution with the numerical solution yielded by the algorithm. The results for  $\alpha_0$  are presented in Table 1.

Table 1. The comparison between analytical and numerical solutions for  $\alpha_0$  for various values of  $\gamma_c$ .

$\gamma_c$	$\alpha_0$	
	Analytical solution	Numerical solution
0.6	1.837200	1.837291
0.7	1.566012	1.566038
0.8	1.393921	1.393923
0.9	1.275978	1.275919
1.1	1.125980	1.125923
1.2	1.075605	1.075623
1.3	1.035258	1.035253

**6. Conclusion**

We have studied the mathematical model of filtration combustion of combustible gas in an autocatalytic reaction case. We have shown that the critical phenomena of the model can be described by the canard. The critical regime is a kind of watershed between the safe processes and thermal explosion regimes. We have established that it is possible to control the mode and,

therefore, the combustion process by adjusting the parameter characterizing the heat removal from the reaction phase to the external environment.

We have developed a new algorithm that allows to compute the critical value of the control parameter by combining analytical methods of the geometric theory of singular perturbations and numerical methods. The presented algorithm can be used in other similar problems for studying critical phenomena in dynamic systems and calculating critical values of control parameters.

## Acknowledgements

The contribution of O. Vidilina was supported by the Russian Foundation for Basic Research and Samara region (grant 16-41-630529-p) and the Ministry of Education and Science of the Russian Federation as part of a program of increasing the competitiveness of SSAU in the period 2013–2020. E. Shchepakina was supported by the Ministry of Education and Science of the Russian Federation (Project RFMEFI58716X0033).

## References

- [1] Spalding von DB. *Combustion and Mass Transfer*. Oxford – New York: Pergamon Press, 1979; 409 p.
- [2] Zeldovich YaB, Barenblatt GI, Librovich VB, Makhviladze GM. *The Mathematical Theory of Combustion and Explosions*. New York: Consultants Bureau, 1985; 597 p.
- [3] Babushok VI, Goldshtein VM, Sobolev VA. Critical Condition for the Thermal Explosion with Reactant Consumption. *Combust. Sci. and Tech.* 1990; 70: 81–89.
- [4] Sobolev VA, Shchepakina EA. *Model Reduction and Critical Phenomena in Macrokinetics*. Moscow: “Energoatomizdat” Publisher, 2010; 320 p. (in Russian)
- [5] Gol'dshtein V, Zinoviev A, Sobolev V, Shchepakina E. Criterion for thermal explosion with reactant consumption in a dusty gas. *Proc. of London R. Soc. Ser. A* 1996; 452: 2103–2119.
- [6] Gorelov GN, Shchepakina EA, Sobolev VA. Canards and critical behavior in autocatalytic combustion models. *Journal of Engineering Mathematics* 2006; 56(2): 143–160.
- [7] Shchepakina E, Sobolev V. Black swans and canards in laser and combustion models. *Singular Perturbation and Hysteresis*. Philadelphia: SIAM, 2005; 207–255.
- [8] Shchepakina E, Sobolev V, Mortell MP. Canards and Black Swans. *Singular Perturbations. Introduction to System Order Reduction Methods with Applications*. *Lecture Notes in Mathematics* 2014; 2114: 141–182.
- [9] Golodova ES, Shchepakina EA. Modeling of safe combustion at the maximum temperature. *Mathematical Models and Computer Simulations* 2009; 1(2): 322–334.
- [10] Gorelov GN, Sobolev VA. Mathematical modeling of critical phenomena in thermal explosion theory. *Combust. Flame* 1991; 87: 203–210.
- [11] Gorelov GN, Sobolev VA. Duck-trajectories in a thermal explosion problem. *Appl. Math. Lett.* 1992; 5(6): 3–6.
- [12] Sobolev VA, Shchepakina EA. Self-ignition of dusty media. *Combustion Explosion Shock Waves* 1993; 29: 378–381.
- [13] Sobolev VA, Shchepakina EA. Duck trajectories in a problem of combustion theory. *Differential Equations* 1996; 32: 1177–1186.
- [14] Shchepakina E. Black swans and canards in self-ignition problem. *Nonlinear Analysis: Real World Applications* 2003; 4: 45–50.
- [15] Shchepakina E. Canards and black swans in model of a 3-D autocatalator. *J. Phys.: Conf. Series* 2005; 22: 194–207.
- [16] Shchepakina E, Korotkova O. Condition for canard explosion in a semiconductor optical amplifier. *Journal of the Optical Society of America B: Optical Physics* 2011; 28(8): 1988–1993.
- [17] Shchepakina E, Korotkova O. Canard explosion in chemical and optical systems. *Discrete and Continuous Dynamical Systems – Series B* 2013; 18(2): 495–512.
- [18] Shchepakina EA. Critical phenomena in a model of fuel's heating in a porous medium. *CEUR Workshop Proceedings* 2015; 1490: 179–189.
- [19] Firstova N, Shchepakina E. Conditions for the critical phenomena in a dynamic model of an electrocatalytic reaction. *J. Phys.: Conf. Series* 2017; 811: 012002.
- [20] Sobolev V. Canard cascades. *Discrete and Continuous Dynamical Systems – Series B* 2013; 18(2): 513–521.
- [21] Gavin C, Pokrovskii A, Prentice M, Sobolev V. Dynamics of a Lotka-Volterra type model with applications to marine phage population dynamics. *J. Phys.: Conf. Series* 2006; 55(1): 80–93.
- [22] Pokrovskii A, Shchepakina E, Sobolev V. Canard doublet in a Lotka-Volterra type model. *J. Phys.: Conf. Series* 2008; 138: 012019.
- [23] Pokrovskii A, Rachinskii D, Sobolev V, Zhezherun A. Topological degree in analysis of canard-type trajectories in 3-D systems. *Applicable Analysis* 2011; 90(7): 1123–1139.
- [24] Sobolev VA. Canards and the effect of apparent disappearance. *CEUR Workshop Proceedings* 2015; 1490: 190–197.
- [25] Benoit E, Callot JL, Diener F, Diener M. Chasse au canard. *Collect. Math.* 1981; 31-32: 37–119. (in French)
- [26] Shchepakina E, Sobolev V. Integral manifolds, canards and black swans. *Nonlinear Analysis. A* 2001; 44: 897–908.

# Mathematical modeling radio tomographic ionospheric parameters reconstruction via nanosatellites constellation for conditions of incomplete source data

O.V. Phylonin<sup>1</sup>, P.N. Nikolaev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The results of mathematical simulation of the formation of initial projection data for the problems of radio-tomography of the ionosphere using navigation satellites GPS and constellation of low earth orbit nanosatellites are presented. It is shown that for ring carriers (spherical layers), in the incompleteness of chordal data, the problems of reconstructing the electron density can be reduced to problems of low-angle tomography and the use of high-speed FBP algorithms.

*Keywords:* radio-tomography of the ionosphere; total electron content; nanosatellites; navigation satellites; FBP

---

## 1. Introduction

One of the promising areas of application of small satellites, including nanosatellites (NS), is their use for solving problems of radio-tomography, lidar-tomography of the ionosphere, analysis of the composition of cosmic radiation, etc. The radio-tomography of the ionosphere (IRT) makes it possible to study various ionospheric structures, namely:

- ionization dips, in particular the total electron content (TEC);
- wave and quasiwave structure;
- traveling ionospheric disturbances (TIDs): "blobs", "patches", "bubbles", "ionization tongue";
- ionospheric "traces" of the corpuscular ionization, etc.

Studies of the structure of the ionosphere are important both for a theoretical understanding of the physics of processes occurring in it, and for many practical problems, since the ionosphere as a medium for the propagation of radio waves significantly affects the operation of various navigation, location and communication systems. The existing radar facilities and ionosondes allow only local diagnostics of the ionosphere. The creation of a fairly dense network of traditional ionosphere probes [1] is very difficult and expensive.

At the same time, existing satellite constellation such as:

- low earth orbit (LEO) constellation - Russian "Cicada" American "Transit";
- high earth orbit (HEO) constellation GPS/GLONASS;
- network of ground-based receivers,

make it possible to sound the ionosphere in different directions and to apply tomographic methods for reconstructing the ionosphere parameters.

In other words, IRT techniques allow to reconstruct the spatial structure of the electron density of the ionospheric plasma [1, 3]. From the beginning of the 1990s, radio-tomography systems operate on the basis of LEO navigation systems. In recent years, radiotomographic studies have been actively carried out using data from HEO navigation systems [4, 5]. To identify different types of ionospheric radio tomography are used here, the terms LEO radio tomography, HEO radio tomography: LEORT and HEORT, respectively.

The radio-tomography of the ionosphere is based on the use, for example, of a two-frequency method, the essence of which can be clarified as follows. When the satellite is moving, ground receiving stations, or other satellites in the same orbit, continuous measurements of the phase delay of signals passing through the ionosphere at two frequencies are conducted. The initial data (chord data in the sense of Radon) are the corresponding phase paths of the radio signals measured in the lengths of the probing waves. If the frequency of the signals is much higher than some, the so-called plasma frequency, then from these data it is possible to determine the integral of the electron concentration along the trajectory of the satellite-receiver beam (the so-called total electronic content - TEC):

$$\int_l N_e(r) dl = \left( \frac{L_1}{f_1} - \frac{L_2}{f_2} \right) \frac{f_1^2 f_2^2}{f_1^2 - f_2^2} \frac{c}{K} + const$$

here:  $c$  - speed of light,  $K = 40.308 \text{ m}^3 / \text{s}^2$ .

Thus, a typical problem of tomographic type arises - the definition of a function of several arguments with respect to a set of linear integrals from it (along the paths of the satellite-receiver). Essential features of this problem are:

- Firstly, the presence of an unknown phase constant for each beam of rays (since only phase change is possible for observations during the span of the satellite in the visibility zone).
- Secondly, the uncertainty of the problem due to the fact that only a small number of beams of satellite-receiver beams intersect the local area of space (and in the case of HEORT, there may also exist areas of total absence of data associated with the unevenness of the network of receiving stations).

- Thirdly, when forming the initial chord data with the help of relatively HEO constellations of satellites, for example, GPS, forming highly stable radio pulses and low-orbiting groups of nanosatellites, recording the radio sounding signals, we deal with *small amounts of chord data obtained in limited angles of convergence on a ring carrier*.

The first problem is solved by using the phase difference approach (as an input difference of the integrals are taken over the neighboring rays, not the integrals themselves). To solve the second problem can be used iterative algorithms to ensure convergence to the normal solution (for a given norm), and also use special grid [6].

Regarding the third aspect it should be noted that the use of navigation satellite constellations such as GPS / GLONASS in combination with LEO nanosatellites constellation requires the development of innovative methods for the formation of the original projection data and algorithms for reconstruction of the desired spatial function distribution - such as electron density. This is due to the fact that the volumes and the possibility of such devices allow only transmit a digital code of the selected chord data in the Mission Control Center (MCC).

## **2. Analysis of the possibilities of using satellite constellations GLONASS, GPS in the problems of radio-tomography of the ionosphere in combination with LEO clusters of nanosatellites**

Methods and means of HEORT, and LEORT allow to recover not only the natural ionospheric irregularities, but also to detect ionospheric disturbances generated by anthropogenic sources. In particular, perturbations caused by the rocket launches, industrial explosions, powerful HF radiation [5, 6]. Methods of RTI using GPS / GLONASS constellation and LEO (250 - 450) km NS clusters also enable to determine the plasma flows, considering the sequence of radio tomography section of the ionosphere.

Input data for ionospheric monitoring problems are measuring the radio signal phase (phase path) when passing them to the path from the satellite to the ground station receiver at two operating frequencies. For GPS systems, these frequencies are  $f_1 = 1575.42$  MHz,  $f_2 = 1227.60$  MHz. Radio signals are continuously emitted by satellite navigation systems; provide many opportunities for the implementation of the ionospheric plasma research using radio-tomography methods. At the same time, the use of LEO and HEO systems leads to two essentially different objectives. The classical methods LEORT let you receive the "instant" two-dimensional cross-section of the ionosphere with high resolution (20 - 30) km. Modern HEORT systems provide four-dimensional (spatial / temporal), the electron density distribution at a lower resolution (up to 30 - 50) km directly dependent on the density of the network of receiving stations in the region.

The next stage of the IRT is a process for sounding of the ionosphere with the help of signals emitted by navigation systems, and the registration of radio signals that have passed a certain layer of the ionosphere, by using LEO constellation of micro and nanosatellites (orbit altitude 220-270 km). The satellites in the GLONASS system are moving in three orbital planes are separated relative to each other along the longitude of the ascending node  $120^\circ$ . The inclination of the orbital plane is  $64.8^\circ \pm 0.3^\circ$ . The orbits close to circular. The average height of orbits is 19,100 km. Each orbital plane is uniformly 8 satellites. Satellite orbital period is 11 hours and 15 minutes  $44 \pm 5$  seconds.

Satellites in the GPS system are moving in six orbital planes are separated relative to each other along the longitude of the ascending node at  $60^\circ$ . The inclination of the orbital plane is  $55^\circ$ . The orbits close to circular. The average height of the orbit 20189 km. Four satellites located in each orbital plane. The orbital period of the satellite 11:00 57 min 59.2 s (half of a sidereal day).

Each satellite has an atomic clock periodically synchronized by commands from Earth. Each satellite clocks synchronized satellite transmission via a special code signal. Before transmitting coded signals are modulated reports of movement trajectories of satellites and satellite parameters of time scales displacement models relative to the system scale. These messages are called navigation messages.

The structure of the signals transmitted by different satellites, such that:

- the receiver has the ability to separate these signals;
- assess their parameters;
- allocate navigation messages independently.

## **3. Features of the formation of a LEO CubeSat cluster in relation to the satellite constellation (GLONASS, GPS) to ionospheric radio tomography study**

To solve the problems of IRT by using navigation satellite constellations (GLONASS, GPS) and LEO clustered systems of CubeSat format, it is obviously necessary to arrange the NS in this orbit so that the conditions for obtaining the initial projection data in the sense of Radon's inverse are satisfied. It is clear that in this sense, the locations of the navigation satellites are rigidly fixed, therefore, the formation of the initial projection data is possible only through the configuration of the orbital constellation of the NS. It should be noted that the IRT using small satellites constellation can be implemented in several ways:

1. By placing constellation of small satellites equipped with transmit-receive systems on the low and medium orbits [7]. Sounding is performed along the chord direction in a ring layer (2D reconstruction problem) using the transmitters and receivers installed on each small satellite. Reconstruction accuracy at radio-tomography such organization depends on the frequency stability of the transmitters emitted radio wave pulses, the accuracy in determining the location of the small satellite in orbit, of the orbit itself forms etc. Furthermore, the pulses emitted by each small satellite transmitter should be identical, since it also determines the accuracy of the reconstruction of the desired functional distributions. It is clear that it is possible to meet these conditions, which allow obtaining an acceptable accuracy of IRT reconstruction [8], only in devices of relatively large mass,



since on their board it is necessary to mark highly stable transmitters, high-precision clocks, sensitive radio signal receivers, microprocessor control modules and preliminary data processing.

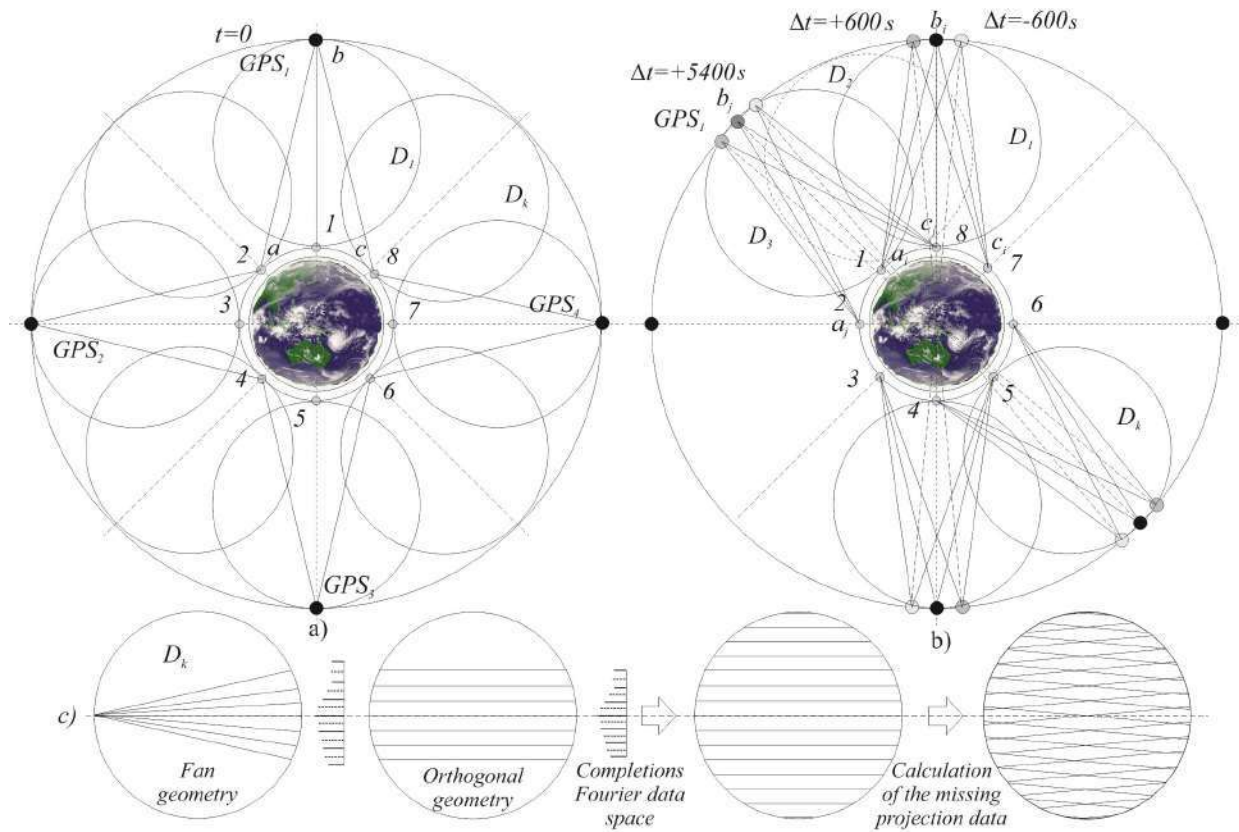


Fig. 1. Example of orbital configuration of GPS and LEO NS cluster for the IRT.

2. At that time, using radio signals of high quality of constellation of navigation satellites and a group of nanosatellites, a completely satisfactory solution of the problems of IRT is possible. Two frequency radio signal receivers, microprocessor control modules and preliminary processing of initial data, as well as transmitters of one-dimensional data sets to the control center, must be installed on board each NS.

Fig. 1 shows an example of the organization of an orbital constellation consisting of four GPS satellites (the average height of the orbit 20189 km) and eight nanosatellites placed in low orbit (220 - 270) km. The orbital plane, in this case the same - it is assumed the two-dimensional solution to the problem of IRT. Recall that the orbital period GPS - the satellite is 11 hours 57 minutes 59.2 seconds (half a sidereal day), and NS treatment period (80÷90) minutes for definiteness choose 90 minutes. It is obvious that for half a sidereal day, each of the NS, who is in a low orbit, make eight orbit pass. At present, many researchers to solve the ionospheric radio tomography problem apply algebraic method of reconstruction of the desired functionality distributions, e.g., the electron density distribution [2]. Indeed, this approach makes it possible to get the maximum resolution and accuracy ionospheric radio tomography problems, but at the same time requires huge computational costs. Significantly lower amounts of computational procedures allow tomography methods based on algorithms of convolution (FBP) and Fourier transforms (FT) [9]. The accuracy and resolution of the reconstruction are quite satisfactory.

When navigation satellites and NS are moving in their orbits, projection data corresponding to a certain angle of convergence - the angle between projections is formed. In particular, after  $\Delta t = 600 c$  time interval the satellite will take the position shown in Fig. 1 a), therefore, the convergence angle of will be  $\Psi = 2 \cdot (\pm 5) = 10$ . Due to the orbital motion of the satellite in a position  $\Psi = +5$  to register the chord data will be NS by numbered 8, 7, 6. Hence the issue for the microprocessor module processing raw data - the redistribution of integral chordates values of the on the corresponding recovery zone  $D_i$ . The signals from satellites are recorded respectively Ns triple numbered 2, 3, 4, ..., 6, 7, 8, taking into account the orbital NS bias. Fig. 1 b) shows the geometry of chord formation data over a time interval  $\Delta t = +5400 c$ .

Note that, as during orbital period navigation satellite, nanosatellites make eight orbit passes, it needs careful conversion of chord data for each of the reconstructed area  $D_i$ . The number of these circular areas should be more than twice as shown in Fig. 1 a).

Fig. 1 c) represented by the ideology of the pre-calculates base projection data to reconstruct the desired functionality distributions, for example, the TEC with the methods of few-view computer tomography. Its essence boils down to the following provisions:

- Implemented conversion of projection data from the fan-beam geometry to orthogonal geometry, but the reconstruction of the circular area is not completely filled with the chords of the projection;

- On the basis of a priori data, using interpolation methods in Fourier space completions produced projection data (on circular harmonics) [9] - to completely fill the circular recovery zone;
- With the help of transmitters installed on each NS, are transmitted one-dimensional projection data in the MCC, and the corresponding received complete a definition data for each projection. The number of the actual results of the projections is not enough to reconstruct the size of the matrix  $n \times n$ .
- Taking into account the a priori data, using the properties of symmetry of the Fourier images, using multiprocessor computing complexes MCC made completions missing projections, so that the condition [9]: the number of projections, in all corners of the projection should be about 1.5 times greater than the dimension recovered format  $n \times n$  (in one dimension -  $n$ );
- The final stage of the procedure is performed each convolution kernel with low-frequency projection and rear projection - the restoration of the desired functional distribution for each of the circular zone  $D_i$ . Then, using interpolation methods, recalculates the required data in the annular zone.

Thus, based on their characteristics produce raw projection data for IRT problems using constellation of navigation satellites and clusters of NS, it is easy to define the conditions to be met by hardware modules installed on each CubeSat:

- On each NS to be installed highly sensitive receivers for frequencies (for the reception of signals from GPS devices);
- Each NS should contain module to determine its location on the orbit by GPS data;
- Computational modules are installed on each NS must provide the required source data processing speed. Perform the appropriate procedures for conversion of projection data of a fan in an orthogonal geometry, redefine the number of chords to fill each round of reconstruction  $D_i$ . Achieving these goals is only possible when using multiprocessor computing modules equipped with the appropriate amount of RAM.
- To send the original one-dimensional projection array to the CPU required transmitters with a wide bandwidth. From the standpoint of reliability, each satellite must contain two such from each transmitter is capable of transmitting data, e.g., each pair of GPS.
- To coordinate referred modules each NS must contain the microprocessor control units.
- Each NS must be a certain way to orient in space and place on a circular orbit at equal distances from each other, for this purpose in each of the NS shall be provided accommodation compact three-axial gyroscope, and a set of actuators.

In order to accommodate the listed modules and accessories to the NS must use 7U CubeSat format. This format is an assembly CubeSat as a Makarov three-dimensional cross and contains 7 1U CubeSat format modules. This design allows you to install it on additional self-extracting solar panels, self-parabolic reflector antennas for receivers and transmitters of GPS signals to communicate with MCC. To increase the power available, the surface of 3D 7U CubeSat satellite is covered with solar panels. Conclusion and startup cluster consisting of those of the NS by means of "Soyuz", in which a transition compartment for up to 4 transport and display systems, each of which contains 4 3D 7U CubeSat format apparatus. The launch of clusters consisting of such NS can be carried out with the help of the Soyuz, in the transition compartment of which it is possible to arrange up to four deployment systems, each of which contains four devices of the 3D 7U CubeSat format.

#### 4. Mathematical model of radio tomography analysis of ionospheric parameters using GPS system and nanosatellites cluster

The effectiveness of the proposed methods radio tomography of the ionosphere using navigation systems such as GPS / GLONASS and LEO NS clusters depends on many factors, such as a function of the gravitational field potential in the plane of the NS orbit, the curvature of the trajectory of the sounded antenna beam as a function of the refractive index, the functional variation of the amplitude, frequency, phase of the beam, etc. Take into account the impact of such factors on the adequacy of the tomographic reconstruction of ionosphere parameters of procedures you can use mathematical modeling techniques. In general, the mathematical model of the ionosphere parameters radio tomography analysis should take into account the following factors:

1. The direct and inverse problems of ionospheric radio sounding is necessary to determine the amplitude change, phase (frequency) radio waves to track the satellite - satellite. To this must be set the dependence of the refractive index of the height  $n(h)$ . This problem has been studied in detail by the authors [10]. The geometry of this problem is shown in Fig. 2 a). The points  $L$ ,  $G$  are located at altitudes  $H_l$ ,  $H_g$  of satellites. Earth Center designated  $O$  point, in general, radial line  $LTG$  in point  $T$  passes at the minimum height above the Earth's surface  $H$ . The radial line at a higher level, in areas  $LL_1$  and  $GG_1$  - straight and in the field  $L_1G_1$  because of the influence of media is rejected by the angle of refraction  $\xi$ . Assuming that the ionosphere is a local spherically symmetric environment can be negligible horizontal gradients of environment (near the point  $T$  - line  $L_1G_1$ ), and assume that the rate  $n(r)$  depends only on the distance  $OC = r = a + h$ . Introduce:  $h$  - the height of a point  $a$  - the radius of the Earth;  $\theta$  - a central angle,  $g_g = a + H_g$ ,  $r_l = a + H_l$ ,  $r_i = a + H$  respectively, of the distance between  $OG$ ,  $OL$  and  $OT$ . For a spherically symmetric medium following formulas [11]:

$$n(r)r \sin \gamma - const, \quad (1)$$

$$P\Delta S - const, \quad (2)$$

here  $\gamma$  - the angle between  $r$  and the unit vector radial lines  $l_0$ . Expression (2) determines the flux density in the cross section of the ray tube  $\Delta S$ , which makes it possible to calculate the change  $P$  due to refraction. Since the refractive index differs only slightly from unity, it is customary [11] to use parameter  $N = n - 1$  which depends on the pressure  $P_a$ , temperature  $T$  and humidity  $w_a$  as follows:

$$N = \frac{77.6}{T} \left( P_a + \frac{481w_a}{T} \right) \cdot 10^{-6}. \quad (3)$$

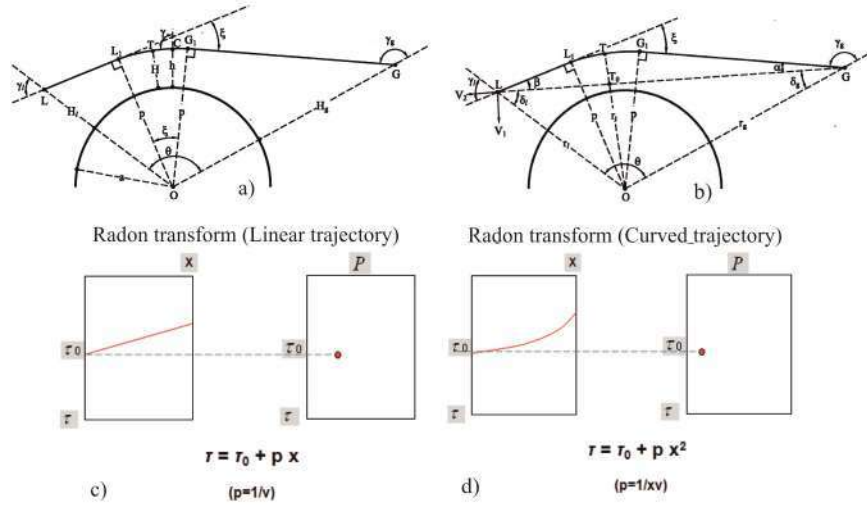


Fig. 2. Illustrate the effect of the ionospheric refractive index of the sounding antenna beam and the Radon image from sounding geometry.

In this model, the height profile of the given refractive index can be approximated by:

$$N(h) \approx N_0 \exp(-b_1 h). \quad (4)$$

when:

$$b_1 = 0.1 \cdot \ln \left( \frac{9.2 \cdot 10^{-5}}{N_0} \right). \quad (5)$$

Since the real profile  $N(h)$  differs from (4), it can be used as an approximation:

$$N(h) = N_0 \exp(-a_1 h^2 - b_1 h). \quad (6)$$

there  $a_1, b_1 - const$  [12].

The above plasma refractive index, for high frequencies:

$$N_p(h) = -\chi N_e f^{-2}, \quad \chi = 40.4 (SI). \quad (7)$$

For the upper part of the ionosphere  $N_e = N_m \exp[-b_2(h - h_m)]$ , here  $N_m$  - the electron density in the main ionospheric maximum at altitude  $h_m$ . Below  $h_m$ , you can use the approximation:

$$N_e(h) \approx N_m \left[ 1 - \left( \frac{h_m - h}{c_2} \right)^2 \right], \quad (8)$$

Where  $c_2$  - the nominal thickness of the lower part of the ionosphere.

The paper [11] shown that for beam line in a spherically symmetric medium satisfies the relation:

$$\operatorname{tg} \gamma = \frac{p}{(r^2 n^2(r) - p^2)^{1/2}} \quad (9)$$

which implies that the radial line defined by the altitude  $n(h)$  profile and the parameter  $p$ .

The radius of curvature of the beam in a spherically symmetric medium:

$$R_0 = \frac{a + h}{\sin \gamma + (a + h) \frac{d\gamma}{dh} \cos \gamma}. \quad (10)$$

Refraction angle:

$$\xi = 2 \int_H^\infty \left( \frac{d\gamma}{dh} + \frac{d\theta}{dh} \right) dh, \quad \rightarrow \quad \xi = -2 \int_H^\infty \frac{1}{n} \frac{dn}{dh} \operatorname{tg} \gamma dh. \quad (11)$$

For model experiments it is advisable to use the approximation:

$$\xi = N_0 (2\pi ba)^{1/2} \exp(-bH) \quad (12)$$

**Note:** In the atmosphere of refraction does not depend on the wavelength and angle  $\xi$  in the ionosphere proportional to the square of the wavelength. The vertical gradient of the electron density:

$$\frac{dn}{dh} = -\chi f^{-2} \frac{dN_e}{dh}.$$

The above expressions for the refractive index, the radius of curvature, and the refraction angle make it possible to simulate in detail the ray lines along satellite-satellite lines. Use as a projection data values integrated radio signal intensity along a curved radial lines, including in cases where NS is provided with receivers are on the horizon line of sight (see the eclipsing of reference. Fig. 2 c) and d)).

2. Refraction attenuation, frequency and phase changes sounding radio waves make it possible within the framework of the model of the ionosphere sounding radio tomography obtain more accurate data about Radon icons is formed in the projection data generated. This, in turn, all other things being equal, allows a high degree of precision to carry out reconstruction of the desired functional distributions. Consider ray tube at the point  $G$  (see. Fig. 2 b)) with an angular size  $d\gamma_g$  in the perpendicular plane of its size  $d\chi$ , and calculate its size at the point  $L$ . From the above geometry seen that  $LL_2 = r_1 d\theta$  the linear dimension of the ray tube at a point  $L$  equal  $LL_3 = r_1 \cos \gamma_1 d\theta$  to one can show that the cross-sectional area at the point  $L$  is:

$$S_1 = r_1^2 \sin \theta \cos \gamma_1 d\theta d\chi. \quad (13)$$

In the absence of refraction ray tube would have a cross-section in the region of point  $L$ :

$$S_0 = L^2 \sin \gamma_g d\gamma_g d\chi, \quad (14)$$

here  $L = \sqrt{r_1^2 + r_g^2 - 2r_1 r_g \cos \theta}$  - the distance between points  $L$ ,  $G$ . Define refractive weakening as the ratio of the power flux density in the presence of refraction  $P_1$  and its absence  $P_0$ :

$$X = \frac{P_1}{P_0} = \frac{S_0}{S_1} = \frac{L^2 \sin \gamma_g d\gamma_g}{r_1^2 \sin \theta \cos \gamma_1 d\theta}. \quad (15)$$

Authors [11] lead (15) to mean:

$$X = \frac{p(r_1^2 + r_g^2 - 2r_1 r_g \cos \theta)}{r_1 r_g \sin \theta \left[ (r_1^2 - p^2)^{1/2} + (r_g^2 - p^2)^{1/2} - \frac{d\xi}{dp} (r_1^2 - p^2)^{1/2} (r_g^2 - p^2)^{1/2} \right]}. \quad (16)$$

In (16) it is assumed that the radio sounding  $n_g = n_l = 1$ . For small angles of refraction at high frequencies, the sounded beam:

$$X = \frac{p(L_g + L_l)^2}{r_g r_l \sin \theta \left( L_g + L_l + L_g L_l \frac{d\xi}{dp} \right)}. \quad (17)$$

From (17) it follows that in the case of sounding system consisting of HEO - LEO satellites  $L_g \gg L_l$ :

$$X \approx \frac{p}{r_l \sin \theta \left( 1 + L_l \frac{d\xi}{dp} \right)}. \quad (18)$$

For small angles of refraction  $p \approx r_l \sin \theta$  that occurs for the chord data in the center of the fan-expression (18) sensing beam can be written as:

$$X \approx \left( 1 - L_l \frac{d\xi}{dp} \right)^{-1}. \quad (19)$$

In the exponential approximation  $N(h)$  we can use the ratio:

$$X \approx \left[ 1 + bL_l (2\pi ba)^{1/2} N_0 \exp(-bH) \right]^{-1}. \quad (20)$$

Doppler frequency change in the ionosphere  $\Delta f_s$ :

$$\Delta f_s = \lambda^{-1} (V_2 \cos \beta + V_1 \sin \beta), \quad (21)$$

where  $V_1, V_2$  - on the satellites velocity projection beam line at the point  $L$ . If there is no change in the frequency of the ionosphere due to the satellite motion is:

$$\Delta f_0 = \lambda^{-1} V_2. \quad (22)$$

Thus, the change in frequency only through the ionosphere:

$$\Delta f = \Delta f_s - \Delta f_0 = \lambda^{-1} [V_2 (\cos \beta - 1) + V_1 \sin \beta]. \quad (23)$$

That is, ionospheric frequency change depends from the angle  $\beta$  and velocity of the satellite components  $V_1, V_2$ . For small angles of refraction and  $r_g \gg r_l$  can be used:

$$\Delta f \approx \lambda^{-1} V_1 \xi. \quad (24)$$

To account for the phase change of the probe beam  $\Delta \varphi = \varphi - \varphi_0$ , where  $\varphi_0$  - the phase for the curved path - phase for the rectilinear propagation of the beam, i.e., when the ionosphere is absent in this model, the following relationships are used:

$$\varphi = \frac{2\pi}{\lambda} \int_G^L n dl = \frac{2\pi}{\lambda} \left( \int_{r_l}^{r_g} \frac{n^2 r dr}{(n^2 r^2 - p^2)^{1/2}} + \int_{r_l}^{\eta} \frac{n^2 r dr}{(n^2 r^2 - p^2)^{1/2}} \right), \quad \varphi_0 = \frac{2\pi}{\lambda} \left[ (r_g^2 - p_0^2)^{1/2} + (r_l^2 - p_0^2)^{1/2} \right], \quad (25)$$

here  $p_0$  - the minimum distance from the line  $GL$  to the center of the Earth. Importantly, ionospheric value  $\Delta \varphi$  proportional to the wavelength, and the refractive attenuation  $X$ , frequency change  $\Delta f$  and phase  $\Delta \varphi$  can be expressed in terms of the angle of refraction, so depending  $X(H)$ ,  $\Delta f(H)$ ,  $\Delta \varphi(H)$  interconnected. This approximate relationship, provided that  $V_1$  it is not time-dependent, is given [11]:

$$X = \left[ 1 - \left( \frac{L_l c}{f V_1^2} \right) \cdot \frac{d(\Delta f)}{dt} \right]^{-1},$$

$$X = \left[ 1 - \left( \frac{L_l c}{2\pi f V_1^2} \right) \cdot \frac{d^2(\Delta \varphi)}{dt^2} \right]^{-1}. \quad (26)$$

**Note:** If the solution of inverse problems of monitoring ionospheric parameters from the experimentally obtained dependences  $\Delta \varphi(t)$ ,  $\Delta f(t)$ ,  $X(t)$  is necessary to reconstruct the profiles  $N(t)$ ,  $N_e(t)$ , in such cases, the satellite coordinates and their speed must be known.

3. The greatest difficulty in solving inverse problems of monitoring ionospheric parameters using navigation satellites - sources of sounding signals and LEO NS clusters - sounding recording radio pulses is the problem of reconstruction of the desired functional distributions  $N(r, t)$ ,  $N_e(r, t)$  in sensing zone. Recall that the sensing zone for 2D problem is an annular carrier (in the form of a spherical layer for applications of direct 3D - reconstruction), determined by the diameter and the orbits of the navigation satellites and NS clusters. To reconstruct the desired functional  $N_e(r, t)$  type distributions by fast algorithms based on the Radon transform must be reformulated Radon's theorem for the carrier ring, or to fill a circular annulus reconstruction carriers  $D_i$ , as it is shown in Fig. 2. In the first case will have to deal with very large data sets while the final resolution in the reconstructed image will be low, of the order (200×200) km, which is much lower than with traditional methods tomography [1]. The second case is for these preferred methods, despite the fact that there is a problem in calculation of the probing beam parameters at the boundaries of each circular zone. Recall that for the inverse Radon transform function  $N_e(r, t)$  at any given time  $t$  can be written as:

$$[R]^{-1} N_e(r, \varphi) = \frac{1}{2\pi^2} \int_0^{\pi} \int_{-\infty}^{+\infty} \frac{1}{r \cos(\theta - \varphi) - l} \frac{\partial N_e(l, \theta)}{\partial l} dl d\theta. \quad (27)$$

Represent the action of the operator  $[R]^{-1}$  in the form of a sequence of simpler operators  $R^{-1} = B H_l D_l$ . The integral in equation (27) is improper, since  $r \cos(\theta - \varphi) = l$  it diverges. It can be calculated by converting the integral in the sense of the Cauchy principal value, after entering the designation  $r \cos(\theta - \varphi) = \tau$ :

$$[H_l(N_e)](l, \theta) = -\frac{1}{\pi} \lim_{\varepsilon \rightarrow 0} \left\{ \int_{-\infty}^{l-\varepsilon} \frac{N_e(\tau, \theta)}{l-\tau} d\tau + \int_{l+\varepsilon}^{+\infty} \frac{N_e(\tau, \theta)}{l-\tau} d\tau \right\} \quad (28)$$

Calculate the integral can, if present in the Hilbert transform (28) in the form of a convolution of two functions  $N_e'(\tau, \theta)$  and  $\chi(l) = -\frac{1}{\pi l}$ , i.e., written in the form of relation:

$$[H_l(N_e')]_l(l, \theta) = [N_e' * \chi]_l(l, \theta). \quad (29)$$

Run directly to the operation (29) cannot be performed, so we approximate the function  $\chi(l)$  by a function  $\chi_A(l)$ , so that the following condition:

$$\lim[N_e' * \chi_A]_l(l, \theta) = [H_l N_e']_l(l, \theta) \quad (30)$$

This approach, defined by (30), as is known [9] is called regularization method and the functions set  $\{\chi(l) / A > 0\}$  - is called a family of regularizing functions. As it is known the integral convolution of two functions can be replaced by multiplication of their Fourier spectra in the frequency domain, i.e.:

$$[N_e' * \chi]_l(l, \theta) = F_{\omega_l}^{-1}[[F_l N_e'](\omega_l) \cdot [F_l \chi](\omega_l)], \quad (31)$$

here:

$$[F_l(N_e')]_l(\omega_l) = \int_{-\infty}^{+\infty} N_e'(l, \theta) e^{-i2\pi\omega_l l} dl, \quad [F_l \chi](\omega_l) = \int_{-\infty}^{+\infty} \chi(l) e^{-i2\pi\omega_l l} dl. \quad (32)$$

A direct calculation of (31) is difficult, as the final function  $N_e'(l, \theta)$  has an infinite range. A simple "truncation spectrum" is unacceptable, because it leads to the generation of noise due to the Gibbs phenomenon. In such cases, it decided to do, taking into account (30), as follows:

$$\chi_A(l) = \int_{-\frac{A}{2}}^{\frac{A}{2}} [F_l \chi](\omega_l) \cdot W(\omega_l) e^{2\pi i \omega_l l} d\omega_l. \quad (33)$$

Window function  $W(\omega_l)$  must satisfy the following conditions:

$$W(\omega_l=0) = 1; \quad W(\omega_l) = 0; \quad |\omega_l| \geq \frac{A}{2} \lim_{A \rightarrow \infty} W(\omega_l) = 1; \quad W(\omega_l) \quad 0 \leq \omega_l \leq \frac{A}{2}; \quad \lim_{A \rightarrow \infty} \omega(l) = 0, \rightarrow \omega(l) = \frac{1}{2\pi} \int_{-\frac{A}{2}}^{\frac{A}{2}} W(\omega_l) e^{2\pi i \omega_l l} d\omega_l$$

Fourier spectrum of the function  $\chi(l)$  is obvious:

$$[F_l \chi](\omega_l) = -\frac{2}{\pi} \int_0^{+\infty} \frac{\sin 2\pi\omega_l l}{l} dl = -\text{sgn}(\omega_l)$$

in view of

$$\omega(l) = \frac{1}{2\pi} \int_{-\frac{A}{2}}^{\frac{A}{2}} W(\omega_l) e^{2\pi i \omega_l l} d\omega_l.$$

Approximating function  $\chi_A(l)$  can be represented as:

$$\chi_A(l) = -2 \int_0^{\frac{A}{2}} W(\omega_l) \sin(2\pi\omega_l l) d\omega_l. \quad (34)$$

If we calculate the limit of approximating function (34), i.e.:

$$\lim_{A \rightarrow \infty} \chi_A(l) = \lim \left\{ \frac{1}{\pi l} [W(\omega_l) \cos(2\pi\omega_l l) - \int_0^{\frac{A}{2}} W^1(\omega_l) \cos(2\pi\omega_l l) d\omega_l] \left( \frac{A}{2} \right) \right\} = -\frac{1}{\pi l}, \quad (35)$$

it becomes apparent that the condition (30) holds, i.e.:

$$[N_e' * \chi_A]_l(l, \theta) = \int_{-\infty}^{+\infty} N_e'(\tau, \theta) \chi(l - \tau) d\tau. \quad (36)$$

Note that the desired function  $N_e(r)$  is defined in a finite region - property section is said to be defined on a finite medium, therefore, the area of its existence can be set in a range bounded by a circle of radius  $R$ , i.e.:  $x^2 + y^2 = R^2$ ;  $\rho_{l, \theta} = 0$ ;  $|l| \geq R$ . With this in mind, we integrate the right side of (40) by parts:

$$[N_e' * \chi_A]_l(l, \theta) = \int_{-\infty}^{+\infty} N_{e\theta}(\tau) \chi'_A(l - \tau) d\tau, \quad (37)$$

We now calculate the derivative of the function  $\chi_A$  under the integral sign in (41), i.e.:

$$\chi'_A(l) = -4\pi \int_0^{\frac{A}{2}} \omega W(\omega_l) \cos(2\pi\omega_l l) d\omega_l. \quad (38)$$

We introduce the notation  $h(l) = \chi'_A(l)$ , then we can write:

$$N_e(l, \theta) = [N_e * h]_l(l, \theta). \quad (39)$$

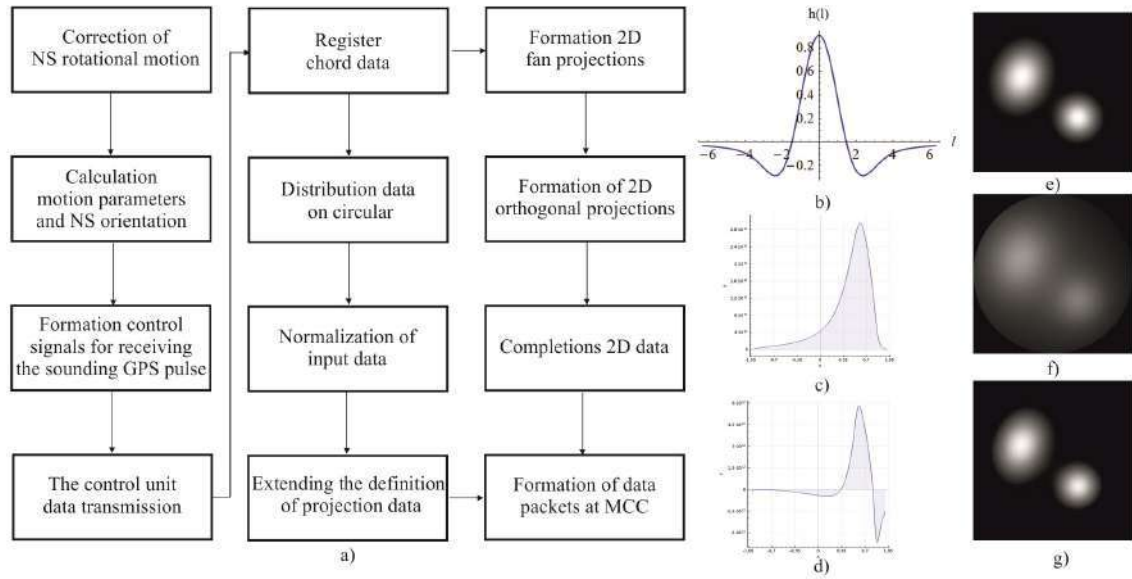


Fig. 3. Illustrations of the stages of mathematical modeling procedures for solving the reconstruction inverse problem of distributions  $N_e(r)$  when used GPS and NS cluster for radio sounding.

Formula (39) in accordance with expression (29) is an approximation to the Hilbert transform provided  $A \rightarrow \infty$ . We use the inverse projection operator, and write:

$$N_e(r, \varphi) = [B(N_e)](r, \varphi) = -\frac{1}{2\pi} \int_0^\pi N_e(r \cos(\theta - \varphi), \theta) d\theta \quad (40)$$

Thus, we illustrate one possible way of approximating the inverse Radon transform, which reduces to two procedures: 1) projection function is convolve with the function  $h(l)$  of certain expressions (39) and (40) respectively. 2) Perform the back-projection procedure.

**Note:** The selection of the appropriate "window" is usually performed fairly subjectively, based on a visual assessment of the quality of the reconstruct image. In this case, a compromise between resolution and noise level is usually sought.

For the purposes of IRT the authors developed low-frequency "window" for fast convolution algorithm applied to the methods of radio sounding using satellite clusters, which are described in detail in [13, 14, 15].

To study the possibilities of using NS clusters capable of detecting the sounding pulses from GPS navigation satellites for solving tomographic problems for reconstructing the desired functional distributions, for example, electron density, the authors carried out a full cycle of mathematical modeling of tomographic reconstruction procedures taking into account the factors considered above. The results are shown in Fig. 3.

Fig. 3 a) shows a block diagram of the control program modules and data pre-processing for the microprocessor-based systems installed on the NS. Low-frequency core view is shown in Fig. 3 b), Fig. 3c) and d) respectively show the generated projection range for the result of convolution of the projection to the core. Fig. 3 e) shows the model function, and Fig. 3 f) and g) of reconstruction by using different cores.

## 5. Conclusion

Creating a mathematical model of ionospheric parameters radio tomography procedures using navigation systems types GPS / GLONASS satellites used as sources of sounding signals and NS cluster, whose satellites are equipped with highly sensitive, stable receivers showed that this approach is a highly effective method of ionospheric research. Method of analysis of ionospheric, allowing near real-time analysis of the distribution function, e.g., the electron density, was proposed by authors. This method is relatively simple to implement, a methodological error of reconstruction by four NS constellation is (15 - 20) % and can be reduced by increasing the number of the NS in the cluster.

## References

- [1] Kunitsyn V, Tereshchenko ED, Andreeva ES. Radio-tomography of the ionosphere. Moscow: fizmatlit, 2007.
- [2] Andreeva ES, Galinov AV, Kunitsyn VE. Tomographic reconstruction of the ionospheric ionization failure. Pis'ma v ZhETF 1990; 52(3): 783–785.
- [3] Bust G, Mitchell CH. History, current state, and future directions of ionospheric imaging. Reviews of Geophysics, 2008; 46 RG1003:1–23.
- [4] Nesterov IA, Kunitsyn VE. GNSS radio tomography of the ionosphere: the problem with essentially incomplete data. Adv. Space Res. 2011; 47: 1789–1803.

- [5] Kunitsyn VE, Nesterov IA, Padokhin AM, Tumanova Yu S. Radio-tomography of the ionosphere on the basis of GPS / GLONASS navigation systems. *Radiotekhnika i elektronika* 2011; 56(11): 1285–1297.
- [6] Kunitsyn VE, Tereshchenko ED, Andreeva ES, Nesterov IA. Satellite radio sounding and radio-tomography of the ionosphere. *Uspekhi fizicheskikh nauk* 2010; 180(5): 548–553.
- [7] Phylonin OV, Talyzin YuB. Mathematical modeling of the processes of studying planetary atmospheres with the help of colonies of small satellites. *Proceedings III All-Russia scientific and technical conference "Actual problems of rocket and space technology"*. Samara, 2013; 245–248.
- [8] Phylonin OV. Inverse ill-posed problems in space research. Samara, 2014; 478 p.
- [9] Phylonin OV. Low-angle reconstructive tomography in a physical experiment. Saarbrücken, Germany: Palmarium Academic Publishing, 2012; 606 p.
- [10] Kalashnikov IE, Matyugov SS, Yakovlev OI. Influence of the ionosphere on the parameters of the signal in the radio decomposing of the Earth's atmosphere. *Radiotekhnika i elektronika* 1986; 31(1): 56.
- [11] Yakovlev O, Pavel'ev A, Matyugov S. Satellite monitoring of the Earth: Radiospheric monitoring of the atmosphere and ionosphere. M.: Knizhnyi dom LIBROKOM, 2010; 208 p.
- [12] Bilitza D, McKinnell L-A, Reinisch B, Fuller-Rowell T. The International Reference Ionosphere (IRI) today and in the future. *Journal of Geodesy* 2011; 85: 909–920.
- [13] Phylonin OV, Nikolaev PN. Monitoring of the state of the earth's ionosphere by a group of small satellites. *Vestnik Samarskogo universiteta. Aerokosmicheskaya tekhnika, tekhnologii i mashinostroenie* 2016; 15(1): 132–138.
- [14] Phylonin OV, Belokonov IV, Nikolayev PN. Mathematical Modeling of Radio Tomographic Ionospheres Monitoring Via Satellite Constellation. *Scientific and Technological Experiments on Automatic Space Vehicles and Small Satellites. Procedia Engineering* 2015; 104: 131–138.
- [15] Phylonin OV, Belokonov IV. Investigation of the possibilities of spatial reconstruction of the parameters of the electronic component of the ionosphere using navigation satellites. *Izvestiya of the Samara Scientific Center of the Russian Academy of Sciences* 2014; 16(4-1): 47–53.



# Modeling control over large space structure on geostationary orbit

V.V. Salmin<sup>1</sup>, A.S. Chetverikov<sup>1</sup>, K.V. Peresypkin<sup>1</sup>, I.S. Tkachenko<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The paper considers the problem of control over a large space structure (LSS) control at a given station on a geostationary orbit (GSO). An observation spacecraft with diffractive optical elements (DOE) is taken as an example of a large space structure. Various perturbing factors influence the motion of a LSS along GSO, most notably solar radiation pressure. Two problems are considered: control over the motion of the center of mass, and control of the motion in relation to the center of mass. The paper gives the results of modeling the process of LSS control, based on developed control algorithms.

*Keywords:* modeling; large space structure; low thrust; terminal control; geostationary orbit; solar radiation pressure; controlling torque

## 1. Introduction

The paper considers an observation spacecraft with diffractive optical elements (DOE). The optical scheme is taken as in DAPRA's MOIRE project [1]. This observation spacecraft can be defined as a large space structure: DOE assembly diameter is 10 meters, and the distance from the optical elements to the body of the spacecraft is 60 meters. The design of the frame that connects DOE to the body is discussed in detail in [2].

Further work on the project resulted in development of a hood that prevents any light that doesn't come from the observed object from reaching the optical elements. The hood is a system of surfaces that are so located that they shield the optical element in the body of the spacecraft from any rays that do not come from the observed object, and also to shield the rear surface of DOEs. The observation spacecraft with the hood is shown in Figure 1.

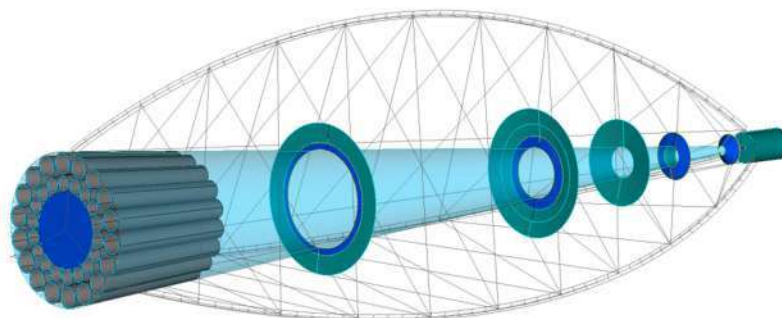


Fig. 1. Observation spacecraft with envelope- and ring-shaped hoods for diffractive optical elements.

The orbital movement of a space structure with such mass and dimensions will be highly influenced by various perturbing factors, particularly by torque and momentum created by solar radiation pressure. In addition, as the spacecraft is moving along its orbit, it will experience quite noticeable perturbing accelerations from the gravity fields of the Sun and the Moon. For these reasons, keeping the orbital structure at the desired station on GSO requires constant trajectory correction. In addition, attitude control is necessary to keep the roll axis of the optical structure under consideration pointed to the Earth at all times.

## 2. Modeling the algorithms of terminal control over the motion of the LSS's center of mass.

### 2.1. Setting the goals of control.

The goal of control is to keep the final state deflection vector  $\Delta X_K$  within the allowable area  $G_D$ . The adjustment maneuver is performed by applying transversal torque with the help of a low thrust electric propulsion (EP) unit. The transversal torque creates acceleration  $a_T$  along the transversal.

This problem is stated as an optimal control problem with the functional

$$I = \Delta x_K^T \Lambda \Delta x_K \rightarrow \min, \quad (1)$$

where  $\Lambda$  is the constant coefficients matrix.

The control is structured as a sequence of lengths of powered and unpowered flight  $u = \{\tau_1, \dots, \tau_i, t_{M1}, \dots, t_{Mn}\}^T$

Deviation of the orbit's semi-major axis  $\Delta A$  of the orbit is equivalent to the deviation of the orbit time of the spacecraft  $\Delta T = T - T_3$ . Here the orbit time of the spacecraft on the geostationary orbit equals star day  $T_3 = 86164.09$  c. In addition, the spacecraft's position on the orbit is determined by longitude  $\lambda$ , which differs from the required value of the station longitude  $\lambda_p$  by  $\Delta \lambda = \lambda - \lambda_p$ .

2.2. Solving the terminal control problem with a multistep algorithm.

The terminal control problem is solved with the help of a multistep algorithm with adjustment of control parameters. Let the control law be set by a sequence of thrust lengths, which is taken as decreasing, and defined by the expression [3]:

$$\tau_i = a \cdot \left[ 1 - \left( \frac{i-1}{n} \right)^b \right], \tag{2}$$

where  $i, n$  are the number of the adjustment and the total number of adjustments respectively;  $a, b$  are the parameters that characterize the law of decreasing lengths of thrusts.

Then the problem of determining the optimal control law is reduced to a two-parameter optimization problem, which is stated in the following way: for the set initial values of the orbital elements, transversal acceleration  $a_T$ , number of corrections  $n$ , lengths of unpowered flight  $t_{II}$  one must find such parameters  $a$  and  $b$  that would ensure the minimum of the functional (see formula (1)).

A peculiar feature of the algorithm presented in this paper is that the control parameters  $a$  and  $b$  are found as the result of minimization of the functional (1) and at the same time, for better precision, the relation of the functional to the parameter  $a$  is approximated by the least squares method. When the control is adjusted (during motion modeling with allowances for perturbations) at every unpowered leg the number of steps  $n$  is also adjusted.

A series of calculations of the control laws for transfer of an EP-powered spacecraft into a given station by longitude and orbit time have been carried out. The delta V expense, depending on the initial value of deviation by orbit time ( $\Delta T_0 = 300 \dots 1000$  c) ranges from 4 to 14 m/s.

2.3. Results of modeling terminal control with the help of a multi-step algorithm.

Tables 1 and 2, and figures 3 and 4 give sample calculations of control parameters (see Formula (2)) for a given station on GSO, for two cases, with and without adjustment. Figure 3 represents phase trajectory of a spacecraft transfer to a given GSO station without adjusting control parameters. The maneuver was modeled with the help of equations of motion in equinoctial elements.

Figure 2 shows the regions of deviation of the final orbit time and longitude divergence values for the EP-powered observational spacecraft transfer control law, without adjustment of control parameters, and with stepwise adjustment at every stage of the transfer. The process was modeled with allowances for perturbing accelerations from the gravitational fields of the Sun, the Moon, the Earth, and solar radiation pressure.

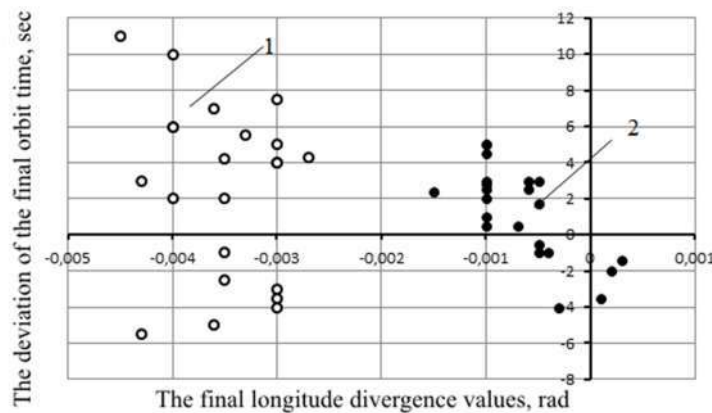


Fig. 2. The area of the final values of  $\Delta T$  and  $\Delta \lambda$  1) without, and 2) with adjustment of control parameters.

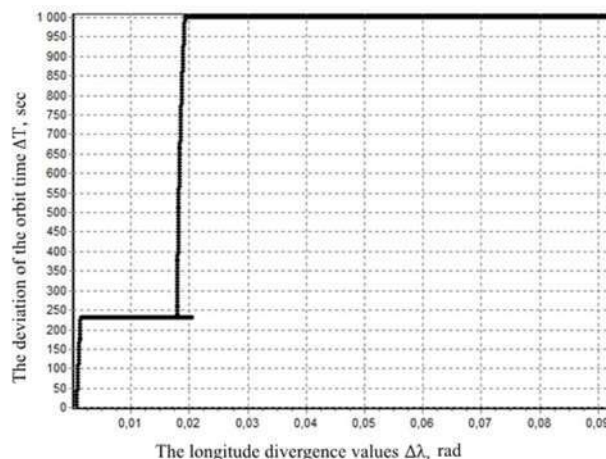


Fig. 3. Phase trajectories of the transfer of EP-powered spacecraft to a given GSO station.

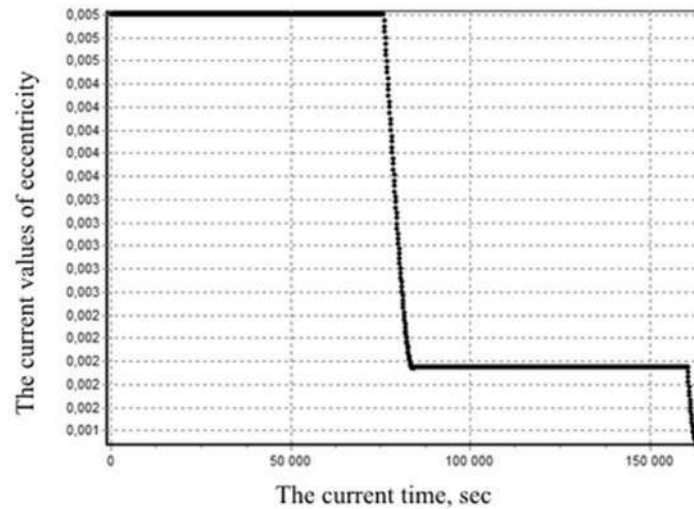


Fig. 4. Changes in eccentricity during transfer of an EP-powered spacecraft to given station at GSO.

 Table 1. Control parameters for transfer to a given station at GSO ( $\Delta T_0 = 1000$  c,  $a_T = 0,001$  M/c<sup>2</sup>,  $t_{II} = 76000$  c,  $\Delta\lambda_0 = 0,092$  rad,  $\lambda_P = 75,1^\circ$ ).

$N\hat{\rho}$	$n$	$a$	$b$	$V_{XK}$ , km/s	$\Delta T_K$ , sec	$\Delta\lambda_K$ , rad	$\Delta e_K$
Without refinement of control parameters							
1	2	8460,5	0,7	0,012	6	-0,0025	0,0018
With refinement of control parameters.							
1	2	8441,5	0,7				
2	2	3105,7	0,1	0,012	2	-0,0005	0,0010
3	1	76,6	2,2				

 Table 2. Lengths of powered flight legs ( $\Delta T_0 = 1000$  c,  $a_0 = 0,001$  M/c<sup>2</sup>,  $t_{II} = 76000$  c,  $\Delta\lambda_0 = 0,092$  rad).

$\tau_1$ , c	$\tau_2$ , c	$\tau_3$ , c
Without adjustment of control parameters		
8460,5	3252,5	-
With adjustment of control parameters		
8441,5	3105,7	76,6

### 3. Modeling rotation of a large space structure about its center of mass

#### 3.1. Setting of the problem

Accelerations on the orbit are determined by maneuvers of positioning the spacecraft to point at the object of observation. The values of the turn angles depend not only on the mutual position of observed objects, but also on perturbing factors that influence the spacecraft's angular position in relation to the Earth. One such factor is the rotation of the spacecraft on its orbit: if the spacecraft does not itself rotate in relation to the inertial system of coordinates, then as it moves along the orbit the optical axis will turn in relation to the Earth. This factor can be negated by ensuring the spacecraft's rotation about its axis at the same rate as the spacecraft's orbit time.

However, other perturbations will alter the angular spin rate of the spacecraft, shifting its optical axis away from the Earth. On a geostationary orbit, with the dimensions of the spacecraft under consideration, the biggest perturbing factor will be solar radiation pressure. Most of the spacecraft's mass is located in its body; in fact, the center of mass is only 3.25 m. from the body. Yet, the main center of solar radiation pressure, according to Figure 1, will be in the neighborhood of the diffractive optical elements, which will lead to production of a substantial moment of rotational force.

To estimate the value of the solar radiation induced turning torque, let us assume that there is no reflection and the solar radiation is completely absorbed by the spacecraft. Then the direction of the solar radiation pressure will coincide with the direction of the sunlight. The area of the radiation beam that falls on the spacecraft and the distance from center of mass to center of pressure depend on the angle of exposure. The value of the rotating torque will be found as the multiplication of these values, and is represented in Figure 5.

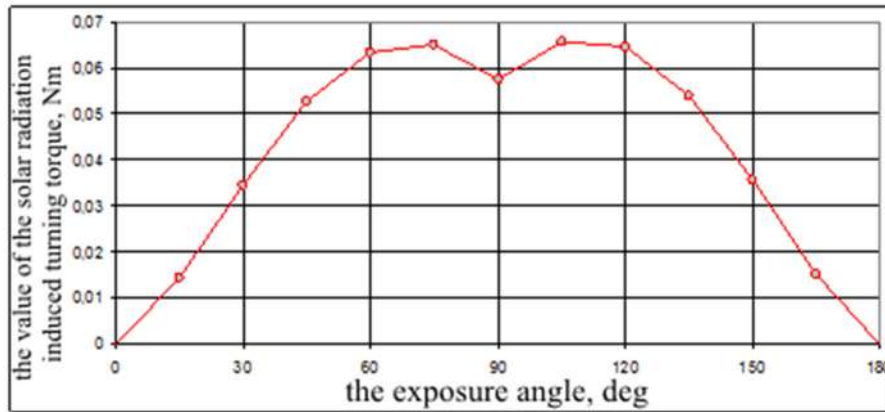


Fig. 5. Relation of rotational torque to exposure angle.

For the structure under consideration, at the maximum torque value the spacecraft will turn by  $0,67^\circ$  in 10 minutes. This is a significant perturbation of motion, and without correction the spacecraft will quite soon turn away from the Earth so much that imaging will be impossible.

Frequent correction is necessary to keep the spacecraft in proper position for imaging. During correction, the controlling impulse will cause vibrations in the spacecraft's frame. Movement of diffractive optical elements in relation to the body creates problems for the spacecraft's optical system. Therefore, it is necessary to select such type of controlling impact that would ensure that the vibrations are damped soon enough not to interfere with the mission of the spacecraft. The control problem for such a large space structure must be solved with allowances for elasticity in its design. Turning of the spacecraft is modeled with the help of the finite elements method.

During the turn, the spacecraft rotates in the inertial system of coordinates, and elastic vibrations occur in its frame. The amplitudes of the vibrations are expected to be small, i.e., not powerful enough to change the elastic and inertial properties of the spacecraft's structure. If the spacecraft is considered in a system of coordinates tied to it, then geometrical non-linearity is absent. Modeling is performed in the inertial system of coordinates for convenience of setting the boundary conditions and analysis of the results. In this case, when the finite-element model of the design turns, one must re-calculate the matrix of masses, dampening, and rigidity for the new orientation of finite elements in space. However, the angle of the turn is small, and the related changes in the matrixes are minor. Let us neglect the impact of the turn of the spacecraft on the matrix coefficients and consider the system as linear. Let us now apply the MSC Nastran linear transition analysis. This analysis performs numerical integration of the main dynamic equation in time [4]:

$$[M] \cdot \{u\} + [C] \cdot \{u\} + [K] \cdot \{u\} = \{P(t)\}, \quad (3)$$

where  $\{P(t)\}$  is the nodal forces vector;  $j$  is the number of the integration step;  $\{u\}$  is the vector of the nodal movement of the model;  $\{u\}$  is the model's nodal velocity vector;  $\{u\}$  is the model's nodal accelerations vector;  $[C]$  is the dampening matrix. The velocities and accelerations are expressed through motion via central-differential approach:

$$\{u\}_j = \frac{\{u\}_{j+1} - \{u\}_{j-1}}{2 \cdot \Delta t}; \quad \{u\}_j = \frac{\{u\}_{j+1} - \{u\}_{j-1}}{2 \cdot \Delta t}, \quad (4)$$

where  $\Delta t$  is the step of integration in time. Then, with averaging the nodal forces vector for three neighboring steps in time, the system (see (Formula3)) is transformed to the following view:

$$[A_1] \cdot \{u\}_{j+1} = [A_2] \cdot \{u\}_j + [A_3] \cdot \{u\}_{j-1} + [A_4(t)], \quad (5)$$

Calculation of motion with the help of a system of linear equations (see (Formula 5)) was performed for initial conditions of  $\{u\}_j = 0$  and  $\{u\}_{j-1} = 0$ , which corresponds to immobile spacecraft at the initial moment in time.

### 3.2. Modeling results

Two laws of controlling torque change were considered:

- 1) Two consequent "square" torque impulses in different directions. This corresponds to minimal values of the controlling torque at a given time and angle of turn;
- 2) Two consequent "smoothened" torque impulses in different directions. The shape of the smoothed impulses is taken as in the formula (6). Smooth change of controlling torque in this case is meant to decrease the amplitude of vibrations of the spacecraft's frame after the turn is completed, as compared to "square" torque impulses.

$$M_{cont}(t) = \begin{cases} 0,5 \cdot M_{max} \cdot \left(1 - \cos\left(4\pi \frac{t}{T_k}\right)\right), & \text{with } 0 \leq t < T_k/2, \\ -0,5 \cdot M_{max} \cdot \left(1 - \cos\left(4\pi \frac{t}{T_k}\right)\right), & \text{with } T_k/2 \leq t \leq T_k, \end{cases} \quad (6)$$

Table 3 gives the parameters of turns for these two torque control laws. The view of the obtained relation of rotation of the DOEs about the body of the spacecraft is represented in Figure 6.

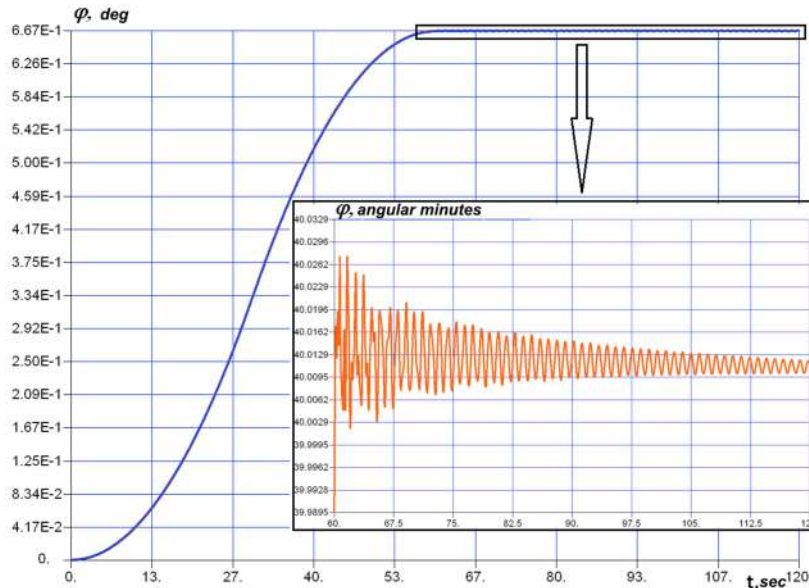


Fig. 6. The angle of the turn of the spacecraft for "square" controlling torque impulses.

Table 3. Parameters of the turn.

Controlling torque change law		
Time of turn [s]	60	60
Angle of turn [degrees]	0.67	0.67
Maximum controlling torque [N m]	13.1	26.2
Amplitude after turn [s of a]	2.34	0.021

As seen from Table 3, adoption of the "smoothened" torque control law reduced the amplitude of vibration by order of two. The maximum controlling torque value, however, increased twofold, which would require the spacecraft to be equipped with more powerful control thrusters. The results lead to the conclusion that the problem of reducing vibrations in the spacecraft's optical system elements can be efficiently resolved in this case by the choice of the type of controlling impact.

#### 4. Conclusion

The process of controlling a large-dimensional structure shown in Fig. 1 was modeled, using the specially developed multi-step terminal control algorithm that allows to account for perturbing accelerations. Adjustment of control parameters during the correction maneuver allows to reduce the final deviations of the orbital parameters (see Fig. 2). However, the considered multi-step algorithm has a disadvantage. To achieve the required eccentricity value at the end of the transfer to the given station point, the lengths of the passive parts of the transfer has to be hand-picked, which makes the search for the solution of the problem more complicated and not always successful.

Dynamic calculation and modeling of a turn of an observation spacecraft were carried out for two variants of torque control law: with "square" and "smoothened" change of the controlling torque.

In the first instance, which ensures the quickest turn, the amplitudes of vibration in the optical elements and the body of the spacecraft are  $\sim 0,039$  minutes of angle immediately after the turn, and  $\sim 2,9E-4$  minutes of angle 60 seconds after the end of the turn. In the second instance the amplitudes of vibration in the optical elements and the body of the spacecraft are  $\sim 3,5E-4$

minutes of angle immediately after the turn, and  $\sim 1,0E-4'$  minutes of angle 60 seconds after. The use of the "smoothened" torque control law increases the maximum value of controlling torque by two (to 26 Н·м for the turn under consideration), but decreases the vibrations produced during the maneuver by order of two.

### Acknowledgements

The research was carried out with financing within the framework of state order № 9.1004.2014/K.

### References

- [1] Atcheson P, Stewart C, Domber J, Whiteaker K, Cole J, Spuhler P, Seltzer A, Smith L. MOIRE - Initial demonstration of a transmissive diffractive membrane optic for large lightweight optical telescopes. Proceedings of SPIE - The International Society for Optical Engineering 2012; 8442: 844221.
- [2] Salmin VV, Karpeev SV, Peresykin KV, Chetverikov AS, Tkachenko IS. Feasibility study and modeling of components for an informational space system based on a large diffractive membrane. CEUR Workshop Proceedings 2016; 1638: 132–148.
- [3] Chernyavsky GM, Bartenev BA, Malyshev V A. Controlling the orbit of a geostationary satellite. Moscow: Mashinostroyeniye 1984; 144 p. (in Russian)
- [4] MSC.Nastran. Reference Manual: The Official Web Site of the Corporation, 2004. URL: <https://simcompanion.mscsoftware.com/resources/sites/MSC/content/meta/DOCUMENTATION/9000/DOC9188/~secure/refman.pdf>.

# On the method of step evaluation in construction descriptive models

T.E. Rodionova<sup>1</sup>, G.R. Kadyrova<sup>1</sup>

<sup>1</sup>Ulyanovsk State Technical University, Severniy Venets St. 32, 432027, Ulyanovsk, Russia

---

## Abstract

Mathematical regression models used to describe technical objects or process the considered. Assumptions violations of regression analysis appearing in different practical data processing, are discussed. The method of step evaluation allowing to overcome negative impact of multicollinearity effect is described. The models got by the analysis of laser and radio interferometric observations and data of physico-chemistry index of water source are cited. Received the ratings are compared with the results got when using the method of least squares and the step estimation method. The choice of the optimal model is made according to the criteria of minimum displacement. The possibility of applying the method of step evaluation to construct descriptive models is proved.

*Keywords:* descriptive models; multicollinearity; the method of least squares; regression analysis; methods of structural identification; step evaluation

---

## 1. Introduction

Let's consider a descriptive (parametrical) regression model applied for the description of relationships of cause and effect of the phenomenon. Let the mathematical model look like:

$$Y = \eta(X, \beta) + \varepsilon \quad (1)$$

Where  $Y$  is a dependent variable;  $\mathbf{X} = (x_0 x_1 \dots x_{p-1})^T$  - is a vector of independent variables;  $\beta = (\beta_0 \beta_1 \dots \beta_{p-1})^T$  is a vector of unknown parameters defined by the result of the experiment;  $\varepsilon$  is a vector of random errors. Variables  $X$  and  $Y$ , included into the model, are the result of a passive experiment, i.e. the measured or calculated values. Vector  $\beta$  in model (1) is supposed to non-changeable in time, i.e. a mathematical model is considered stationary according to parameters [1,13].

In practice, to estimate parameters of such mathematical models methods of regression analysis are used, in particular the method of least squares. Thus it's necessary to consider possible infringements of conditions of application of this method. The application of regression modelling in the task considered means research and selection of optimal methods to obtain the best linear estimates of parameters and check the effectiveness of the resulting model according to the relevant criteria [3,12].

We can identify the following violations of applying an ordinary method of least squares in solving practical problems. They are as follows:

- Models contain insignificant (noise) terms;
- The model parameters correlate with each other (multicollinearity effect);
- The residuals can also be further distorted by autocorrelation and other systematic errors.

In general, the choice of the way to adapt violations of conditions of regression analysis by the method of least squares depends on the type of model investigated [9, 14]. In this case, the object of attention is directly the parameters of the model, rather than the results of prediction. The ultimate goal of adaptation is the best linear estimates, e.g. evaluations not burdened by notable systematic and random errors. They can be such values, at least, in case of statistical significance and, what in most important, when parameters of the model are independent on each other. It's obvious that adaptation to the first violation mentioned by simply removing insignificant terms is difficult for a very simple reason: some of them can be interconnected with significant ones.

To overcome the effect of multicollinearity and reduce the number of insignificant terms in descriptive models it's proposed to use the step evaluation method.

## 2. The description the step evaluation method

According to the method the phased partitioning is carried out not by individual variables (like in step regression) but by their groups consistently formed as subsets of variables with insignificant pair correlation coefficients  $r_{ij}$ . That means that the groups are formed not by the degree of correlation with the sequentially formed by responses, but in the form of separate structures in almost orthogonal basis [2,4]. A brief description of the algorithm of the step estimation method:

1. The estimation valuation an original model using one of computing schemes of least squares method is calculated:

$$\Delta = (X^T X)^{-1} X^T, \quad (2)$$

Its covariance matrix is

$$D(\Delta) = (X^T X)^{-1} \sigma^2, \quad (3)$$

and different statistics, allowing to estimate the statistical value of each term and whole model, including values of t-statistics and coefficients of pair correlations  $r_{ij}$  are calculated.

2. The first subset of corrections  $\Delta_1$ , that got insignificant values  $r_{ij}$  is formed using comparison of values  $r_{ij}$ .
3. The parameters of the orthogonal structure are estimated



$$Y = X_1 \Delta_1, \quad (4)$$

the first vector of residues is calculated

$$e_1 = Y_1 - \hat{Y}_1, \quad (5)$$

which is regarded as another response vector forming the next subset of corrections from the set of remained ones.

4. Parts 2 and 3 are repeated until the process of forming subsets  $\Delta_1, \Delta_2, \dots, \Delta_k$  is finished.

To improve the quality of the estimates obtained, the orthogonal transformation of the Householder is included in the calculation scheme. In this case, numerical stability, characteristic for orthogonal transformations, is combined with flexibility, which makes it easy to adapt to the consistent accumulation of data, which is very important for solving large-dimensional problems. In addition, the requirement for computer memory is reduced, execution speed and accuracy are increased. The protection from "machine zeros" and overflow should be noted. With the help of the first strategy of this algorithm, an attempt was made to evaluate the interrelated regressors separately by evaluating them at different stages of this method (MSE1). As a drawback of such an algorithm, it can be noted that among the estimated parameters there are also insignificant statistics according to the Student's statistics. As a defect of such an algorithm, it can be noted that among the estimated parameters there are also insignificant statistics according to the Student's statistics.

The second strategy of this method including to the final model only those regressors that turned out to be significant according to the t-criterion at each stage of the work (MSE2). It's very similar to the step regression method, but due to the fact that the calculation is based on individual subsets, it's possible to estimate many times the parameters of the initial model – since the regressor insignificant at one stage may turn out to be significant in subsequent ones. This is very important for the problem of parametric estimation, where is necessary to obtain the most possible complete model. The defect of this strategy is the lack of analysis of the interdependence of the included parameters.

The third strategy is the combination of the first and second ones. The selection to the set of evaluated parameters is performed immediately according to two grounds; namely the significance and orthogonality (MSE3).

### 3. The description of the initial data and revealed violations of regression analysis assumptions

For approbation of this method of estimation of the parameters of a mathematical model the following data were used: the data on the laser location of the moon; VLBI observations of extragalactic sources; the results of physical and chemical control of drinking water.

The processed biennial radar data were got by using to angle reflector from the spaceship "Appollo-15" in McDonald observatory (Texas, USA) from August 1971 to November 1973 (549 observations at all). The source data in the form of coefficients of the conditional equations were prepared by the staff of the Institute of Theoretical Astronomy of USSR Academy of Sciences.

The considered VLBI observations are 1262 conditional equations for determining 203 corrections of the constant theory of the orbital motion and rotation of the Earth. To these data have been added 4 coupling equations, were added to these data. They determine the equality of the parallel transfer of the earth's coordinate system and the rotation of the earth's and celestial coordinate systems to zero, as well as the constraints imposed on the basis vectors. The data for the calculations were prepared by Professor V.E. Zharov (The State Astronomical Institute named after P.K. Shernberg, Moscow State University) [6,7].

As a third example, the results of physico-chemical control of drinking water (responses  $y_1 - y_7$ ) and water from water source (estimated parameters  $x_1 - x_8$ ), used to clean water were considered [8,10]. The original file is a result of the parameters control during a year.

As a first problem in the data processing we can name the problem of the sufficiency of the observations scope. In the research we are dealing with the following situation: laser data on the Moon to determine 24 unknown corrections contain 549 conditional equations, e.g. they exceed the number of estimated amendments 22 times; according to radiointerferometric data on the Earth, we have the ratio of 203 unknown corrections and 1289 conditional equations (including 27 coupling equations), so the number of observations is only 6 times as many as the number of parameters; according to the water source 365 observations are available to determine 8 water parameters (the number of observations is 45 times as many as the number of parameters). In the regression analysis between the number of determined parameters  $p$  and the number of observations  $n$  must be satisfied, during the experiment the ratio  $n = 5p \div 15p$ .

Data research began with the analysis of the model obtained by the multiple regression method. The number of insignificant parameters of the model and the matrix of pair correlation coefficients were considered. For this purpose, the SPOR package was used, which makes it possible to obtain regression models and determine their quality measures [5,11,15].

The presence of abnormal observations in the sample can be considered as the second problem facing a researcher. In the considered initial data four abnormal observations were removed from the laser observation file of the Moon, 18 outliers were found in the file with VLBI observations, and in the data for the water source for different responses, the amount of emissions varies from 1 to 4.

The next problem is directly related to the matrix of its original data: among the arguments (variables) should not be linearly dependent ones. However, in practice, this assumption is not always observed. When this condition is violated, the linear functional or statistical relationship exist between the analyzed variables. This phenomenon is called multicollinearity and has very negative consequences for estimation the regression coefficients. In computational mathematics, these concepts correspond to the degeneracy and poor conditionality of the matrix  $X^T X$ , i.e. for the latter there doesn't exist  $(X^T X)^{-1}$  and its determinant is close to zero. Consequences of this violation are particularly serious for models whose estimated parameters are subject to



physical interpretation. One of the ways to solve the multicollinearity problem may be that the equation must contain only terms that uncorrelate with each other.

In the analysis of radiointerferometric data, it was revealed that the matrix of correlation coefficients contained 76 coefficients exceeding modulo 0.5. 30 of these values of coefficients are greater than 0.95, which indicates an almost linear relationship between the estimated parameters. During the research of data of water source according to  $y_1$ - $y_7$ , from 1 to 3 correlating parameters of the model were revealed. It should also be mentioned that in the mathematical models under consideration the data on the factors and on the response have a different physical meaning and different physical dimensions. This causes computational inconvenience, because you have to work with both very large and very small numbers which can lead to computational errors.

Thus, the presence of insignificant terms in the obtained models, as well as the presence of a mutual correlation between the estimated parameters of the anomalous observations makes it possible to conclude that the assumptions of the regression analysis are violated.

#### 4. The step evaluation method used for adaptation to identified violations

To eliminate the effect of multicollinearity and the presence of insignificant parameters in the models, the step estimation method described above was applied. Further the results of application the step evaluation method for processing different data sets are given. The main task in creating descriptive models is to determine the maximum number of parameters with the highest accuracy. For laser data, the application of the step orthogonalization method (MSE1) allowed to estimate all parameters of the model. During selection only significant parameters of the step estimation method (MSE2), 10 corrections were obtained, and the MSE3 algorithm gave estimates for 9 corrections. The step regression method, which was used for comparison allowed us to estimate only 9 parameters out of 24 possible ones.

For radiointerferometric data: the application of the step regression method resulted in a model, containing estimates of 6 significant parameters out of 203 possible ones; 188 amendments were identified in the MSE1 strategy of the step-by-step assessment method, the MSE2 strategy gave an assessment of 136 amendments, and the MSE3 strategy gave 51 amendments.

Response	SR	MSE1	MSE2	MSE3
$y_1$	1, 2, 3, 5	3, 4, 5, 6, 7, 8	-	-
$y_2$	2, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	2, 4, 6, 7	2, 3, 6, 7
$y_3$	3, 5, 7	3, 4, 5, 6, 7, 8	1, 2, 4, 6, 7	1, 2, 3, 4, 6
$y_4$	2, 3, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 3, 4, 6, 7, 8	4, 6, 8
$y_5$	1, 2, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 7	1, 2, 3, 4, 5, 6, 7, 8
$y_6$	2, 3, 4, 5, 6, 7	3, 4, 5, 6, 7, 8	4, 6, 7	-
$y_7$	2, 5, 7	3, 4, 5, 6, 7, 8	2, 5, 6, 7	5, 6, 7

Table 1 lists the sets of parameters included in the model obtained by different computational schemes (SR is a step regression) as part of the treatment of water purification data. It can be seen from the table that the strategy of step evaluation method (The MSE1 is the selection of only orthogonal parameters) allows to estimate more model parameters than a step regression, which is very important in describing the technological process. For the considered data set, the MSE2 strategy (only significant ones selection at each step) and MSE3 (choosing significant and simultaneously orthogonal parameters at each step) did not allow to obtain models better than a step regression. The table shows the structure of the model and which of the eight regressors are significant and are the part of the model. The above data allows us to conclude that for various samples in the SR model different parameters were introduced, while neither of the models included either of the controlled parameters  $x_7$  and  $x_8$ . In the model obtained by the MSE1 strategy for all indicators of quality operation of the object  $y_1$ - $y_7$  the set of indicators is the same and practically in all cases  $x_7$  and  $x_8$  significant [16].

Comparing the estimates for the same parameters, obtained by various estimation methods, we can conclude that we have obtained values sufficiently close to each other. If we take the values obtained by the step regression method as the standard, then a very small number of estimates obtained by other methods is greatly different from the standard. Considering the ratio of the standard errors of the above estimates obtained by different estimation methods, we see that the accuracy of estimation of the unknown parameters in the considered methods of SR and MSE1 practically coincides. Thus, it can be concluded that the obtained models are applicable to the description of this technological process.

The next stage of the research is the task of choosing the best descriptive model. When solving it, it should be borne in mind that the internal criteria, i.e. criteria that don't use any additional information, in the presence of interference, can not solve the problem of choosing the best descriptive model. When using external measures, it's very important to split the initial sample into two parts. It's necessary to take into account the physical meaning and time of observation, since the initial data is a combination of several samples. It's proposed to choose a model by the criterion of minimum displacement-in consistency, which demands the model obtained from the training set, to be at least as possible different from the models obtained for the test sample. Analyzing the obtained results of processing radiointerferometric, laser observations and data on water purification, we can conclude that the methods of step estimation are effective.

## 5. Conclusion

The numerical experiments carried out make it possible to make the following conclusions:

- the step evaluation method allows to evaluate a larger number of model parameters;
- estimations of the step evaluation method are close to the estimations of step regression. Thus, the step evaluation method can be used for evaluating the parameters of a mathematical model, as well as for describing technical objects and technological processes. Analyzing the obtained values of the minimum displacement criteria, for the indicated observations (both radiointerferometric and laser observations and data on water purification), it can be concluded that the step evaluation methods are effective and allow us to describe the object under investigation with sufficient accuracy.

## References

- [1] Valeev SG, Kadyrova GR. Optimal regression search system: tutorial. Kazan: FEN, 2003; 160 p.
- [2] Valeev SG, Rodionova TE. The method of stepwise orthogonalization of the and its using during least-squares taskio Izvestiya Vuzov. Geodezy and Aerophotography 2003; 6: 3–14.
- [3] Valeev SG, Rodionova TE. Analysus of methods for parameters rating at multicollinear values. Izvestiya Vuzov. Geodezy and Aerophotography 1999; 5: 20–28.
- [4] Valeev SG, Rodionova TE. Software for solving task of structure-parametrical ranking during data processing. Izvestiya Vuzov. Geodezy and Aerophotography 2004; 1: 25–34.
- [5] Valeev SG, Kadyrova GR. Automatic system for solving least-squares method tasks. Izvestiya Vuzov. Geodezy and Aerophotography 1999; 6: 124–130.
- [6] Valeev SG, Rodionova TE, Zharov VE. Methodic of statistical processing of RSDB-observings. Izvestiya Vuzov. Geodezy and Aerophotography 2008; 1: 13–18.
- [7] Valeev SG, Rodionova TE, Zharov VE. Computational experiments for processing of RSDB-observings. Izvestiya Vuzov. Geodezy and Aerophotography 2008; 2: 94–100.
- [8] Rodionova TE. Using adaptive-regression modelling for describing the functioning of technical object. Izvestiya of the Samara Russian Academy of Sciences scientific center 2014; 16(6-2): 572–575.
- [9] Kadyrova GR. Estimation and prediction of the state of a technical object based on regression models of regressions. Automation of management processes 2015; 4(42): 90–95.
- [10] Rodionova TE, Klyachkin VN. Statistical methods of estimation the drinking water quality. Reports of the Academy of Sciences of the Russian Federation 2014; 2-3: 101–110.
- [11] Valeev SG, Kadyrova GR, Turchenco AA. Software system for optimal regression searching. Issues of modern science and practice. Technical science 2008; 4(14): 97–101.
- [12] Kadyrova GR. Modification of the stepwise regression method for obtaining mathematical models for predicting the behavior of an object. Automation of management processes 2016; 3(45): 65–70.
- [13] Valeev SG, Rodionova TE. Sequential orthogonalization of a basis in problems of the least squares method. Messenger of the Ulyanovsk state technical university 1999; 1(6): 4–9.
- [14] Kadyrova GR. Software System of searching for optimal regression models of forecast . Way of science 2014; 7 (7): 10–11.
- [15] Kadyrova GR. The system of searching for the optimal model. State of affairs and development prospects. Modern science potential 2015; 4(12): 8–10.
- [16] Rodionova TE. Comparison of regression indicator models of the drinking water quality. Materials of 3-rd science-practical internet-conference “Interdisciplinary research in the field of mathematical modeling and informatics”. Tolyatti, 2014; 159–162.

# Server hardware resources optimization for virtual desktop infrastructure implementation

K. Makoviy<sup>1</sup>, D. Proskurin<sup>1</sup>, Yu. Khitskova<sup>1</sup>, Ya. Metelkin<sup>1</sup>

<sup>1</sup>*Voronezh State Technical University, 20 let Oktyabrya str., 84, 394006, Voronezh, Russia*

---

## Abstract

A new model of capacity planning problem applied to Virtual Desktop Infrastructure implementation is proposed. The possibility of applying the methods of integer mathematical programming to the problem of optimizing server set providing the predetermined number of virtual machines operating.

*Keywords:* capacity planning; virtual desktop infrastructure; integer programming; server hardware; equipment costs

---

## 1. Introduction

Virtualization is a common concept for concealing the real structure that is used for creating virtual hardware and operating system, virtual storage and network resources. Most organizations of different sizes and income have implemented server virtualization over the past 10 years. Server virtualizations is based on the hypervisor technology, which creates a thing interlayer between hardware and guest operation system.

On the next step of developing IT infrastructure, organizations address to the technology of centralized desktop execution enhancing end user experience and IT management of desktops. While implementing desktop virtualization it is essential to understand that this solution requires not only adequate planning but also significant financial costs. Value of hardware for physical servers makes considerable contribution for the investment costs [1] whereas optimal configuration of the servers purchased can save considerable funds.

We offer a mathematical model for solving the optimization problem of server resources needed for desktop virtualization implementation and present computing results.

## 2. The object of the study

Server virtualization is essentially a server consolidation, i.e. an approach to the efficient usage of physical server resources. This technology allows several operation systems to run on one physical server and isolate applications from each other's influence, minimize investment and operational costs, avoid overprovisioning.

Desktop virtualization or Virtual Desktop Infrastructure (VDI) uses advantages of server virtualization and cloud technologies bringing together the benefits gained from hypervisor-enabled virtualization and modern display network protocols. Desktop operating systems run on a physical server under control of host operating system i.e. 'hypervisor' whilst screen image is delivered by a network protocol to a client device which may be a Personal Computer (PC), Thin Client, laptop, tablet, etc. There are several commercial software products for implementing Virtual Desktop Infrastructure. The most popular ones are VMWare Horizon View, Microsoft VDI, XenDesktop from Citrix. There is also freeware product on the base of Linux KVM.

One of the key perspectives of VDI implementation is a possibility to execute any application on any device for which there is a VDI client since applications are executed on the operating system running on the server, not on the device itself. Thus, desktop virtualization provides the basis for extremely promising technology allowing creation a common learning environment - BYOD (Bring Your Own Device) – a new initiative giving opportunity to use wide variety of client's personal devices in a corporate environment.

The number of client computers in a typical organization far exceeds the number of servers therefore this is so important to be able to assess server resources required to run client virtual machines. No less important is to be able to choose the optimal set of hardware servers, for example, from the range of particular vendor. The key moment to minimize expenses of hardware procurement is a clear view of hardware server set needed to provide execution of required number of virtual desktops. We consider VDI implementation in a high school institute, namely the Voronezh State Technical University, which has already a centralized server infrastructure and well-designed network. Desktop infrastructure in an educational institution contains as a rule several sets of identical computers that placed in computer labs. Definitely apart from desktops in computer labs there are a large number of computers with diverse software, which are used by staff. This type of computers is not the best choice to being virtualized at the first stage of the project. In our model, we assume a number of identical desktops that should be placed optimally on hardware servers. Identity we understand like equivalence in their performance requirements specifically memory needs for running desktop applications.

### 3. Methods

The problem of virtualized server optimization was considered previously in two aspects – static and dynamic. Static Server Allocation Problem is an approach based on a service concept, the model was introduced in [2] and designed to optimally allocate source servers to physically target servers and was proven that this model is NP-hard problem, heuristic solution based on bin packet problem is offered. Another option of using linear programming methods for virtualized system placement representing the dynamic aspect of the problem is used for creating application placement controller pMapper [3].

There are several attempts to solve the problem of dynamic replacement of virtual machines on existed physical server infrastructure in datacenter to optimize energy consumption, minimize administrative efforts, increasing server utilization. An approach of dynamic resource allocation for large Internet-oriented data centers bases on queuing theory and Erlang's loss formula represented in [4]. On the other hand it is proposed to use a genetic algorithm based approach, namely GABA, to adaptively self-reconfigure the VMs (Virtual Machines) in large-scale data centers [5]. All the models proposed focuses on the server virtualization not the desktop virtualization. As for desktop virtualization an allocation algorithm based on a bin-packet problem is developed [6]. It is mainly focused on achieving a balance between resource usage optimization and user satisfaction.

In this work we concentrated on the problem of server hardware assessment optimization in order to reduce financial costs while implementing desktop virtualization at the university. To achieve this goal we have to analyze resource requirements of VMs that will be used, number of VMs, and range of hardware servers of the vendor then solve optimization problem to choose a set of optimal server models and their configuration to minimize total cost.

#### 3.1. Model description

For the model we assume a particular number of the same virtual desktops. We plan to use them for computer labs at the university and actually we probably will have a need of several types of virtual machines for different labs but for the first approximation, we will consider all virtual machines have exactly the same resources requirements.

We consider discrete set of server platform models, each of them may be supplemented by additional RAM (Random Access Memory) modules. We can extend RAM with additional memory modules that have various amounts and prices. We assume also that performance of the server is acceptable if RAM amount is sufficient for running VMs only in virtual memory not using as a rule a paging file. In this approximation, we do not consider the processor load since the main purpose of this model is minimizing total costs at the very start of VDI implementation project.

For the model description, we introduce the following variables:

$\bar{S} = \{S_1, S_2 \dots S_m\}$  – vector of server platform models that can be used for the hardware servers, where  $m$  – total number of server platform models selected for consideration;

$\bar{C} = \{C_1, C_2 \dots C_m\}$  – vector of values of server platforms  $\bar{S}$ , where  $C_i$  - is a value of  $S_i$ ,  $S_i \in \bar{S}$ ,  $i = 1..m$ ;

$\bar{N} = \{N_1, N_2 \dots N_m\}$  – numbers of servers of server platform model  $S_i$  that will be used in a final set;

$\bar{P} = \{P_1, P_2 \dots P_m\}$  – vector of memory slots in the server  $S_i$ , this is a maximum number of memory modules that can be used for the server  $S_i$ ;

$\bar{M} = \{M_1, M_2 \dots M_m\}$  – vector of maximum RAM amounts that can be added to the server platform  $S_i$ ;

$\bar{R} = \{R_1, R_2 \dots R_k\}$  – amount of memory module  $j$ ,  $j = 1..k$ , where  $k$  – is the number of types of RAM modules;

$\bar{Cv} = \{Cv_1, Cv_2 \dots Cv_k\}$  – value of memory module  $j$ ,  $j = 1..k$ .

Because our goal is to minimize costs then we determine an objective function reflecting the total cost of the hardware server set. The total cost of the server consists of the value of based server platform model ( $C_i$ ) and the cost of additional RAM

modules ( $\sum_{j=1}^k Cv_j n_{ji}$ ), where  $n_{ji}$  - number of RAM modules  $j$  on the server  $S_i$ . Thus, the objective function is the following:

$$F = \sum_{i=1}^n (C_i + \sum_{j=1}^k Cv_j n_{ji}) N_i \quad (1)$$

In the following we present constrains for the objective function:

1. The total amount of RAM should not exceed the one supported by this server platform model:

$$\sum_{j=1}^k R_j n_{ji} \leq M_i \quad (2)$$

where  $n_{ji}$  - number of RAM modules  $j$  on the server  $S_i$ ,  $j = 1..k$ ,  $i = 1..m$ .

2. The total number of RAM modules cannot exceed the number of hardware server model memory slots:

$$\sum_{j=1}^k n_{ji} \leq P_i, \quad (3)$$

3. The total amount of RAM memory on all servers out of server set should provide enough memory to run necessary number of VMs:

$$\sum_{i=1}^n ([\sum_{j=1}^k R_j n_{ji}] / V) \geq N_V, \quad (4)$$

where  $N_V$  – is a number of VMs,  $V$  – memory needed for one virtual machine.

4. To get a solution that makes a sense we will add a constrains for numbers of servers and RAM modules to be integer:

$$N_i, n_{ji} \geq 0, i = 1..m, j = 1..k, N_i, n_{ji} - \text{integer} \quad (5)$$

The model proposed makes it possible to solve the problem of selecting the optimal set of server hardware equipment necessary for Virtual Desktop Infrastructure deployment. This model can be refined to allow sets of virtual machines that differs by hardware resources requirements and expand the range of considered hardware resources types.

### 3.2. Model solution.

In order to obtain a solution we divided this problem into two parts:

1. On the first step of calculation, we create optimal filling of the server slots by RAM modules, analyzing filling for 25%, 50%, 75% and 100% of the maximum amount. Objective function  $C_{v_i}^p$  reflects the cost of RAM added to the  $S_i$  server platform model filled with RAM modules by part equal to  $p$  of the maximum and is the following:

$$C_{v_i}^p = \min \sum_{j=1}^k C_{v_j} n_{ji} \quad (6)$$

subject to:

$$\begin{cases} \sum_{j=1}^k R_j n_{ji} = M_i p \\ \sum_{j=1}^k n_{ji} \leq P_i \end{cases}, \quad (7)$$

where  $p$  – part of the maximum RAM amount, which can be either 0,25, 0,5, 0,75 and 1. This is linear programming problem, which was solved by branch and boundary method [7]. For each filling percentage, we get the optimal set of memory modules for every server platform model. Thus, we get four hardware servers for selection instead of one server platform model.

2. On the second step we form a final set of servers minimizing the following objective function:

$$\min \sum_{i=1}^m \sum_{j=1}^4 (C_i + C_{v_i}^{p_j}) N_{ij}, \quad (8)$$

where  $C_{v_i}^{p_j}$  is a result of (6), i.e. cost of additional memory of server  $S_i$ , filled by memory modules on  $p_j$  part, subject to (4) and  $N_{ij} \geq 0, i = 1..m, j = 1..4, N_{ij} - \text{integer}$ .

## 4. Results and Discussion

In the following, we provide the numerical results of applying this model to a set of servers for deploying different numbers of virtual machines. Any set of hardware server platforms by one vendor can be used as an initial set of server platform models. The servers used in calculation presented in the table 1. These are 11 models of HP ProLiant Servers of ML product line. One of the main reasons to use these servers was the fact that HP ProLiant Servers are used in Voronezh State Technical University IT-infrastructure. The cost and configuration of server platform models were taken from the site of one of the server distributors [8]. It happens that two models differ only in the processor that is why CPU model is also presented in a table 2.

For the problem solution, we used MatLab realization of the brunch and bound method [7]. As for amount of memory needed for one virtual machine we assume it is 4Gb. This the amount recommended by vendors of VDI software is used as a first approach. Further investigations of the memory amount necessary for one virtual machine should base on performance counters analysis in a pilot project. Some software products can help to estimate the required amount of memory, for example, VMWare View Planner.

There are five types of RAM modules available for HP ProLiant Servers: 2Gb, 4Gb, 8Gb, 16Gb, 32Gb value 26, 136, 215, 315, 840 USD respectively. For each server platform model the problem of optimal memory filling up to 25, 50, 75 and 100% of

maximum amount is resolved. It was not always possible to get 100% of maximum possible memory capacity because of the pre-installed small RAM modules. In this case, the maximum possible amount of memory was considered. The result of optimal filling the server slots by RAM modules to minimize cost while maximizing the amount of memory is in Table 2.

Table 1. Initial set of hardware server models.

№	Name	Initial RAM (Gb)	Number of RAM modules (pcs.)	Max amount of RAM (Gb)	RAM slots (pcs.)	CPU	Cost (USD)
1	ML150 Gen9 NHP		4	1	512	16	E5-2603v3 - 1.60 1580
2	ML150 Gen9 Hot Plug		8	1	512	16	E5-2609v3 - 1.90 1700
3	ML150 Gen9 NHP		8	1	512	16	E5-2609v3 - 1.90 1960
4	ML350p Gen8		8	2	384	24	E5-2620 - 2.00 3300
5	ML350p Gen8		8	2	384	24	E5-2630 - 2.30 4300
6	ML350p Gen8		32	4	384	24	E5-2620 - 2.00 4440
7	ML350e Gen8 Hot plug		8	2	192	12	E5-2420 - 1.90 1874
8	ML350p Gen8 E5-2620 Hot Plug		16	2	384	24	E5-2620 - 2.00 3556
9	ML350p Gen8 E5-2620		8	2	384	24	E5-2620 - 2.00 3169
10	ML350e Gen8 Hot plug		2	1	96	12	E5-2407 - 2.20 1624
11	ML350p Gen8 HPM		16	2	384	24	E5-2640v2 - 2.00 7100

Table 2. Optimal filling of server slots for several server platform models.

Percentage of maximum RAM supported by hardware server (%)	Number of RAM modules 2Gb	Number of RAM modules 4Gb	Number of RAM modules 8Gb	Number of RAM modules 16Gb	Number of RAM modules 32Gb	Cost (USD)
<b>ML150 Gen9 NHP</b>						
25	6	0	0	7	0	2361
50	0	1	1	11	2	3816
75	0	0	1	3	10	9560
100	0	0	0	0	15	12600
<b>ML150 Gen9 Hot Plug</b>						
25	4	0	0	7	0	2309
50	0	0	1	13	1	5150
75	0	0	1	5	9	9350
100	0	0	0	0	15	12600
<b>ML350e Gen8 Hot plug</b>						
25	4	0	0	2	0	734
50	4	0	0	5	0	1679
75	0	0	1	8	0	2735
100	0	0	1	7	2	4100
<b>ML350e Gen8 Hot plug</b>						
25	11	0	0	0	0	286
50	7	0	0	2	0	812
75	3	0	0	4	0	1338
100	5	1	0	5	0	1841

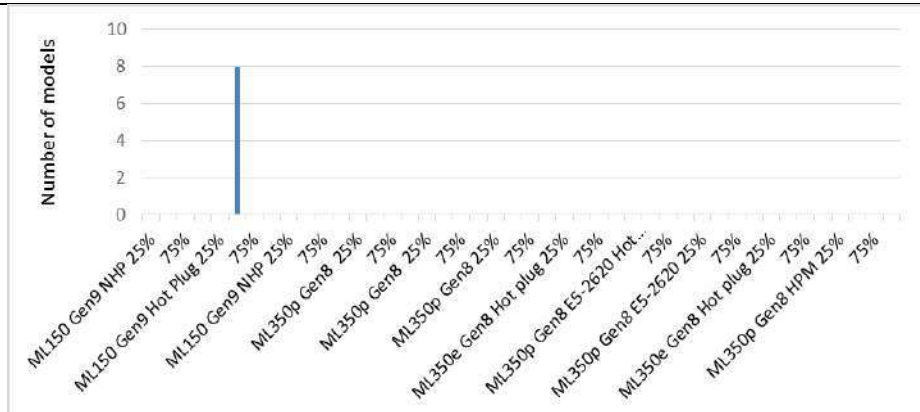


Fig.1. Problem resolution for 500 virtual machines.

The obtained results are used in the second part of solution that implement selection of optimal final set of servers from the variety of servers determined on the first part of solution. The result of model solution for 500, 700, 900 and 1000 VMs is presented on the figures 1-4.

According to figure 1 for placing 500 virtual machines, it's optimal to use eight servers ML150 G9 Hot Plug with 50% of memory filling. Table 2 shows that for this filling it is necessary to add to the base model one RAM module of amount of 8Gb, 13 modules of 16Gb and one of 32Gb.

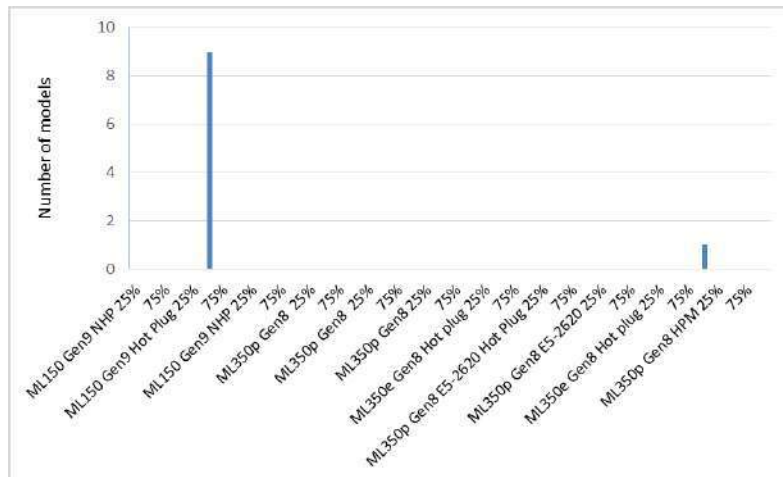


Fig.2. Problem resolution for 600 virtual machines.

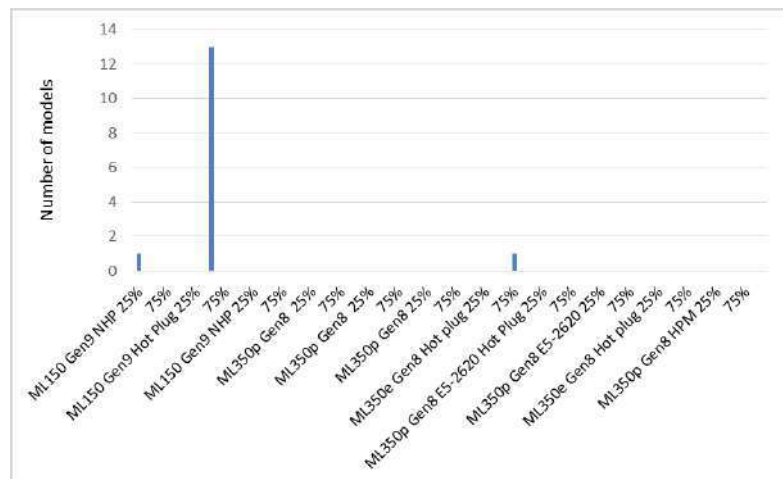


Fig. 3. Problem resolution for 900 virtual machines.

For 600 virtual machines it's optimal to use nine servers ML150 G9 Hot Plug with the following set of RAM modules: 1 x 8Gb, 13 x 16Gb and 1 x 32Gb and one server ML350e G8 Hot Plug with RAM: 5 x 2Gb, 1 x 4Gb and 5 x 16Gb.

For 900 virtual machines the following final is obtained (figure 3, table 2):

- 1 x ML150 G9 NHP server model with RAM modules:
  - 6 x 2Gb, 7 x 16Gb;
- 13 x ML150 G9 Hot Plug server with RAM:
  - 1 x 8Gb, 13 x 16Gb, 1 x 32Gb;
- 1 x ML350e G8 Hot Plug server model with RAM modules listed below:
  - 1 x 8Gb, 8 x 16Gb.

For 1000 virtual machines we receive the following result from calculation results presented on figure 4 and table 2:

- 1 x ML150 G9 NHP server with RAM:
  - 6 x 2Gb, 7 x 16Gb
- 14 x ML150 G9 Hot Plug server with RAM:
  - 1 x 8Gb, 13 x 16Gb, 1 x 32Gb
- 1 x ML350p G8 Hot Plug server
  - 4 x 2Gb, 17 x 16 Gb.

Analyzing the results received we noticed that server platform ML150 G9 Hot Plug occurs more frequently than the others. This can be interpreted as the server platform model with optimal price quality ratio. This means that an analysis of large number of server platform models can reveal the ones preferred for procurement of equipment in the enterprise.

Application of this model to hardware server procurement can minimize costs during implementation of Virtual Desktop Infrastructure. This model may be extended by considering several types of VMs requiring different amount of memory, taking into consideration processor power and Fault Tolerance demands.

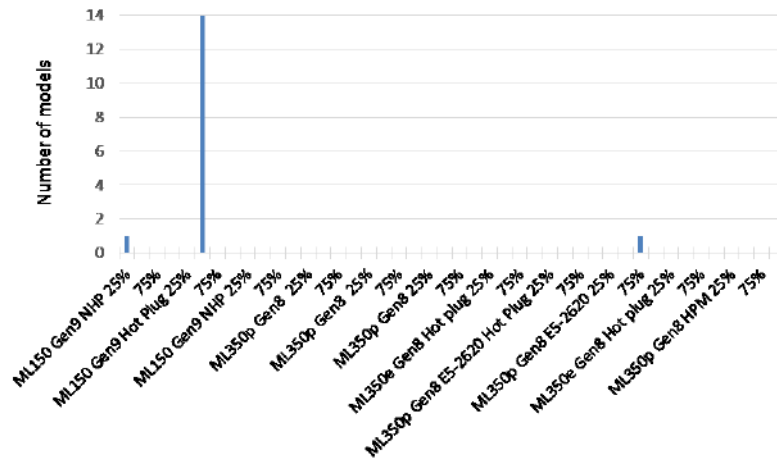


Fig. 3. Problem resolution for 900 virtual machines.

## 5. Conclusion

The interest to VDI technology grows fast because of popularity of cloud computing. Desktop virtualization implementation is a next step in centralizing IT infrastructure that brings both management advantages and academic benefits creating a convenient integrated educational environment. The new model for the optimizing acquisition costs of server hardware purchased for VDI implementation is offered. The results for numerical calculation shows also server platforms models with best price-quality ratio.

## References

- [1] Makoviy KA, Khitskova YuV. Economic basis of VDI deployment in institution of higher education IT-infrastructure. *Modern economy: problem and solutions* 2015; 2(62): 75–81.
- [2] Speitkamp B, Bichler M. A mathematical programming approach for server consolidation problems in virtualized data centers. *IEEE Trans. Services Comput.* 2010; 3(X): 266–278.
- [3] Verma A, Ahuja P, Neogi A. Mapper: power and migration cost aware application placement in virtualized systems. *Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware* 2008; 243–264.
- [4] Song Y, Zhang Y, Sun Y, Shi W. Utility Analysis for Internet-Oriented Server Consolidation in VM-Based Data Centers. *Proceedings of the IEEE International Conference on Services Computing* 2009; 1–10.
- [5] Mi H, Wang H, Yin G, Zhou Y, Shi D, Yuan L. Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers. *Proceedings of the IEEE International Conference on Services Computing* 2010; 514–521.
- [6] Armstrong D, Espling D, Tordsson J, Djemame K, Elmroth E. Contextualization: dynamic configuration of virtual machines. *Journal of Cloud Computing: Advances, Systems and Applications* 2015; 4–17.
- [7] Taha HA. *Operation research: An introduction*. Moscow: Publishing House Willams, 2001; 912 p.
- [8] Servers and Accessories of Hewlett-Packard. URL: [http://www.proliant.ru/files/File/HP\\_proliant\\_price\\_09\\_15.xls](http://www.proliant.ru/files/File/HP_proliant_price_09_15.xls).



# On delayed loss of stability in one mechanical problem

E.V. Shchetinina<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

Consider a double pendulum under the external force. This model can be described by the multi-speed systems of ordinary differential equations. We study the existence of stable motions depending on the parameters of the system and on the value of the external force. Critical values of the parameters are defined.

*Ключевые слова:* slow-fast systems; delayed loss of stability; double pendulum

---

## 1. Introduction

Numerical investigation of dynamical systems can be very complicated due to the high dimensions of systems and presence of big and small parameters. In such cases it is reasonable to use combination of numerical and analytical methods. By analytical methods we can decrease the dimension of the model, restrict the domain of the investigation, define the different types of behavior depending on the parameters values. And then by using numerical methods we are able to investigate specific solutions and manifolds of them.

One such analytical method is the method of the integral manifolds. With the help of it we are able to decrease the number of variables, to restrict the domains of the investigation, to predict different types of behaviour of the solutions to the system.

In this paper we consider a double pendulum with elastic hinges under the external force. We assume that the force is slowly changing in time. Our goal is to investigate the behavior of the solutions for different values of the parameters of the system and to find the influence of the force growth to the pendulum motions.

## 2. Integral manifolds

Integral manifolds method is an efficient tool for studying complicated dynamical systems. It was developed by many authors. This method is quiet successful for studying multi-scaled systems (see, e.g. [3]). The main ideas if this method are as follows.

Consider the singularly perturbed system:

$$\frac{dx}{dt} = f(x, y, a, \varepsilon), \varepsilon \frac{dy}{dt} = g(x, y, a, \varepsilon), \quad (1)$$

with  $x \in R^n, y \in R^m, t \in R, a$  is a parameter,  $0 < \varepsilon \ll 1$ . Here  $x$  is a slow variable,  $y$  is a fast variable. Integral manifold of such a system is an invariant set of the system. We are interested in the integral manifolds of the form  $y = h(x, \varepsilon)$ , with  $h$  smoothly depending on  $\varepsilon$ . This type of manifolds is called slow integral manifolds. The motion on this manifolds is provided by the equation  $\dot{x} = f(x, h(x, \varepsilon), \varepsilon)$ .

Suppose that the degenerated equation  $g(x, y, a, 0) = 0$  has an isolated root  $y = h_0(x, a)$ . The surface defined by the relation  $y = h_0(x, a)$  is called slow manifold. Consider the Jacobi matrix  $\partial g / \partial y(x, y, a)$  with  $y = h_0(x, a)$ . If all eigenvalues of the Jacobi matrix are in the left open complex half-plane, then the slow manifold is attractive. If there exists at least one eigenvalue in the right complex half-plane, then the slow manifold is repelling. In the  $\varepsilon$ -neighbourhoods of the attracting and repelling slow manifolds there exist attracting and repelling slow integral manifolds.

The surface with the condition that the eigenvalues of the Jacobi matrix are on the imaginary axis is called the surface of change of attractivity. Due to the presence of the additional parameter  $a$  we are able to glue together attracting and repelling slow integral manifolds at one point. Thus we obtain the solution that follows first an attractive slow manifold and then repelling slow manifold. Such a solution is called a canard solution.

Also there exists another type of change of attractivity. Suppose that in the spectrum of the Jacobi matrix there exists a pair of complex conjugated eigenvalues which cross an imaginary axes from the left to the right with nonvanishing speed. In this case the trajectories of the system starting in the small neighbourhood of the attracting part of the slow manifold follow it until the point of change of attractivity. But after crossing this point the trajectories do not leave this small neighbourhood immediately. The loss of attractivity is delayed: for some time the trajectories stay near repelling part of the slow manifold and then jump away. This phenomenon was described in [2]. Let us mention that the presence of the additional parameter allow us to change the time of the attractivity loss delay [4].

## Ziegler pendulum

We consider the Ziegler system which describes a double pendulum with elastic hinges under the external force  $P$  (fig. 1) [5]:

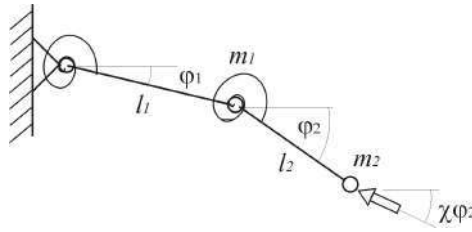


Fig. 1. Ziegler pendulum.

The behaviour of such a system can be described by the Lagrange equations. In the case that the spring rigidity is high, then the system has a big parameter. Suppose that the strength of the force  $P$  is slowly changing according to the law  $p = p_0 + \varepsilon\tau$ , where  $0 < \varepsilon \ll 1$ . It means that the system becomes multispeed. By using the integral manifold theory we obtain the following system of ordinary differential equations

$$\begin{aligned} \dot{p} &= \varepsilon, \\ \dot{\varphi}_1 &= -\varepsilon \left( \varphi_1 + \kappa \left( (1+\mu)\cos(\varphi_1) + \cos(\varphi_2) \right) + p \left( \sin(\chi\varphi_2 - \varphi_1) - \sin((\chi-1)\varphi_2) \right) \right), \\ \dot{\varphi}_2 &= -\varepsilon \left( \varphi_2 + \kappa \left( 2\cos(\varphi_2) + (1+\mu)\cos(\varphi_1) \right) + p \left( 2\sin((\chi-1)\varphi_2) - \sin(\chi\varphi_2 - \varphi_1) \right) \right). \end{aligned}$$

Here  $\kappa$  is a small parameter inversely related to the spring rigidity,  $\mu = m_1/m_2$ . The system possesses the stationary state  $\varphi_1 = \varphi_2 = 0$ . Linearization of the fast subsystem near this manifold leads to the form

$$\begin{aligned} \dot{p} &= \varepsilon, \\ \dot{\varphi}_1 &= -\varepsilon \left( \varphi_1(p-1) + \varphi_2 p(1-2\chi) \right), \\ \dot{\varphi}_2 &= -\varepsilon \left( p\varphi_1 - \varphi_2(1+p(3\chi-2)) \right). \end{aligned}$$

The characteristic equations is

$$(\lambda + 1 - 3/2 p(1 - \chi))^2 + p^2/4 (1 - \chi)(9\chi - 5) = 0.$$

From this, it follows that for  $0 < \chi < 5/9$  the last equation has two real eigenvalues, and for  $5/9 < \chi < 1$  it has a pair of complex conjugated eigenvalues. In the case of real eigenvalues there exists the point  $p_{cr}$  defined by

$$p_{cr} = 3/2 - 1/2 \sqrt{(5 - 9\chi)/(1 - \chi)}$$

Such that for  $p < p_{cr}$  all eigenvalues are negative, for  $p = p_{cr}$  one eigenvalue is zero and another one is negative, for  $p > p_{cr}$  one eigenvalue becomes positive. Therefore the slow manifold  $\varphi_1 = \varphi_2 = 0$  is attracting for  $p < p_{cr}$ , and repelling for  $p > p_{cr}$ . Therefore the solution  $\varphi_1 = \varphi_2 = 0$  is a canard solution. All other solutions, starting for  $p < p_{cr}$  approach the small neighbourhood of the attractive part of the slow manifold and follow it until the point  $p = p_{cr}$ . After that they follow for some time the repelling part of the slow manifold and then jump away.

In case of complex eigenvalues the value of  $p_{cr}$  is defined by the equation

$$2 - 3p(1 - \chi) = 0.$$

Then the eigenvalues are in the left half plane for all  $p < p_{cr}$ , for  $p = p_{cr}$  the system has a pair of pure imaginary eigenvalues, and for  $p > p_{cr}$  the real part of them becomes positive. It means that the solutions of the system possess delay of loss of attractivity. The trajectories starting at the point with  $p < p_{cr}$  approach the attractive part of the slow manifold and follow it until the point  $p = p_{cr}$ . After that for some time they follow the repelling part of the slow manifold and then jump away.

We note that in both cases as farther the trajectory starts from the change stability point as longer it will follow the repelling part of the slow manifold.

Next pictures show the different solutions to the system for different starting points  $p$  and different values of the parameter  $\chi$ .

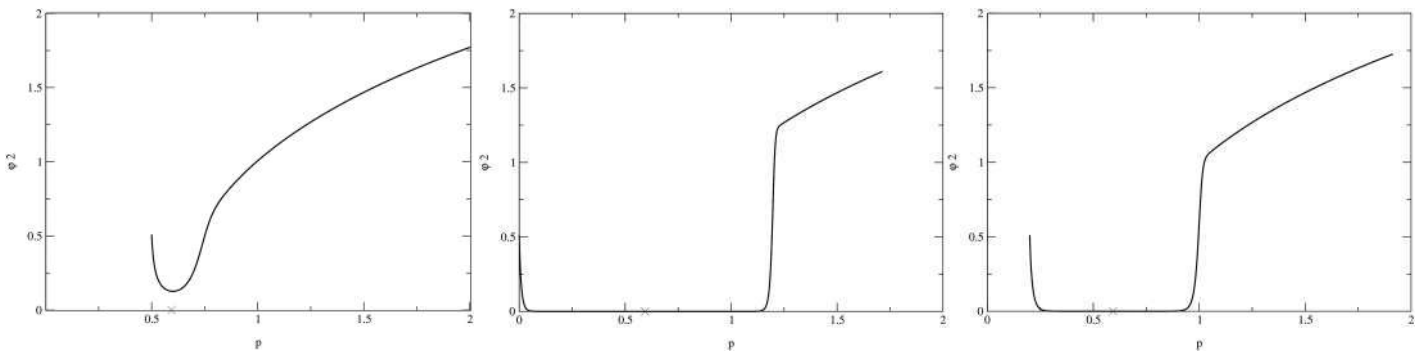


Fig. 2. Canard solutions for different initial value of  $p$  in the case  $\chi < 5/9$ .

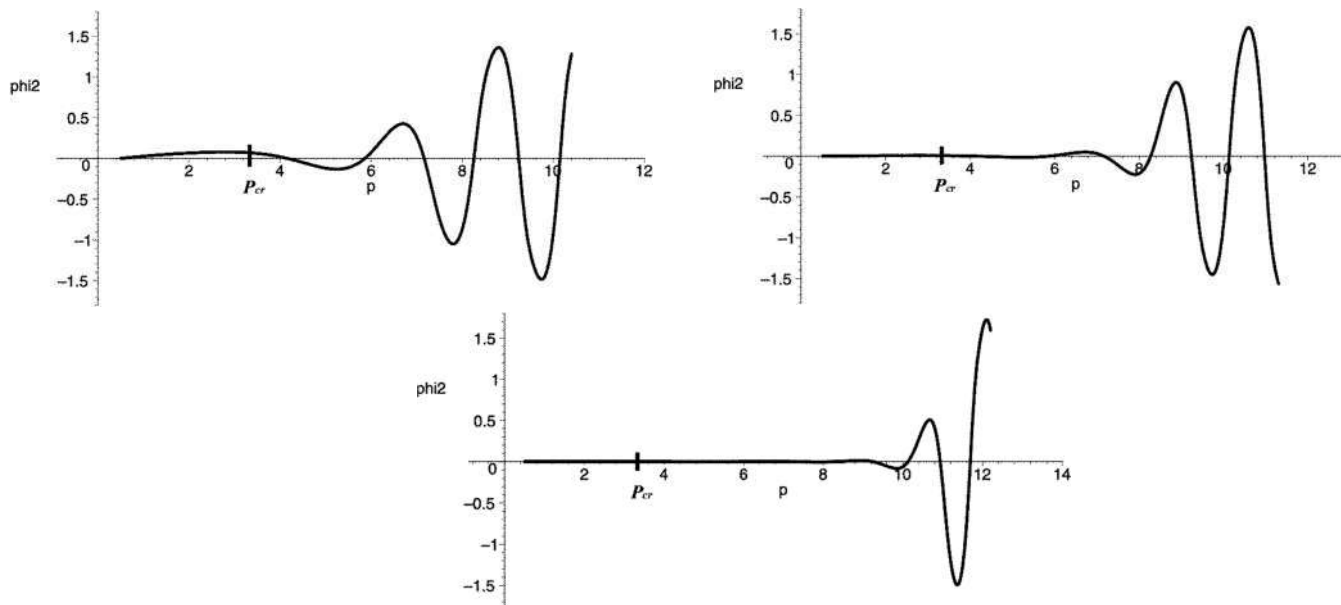


Fig. 3. Solutions with delayed loss of attractivity for the case  $\chi > 5/9$ .

### References

- [1] Gorelov GN, Sobolev VA, Shchepakina EA. Singularly perturbed combustion models. Samara, 1999; 198 p. (in Russian)
- [2] Neishtadt AI. Persistence of stability loss for dynamical bifurcations, I. Differents. Uravn, 1987; 23(12): 2060–2067. II. Differents. Uravn, 1988; 24(2): 226–233. (in Russian)
- [3] Strygin VV, Sobolev VA. Separation of motions by the integral manifolds. M.: Nauka, 1988; 256 p. (in Russian)
- [4] Shchetinina EV. Integral manifolds for slow-fast systems and delayed loss of stability. Vestnik SGU 2010; 6(80): 93–105. (in Russian)
- [5] Ziegler H. Die Stabilitaetskriterien der Elastomechanik. Ingenieur-Archiv, 1952; Band XX: 49–56.

# Characteristics comparison of DTN networks routing protocols using hybrid model of nodes' mobility

A.A. Tsarev<sup>1</sup>, A.Yu. Privalov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The results of simulation of popular routing protocols in DTN wireless networks with the hybrid mobility model of DTN's nodes are presented. The purpose was to evaluate the message delivery probability and the average time of message delivery. The simulation model is implemented in the OMNeT++ simulation system. From the simulation experiments it has been found that the daily periodic repeatability of node's movements has a sufficient influence on the performance of routing protocols with different principles of route determination.

*Keywords:* delay tolerant network; routing protocols; human's mobility model; simulation modeling; OMNeT++

---

## 1. Introduction

Due to the great complexity of modeling of mobile wireless networks in general, and so-called delay-tolerated networks (DTNs) in particular, computer simulation plays a leading role in the study of such networks, including the characteristics of routing protocols. It is obvious that the mobility model used in simulation of such networks has a very strong effect on the considered protocol characteristics. Therefore the mobility model should reflect the features of network nodes real mobility as closely as possible.

As a results of human's mobility researches, which attracted much attention of the scientific community in the last decade, a number of important features were revealed. These features are: the clustering of waypoints in real mobility traces, the Levy distribution of the distances between waypoints, and the so-called persistence (i.e. approximate constancy) of the daily routes of one user, if the system is considered for several days (see, for example, [1-4]). These features must be captured by an adequate model.

In [9, 14] we proposed the hybrid model of human mobility that combines all the important features of human mobility listed above. Our models are based on the models proposed in [7, 8], but more effective in the simulation. Also in the model presented in this report, the persistence of individual routes was more consistently captured by introducing a special characteristic – *the coefficient of persistence*.

In this report we present the results of implementation of our mobility model in the OMNET++ simulation system. The characteristics of some popular routing protocols of DTN networks are investigated, namely, the LET (Last Encounter Time) protocol, the MFV (Most Frequent Visible) protocol, and the PROPHET (Probabilistic Routing Protocol using History of Encounters and Transitivity) protocol [15]. For these protocols, the message delivery probability  $Pr(\textit{delivery})$  and the average message delivery time ( $\overline{TTL}$ ) for various network scenarios are evaluated and compared to find the most effective protocol for considered scenarios.

## 2. Short description of the hybrid mobility model

The detailed description of the hybrid model is given in [9, 14], so here we will only briefly describe its main features and some details which were not considered previously so much.

Movements of one single node (i.e., a person) are considered as straight-line movements between waypoints, where nodes stop at a random time. Lengths of these displacements and pause times are random variables, with distributions close to the Levy distributions. Moreover, waypoints are grouped into clusters, also called hot spots, because they actually correspond to places (we will call them locations) where people spend considerable time during their daily activities (for example, buildings, where they work). In this case, to take into account the routes' persistence, the coefficient of persistence is introduced, which is equal to the fraction of locations that are stored in the routes of one node in different days. Since the modeling of mobility in a period of several days in previous works was considered small, here we will consider it in more detail.

For better adequacy of the hybrid model, both the locations and durations of the daily nodes' routes are taken from the real data from [13]. In particular, we used the traces dataset from the territory of KAIST. It contains data of the one-day travel of several dozen students across the campus of the Korea Advanced Institute of Science and Technology.

In the records of real routes, the movements take an average of 12 hours. The minimum route takes 4.2 hours and the maximum route – 23.3 hours. Because of durations of the routes are different, the model allows the forced end of the route, which happens after a predetermined constant time interval – a *model day*. The duration of model day  $d$  can be changed, depending on the nature of movements in a particular area. This article presents the simulation results with duration of the model day equal to the average route's duration of all nodes from real traces.

Due to the fact that different routes take different time, each user who finishes his route (including compulsory endings), returns to the home location – the location from which the movement is started. If the user's route is longer than the model day,

then at the end of the model day the user should go home. After returning to the home location the user “falls asleep” until the next model day, i.e. ceases to be the source of messages on the network.

Before the start of a new day, routes are changed according to the coefficient of persistence  $p$  – which means the proportion of replaceable visited locations in the entire route. The coefficient  $p$  was introduced to allow the route to be changed day by day for the purpose of simulation the changes in daily route in reality.

Each location in a real route can be visited several times by the user. The number of such possible visits is called the *multiplicity* of the location. While forming a new route, all visits from all locations are summed up and part of the total sum is excluded from new route (this part is determined by the coefficient of persistence  $p$ ). Further, the number of waypoints in the excluded locations is counted to be added later. After deleting visits, some locations are added randomly in the route from a set of all locations, except remaining in the route after the deletion. Adding locations can be repeated to reproduce the multiplicity of locations. Waypoints in new locations are added based on the previously calculated sum of excluded waypoints, in order to save the total number of waypoints in the new route the same as before the change. This logic of generating a new route is designed to ensuring that new route’s duration is close to the duration of the previous route.

However, in addition to fitting new generated routes to previous routes (or to the first one), the first model route should be fitted to the real route by duration. For this purpose, the parameters of the pause time generator are used. The durations of the pauses between have Levy distribution [9, 14] with the parameters  $c$  and  $\alpha$ . To change pause times, the scaling parameter  $c$  was chosen to minimize the mean square deviation of the route durations for the first day of real traces from the route durations of the first day for simulated traces.

### 3. Routing protocols

To describe the routing protocols discussed in this report, we introduce several definitions. *Direct neighbors* or simply *neighbors* are those nodes that have an active network connection with the current node at a given time (i.e., in the range of the communication device). The process of packet routing is that it is necessary to determine which of the neighbors at the given time is the most profitable to transfer this packet, so that it subsequently reaches the target node with the greatest probability if there is no direct connection with the target node at the given time.

First of all, all DTN protocols use packet transfer logic in *one hop* – if node  $i$  has a packet addressed to node  $j$ , then check the direct connection to node  $j$  and the packet is transmitted to it if there is a connection. If there is no target node in the number of neighbors, but there is a neighbor who is also a neighbor for the target node, then the packet is sent to this neighbor (if there are several of them, then any of them). This is a *two-hop* packet transmission logic, which is also always used.

If simple logics cannot find the target node or a suitable transit node, then one of the protocols starts (for example, [10]):

- The Last Encountered Time (LET) protocol;
- More Frequently Visible (MFV) protocol;
- proposed LET-MFV protocol with switching threshold (*hybrid protocol*);
- PROPHET protocol [15].

The LET protocol sends a packet from node  $i$  to that neighbor, who later than another “saw” the target node  $j$  (i.e., had an active network connection with this node). Comparison is made also with the current node  $i$ . If there are several such nodes, then the packet is sent to the random one. If none of the neighbors have “seen” the target node, then the packet is not being transferred to anyone. In general, during the routing process, the packet “strives to catch up” with its target node.

The MFV protocol works by using the history of the frequency of meeting nodes with each other. The packet is sent from node  $i$ , to that transit node  $k$ , which more often sees the target node  $j$ . The measure of the meeting frequency between nodes is defined as the ratio of the total duration of the network connection between two nodes to the entire simulation time. The total duration is calculated by the width of the sliding “window” in simulation days. The width of this sliding “window” (in model days) is a parameter of the model.

For the availability of the MFV protocol, firstly we have to collect a story about the frequency of meetings between nodes. For this purpose, the model has the download phase, during which the collection of statistics is disabled. The duration of this phase is equal to the width of the sliding “window”, i.e. as soon as the required number of days has passed, equal to the width of the window, statistics collection begins.

The hybrid protocol based on LET and MFV (LET-MFV) protocols is implemented. It uses LET only up to a certain time threshold, after which the MFV protocol starts to work. First the LET protocol tries to find a solution about the best transit node. If all neighbors for the current node “saw” the target node later than the threshold, then the protocol switches to the MFV part. Such logic should make the routing situation more optimistic, because of it simulates the accounting for obsolescence of information about when the nodes “saw” each other. After the threshold has been reached the MFV protocol based on the collected statistics about the frequency of meetings starts to work.

Finally, a simplified version of the PROPHET protocol is implemented. Instead of doing unassembled replication of packets on network nodes during the distribution of packets, as simple protocols based on replication do, PROPHET implements “probabilistic routing” [15].

### 4. Experimental results

Hybrid model was implemented in the OMNeT ++ simulation environment [11] using the INET framework [12] (more detailed in [9, 14]) to compare simulation results of routing protocols. The purpose of the experiments is to research the

behavior of the LET, MFV, LET-MFV, and PROPHET protocols depending on the number of nodes  $N$  and the coefficient of persistence  $p$  of traces. Research provided by comparing the target characteristics of the routing protocols: the PDF of packet's delivery delay or time of live of packet  $CCDF(TTL)$  and probability of delivery  $Pr(delivery)$ .

This report uses the traces dataset from the territory of KAIST from collection [13]. All traces were collected in the same way: a number of volunteers (university students) wore GPS receivers in their pocket during the day and these receivers record their position every 30 seconds. These data were used to find real waypoints, waypoint clusters and other parameters for the hybrid model.

For the MFV part of hybrid protocol: the width of the "window" during which the information about the meetings is collected is equal to 5 model days. The duration of the threshold is also equal to 5 model days. As mentioned above, this threshold is required to "load" the information about the meetings before the MFV part starts to work. The parameters of the generator of pause times are  $c = 18$  and  $\alpha = 0.5$ . The parameters of the generator for movements' length are the same as in works [9, 14]. The radius of the transmitters of nodes is 100 meters.

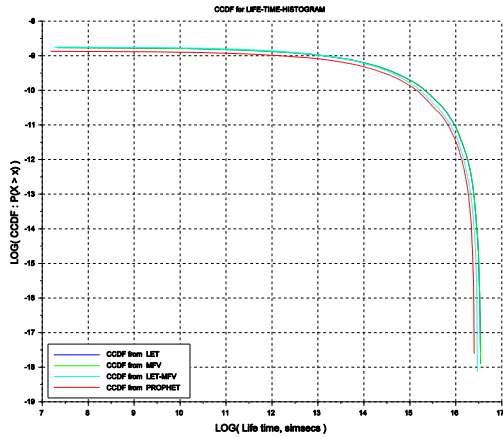


Fig. 1. Comparison of distributions  $CCDF(TTL)$  for  $N = 12$  and  $p = 0.5$ .

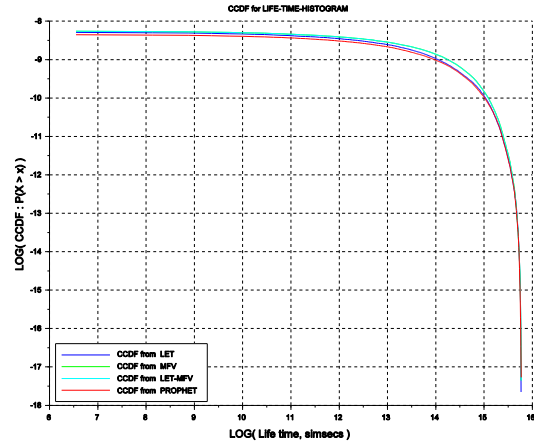


Fig. 2. Comparison of distributions  $CCDF(TTL)$  for  $N = 23$  and  $p = 0.5$ .

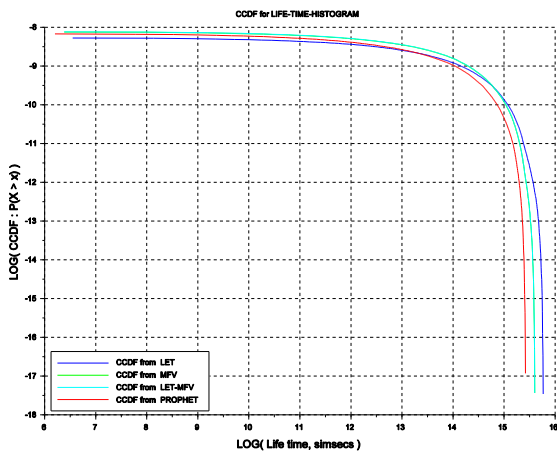


Fig. 3. Comparison of distributions  $CCDF(TTL)$  for  $N = 46$  and  $p = 0.5$ .

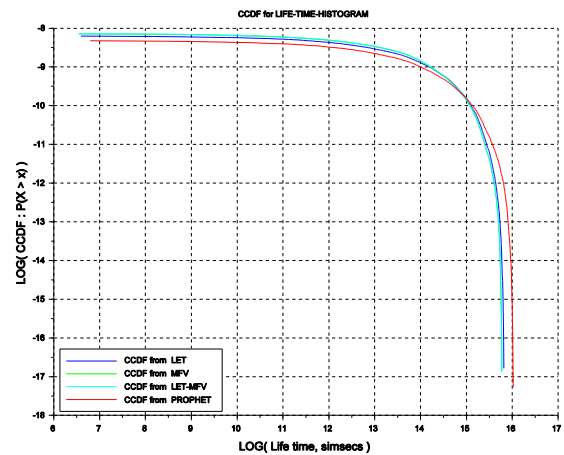


Fig. 4. Comparison of distributions  $CCDF(TTL)$  for  $N = 12$  and  $p = 0.9$ .

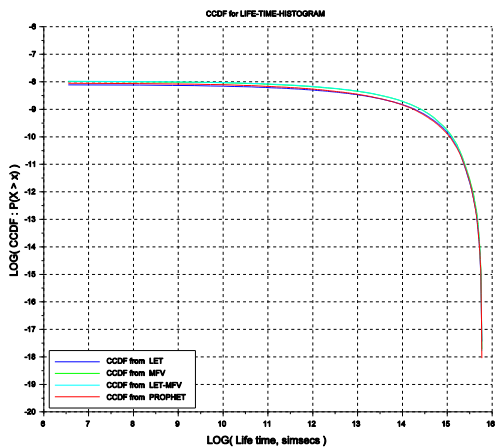


Fig. 5. Comparison of distributions  $CCDF(TTL)$  for  $N = 23$  and  $p = 0.9$ .

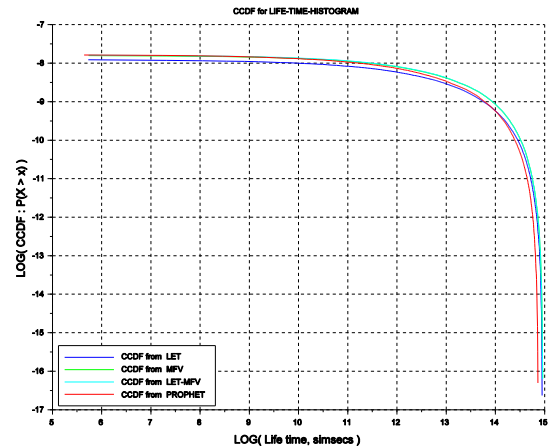


Fig. 6. Comparison of distributions  $CCDF(TTL)$  for  $N = 46$  and  $p = 0.9$ .

The experiments were made with a coefficient of persistence  $p = 0.5$  and  $p = 0.9$ . The number of nodes  $N$  is varied in the experiments: 12, 23, and 46. The duration of model day  $d$  was equal to 12 model hours – this value based on a selective average of the durations of all real routes from the given territory. In figures 1, 2, and 3 the  $CCDF(TTL)$  functions for packet's PDF of time to live for a different number of nodes  $N$  with a coefficient of persistence  $p = 0.5$  are shown. In figures 4, 5 and 6  $CCDF(TTL)$  functions for a different number of nodes  $N$  with a coefficient of persistence  $p = 0.9$  are shown. The estimations of the average packets' time to live  $\overline{TTL}$  are presented in Table 1. The estimated probabilities of packets' delivery  $Pr(delivery)$  for all runs of models are shown in Table 2.

Table 1. Estimation of average time-to-live  $\overline{TTL}$  (measured in simulation seconds).

Count of nodes ( $N$ )	$p = 0.5$				$p = 0.9$			
	LET	MFV	LET-MFV	PROPHET	LET	MFV	LET-MFV	PROPHET
46	52060.2	49784.8	<u>49575.9</u>	56642.8	45952.1	43695.7	<u>43022.3</u>	43560.5
23	52522.1	50324.6	<u>50176.4</u>	55306.7	43013.1	35147.3	<u>34871.2</u>	37608.1
12	63177.9	63177.9	<u>62928.5</u>	80434.9	45022.8	42577.7	<u>42405.2</u>	46661.0

Table 2. Probability estimation of delivery of packets  $Pr(delivery)$ .

Count of nodes ( $N$ )	$p = 0.5$				$p = 0.9$			
	LET	MFV	LET-MFV	PROPHET	LET	MFV	LET-MFV	PROPHET
46	0.6786	<u>0.6939</u>	0.6938	0.6695	0.7945	0.8006	0.8026	0.7742
23	0.6390	<u>0.7375</u>	0.7361	0.6963	<u>0.8072</u>	0.7909	0.7923	0.7575
12	<u>0.6951</u>	<u>0.6951</u>	0.6935	0.8023	<u>0.7387</u>	0.7009	0.7011	0.7077

## 5. Conclusion

The results of simulating of popular routing protocols in DTN networks with a hybrid mobility model of nodes are presented. The message delivery probability and average time-to-live of message was evaluated. As a result of the experiments, it was found that with a small average density of nodes and with an average persistence coefficient, the MFV protocol surpass the other protocols in case of the probability of message delivery. With a large density of nodes and a large coefficient of route persistence, the LET protocol has the advantage in the probability of message delivery, however the best protocol in case of the average delivery time for all considered parameters of nodes' mobility is the protocol LET-MFV.

PROPHET protocol has worse characteristics than others, but it is necessary to note that our implementation of this protocol was simplified (for example, without implementation of replication of packets) and we used the recommended parameters in [15] without deep optimization. So, investigation of applicability of our hybrid mobility model and more deep comparison of considered routing algorithms is the direction of our further research.

## References

- [1] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. *Nature* 2006; 439: 462–465.
- [2] Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. *Nature* 2008; 453: 779–782.
- [3] Rhee I, Shin M, Hong S, Lee K, Chong S. On the Levy walk nature of human mobility. *Proc. IEEE INFOCOM*, Phoenix, AZ 2008; 924–932.
- [4] Rhee I, Shin M, Hong S, Lee K, Kim SJ, Chong S. On the Levy-walk nature of human mobility. *IEEE/ACM Trans. on Networking* 2011; 19(3): 630–643.
- [5] Lim S, Yu C, Das CR. Clustered mobility model for scale-free wireless networks. *Proc. IEEE LCN Tampa, FL 2006*; 231–238.
- [6] Ghosh J, Philip SJ, Qiao C. Sociological orbit aware location approximation and routing (solar) in MANET. *Ad hoc Netw.* 2007; 5: 189–209.
- [7] Lee K, Hong S, Kim SJ, Rhee I, Chong S. SLAW: Self-Similar Least-Action Human Walk. *IEEE/ACM Trans. on Networking* 2012; 20(2): 515–529.
- [8] Lee K, Hong S, Kim SJ, Rhee I, Chong S. Demystifying Levy Walk Patterns in Human Walks. Technical Report in CSC, NCSU 2008. URL: [https://www.csc.ncsu.edu/research/tech/reports.php/Demystifying\\_Levy\\_Walk\\_Patterns.pdf](https://www.csc.ncsu.edu/research/tech/reports.php/Demystifying_Levy_Walk_Patterns.pdf) (28.01.2017).
- [9] Privalov AY, Tsarev AA. Hybrid Model of Human Mobility for DTN Network Simulation. *Proceedings of 30th European Conference on Modelling and Simulation (ECMS2016)*. Regensburg university of applied sciences, Regensburg, Germany 2016; 419–424.
- [10] Dubois-Ferriere H, Grossglauser M, Vetterli M. Age matters: Efficient route discovery in mobile ad hoc networks using encounter ages. *Proc. ACM MobiHoc*, Annapolis, MD 2003; 257–266.
- [11] Varga A. The OMNeT++ discrete event simulation system. *Proceedings of the European simulation multiconference 2001*; 9(185.sn).
- [12] Till S, Kenfack HD, Korf F, Schmidt ThC. An extension of the OMNeT++ INET framework for simulating real-time ethernet with high accuracy. *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques, ICST*. Brussel, Belgium 2011; 375–382.
- [13] Kotz D. Community Resource for Archiving Wireless Data at Dartmouth. Dartmouth College 2015. URL: <http://www.crowdad.org/index.html> (28.01.2017).
- [14] Tsarev AA, Privalov AY. Hybrid Model of Human Mobility for DTN Network Simulation in Comparison with SLAW-type Model. *Proceedings of 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP16)*. Czech Republic 2016. URL: <http://www.csndsp16.com/csndsp16.zip> (28.01.2017).
- [15] Lindgren A, Doria A, Davies E, Grasic S. Probabilistic Routing Protocol for Intermittently Connected Networks 2012. URL: <https://tools.ietf.org/html/rfc6693> (28.01.2017).

# Two-stage approach to real-time assignment of Web Studio customer support tasks

S. Begenova<sup>1</sup>, T. Avdeenko<sup>1</sup>

<sup>1</sup>Novosibirsk State Technical University, Prospekt K. Marksa 20, Novosibirsk, 630073, Russia

---

## Abstract

Many Web-site owners turn to Web Studios for help in solving the problems of their web projects' support, mostly on a subscriber basis. If the Web Studio is professionally engaged in technical support of web projects, then it can have dozens and hundreds of clients, who create hundreds of job requests per month. However, Web Studio's human resources are usually limited. Therefore, the resource allocation problem is of great current interest (by resources we mean the programmers involved in the implementation of this type of work). In this article, we propose two - stage approach for determining an optimal schedule of Web Studio's customer support tasks. The algorithm of dynamic assignment of tasks in real time is implemented taking into account the dynamic character of this process. The performance of the proposed approach is investigated.

*Keywords:* dynamic scheduling theory; mathematical model; customer service; real time assignment; Web Studio; optimization

---

## 1. Introduction

Wide range of specialists is involved while working on the projects implemented by Web Studios. They are managers, programmers, business analysts, designers and other technical personnel. However, according to the Web Studio managers, the main force on which projects are held, and the main critical resource, are programmers.

In accordance with [1] the main areas, the programmers of Web Studios are working on, are:

- Web - sites development (requirements gathering, development of key pages' prototypes, technical specifications development, development of Web-site design, layout of Web-site design, programming and setup, testing and project implementation);
- Guarantee (subscriber) service.

Many Web - site owners turn to Web Studios for help in solving the problems of their web projects' support, mostly on a subscriber basis. Subscriber service means site's technical support and solution of tasks set by the developer or the client within the paid hours. If the Web Studio is professionally engaged in technical support of web projects, then it can have dozens and hundreds of clients, who create hundreds of job requests per month. In this case the support contract usually fixes the maximum response time to the client's request. Web Studio human resources are usually limited. Therefore, the resource allocation problem is of great current interest (by resources we mean the programmers involved in the implementation of this kind of work).

Resource allocation problem is usually used to solve the problem of assigning resources (machines, programmers) to the tasks that need to be performed [2 – 5]. However, in the case of solving the problem of customer support allocation tasks, the usage of static methods of scheduling theory is not enough. In most real-world situations, unforeseen circumstances continually arise that require schedules revision or modification. Such circumstances can concern either resources or operations. Event related to operations is, for example, changing deadlines, canceling orders, late arriving orders, changes in the production process due to the replacement of resources, etc. Thus, the schedule often becomes irrelevant even before its completion.

Such situation forced the development of the so-called dynamic scheduling theory [6 – 8], a set of approaches that respond to unexpected events, either by adjusting the existing schedule, or by rescheduling the remaining operations. It is more efficient to use the dynamic scheduling theory within the terms of present dynamic environment. The present article considers and analyzes two - stage approach to scheduling Web Studio customer support tasks. At the first stage of the approach we adapt mathematical model for multi-skilled project scheduling [9 – 12], the solution of which allows us to construct the initial static schedule. At the second stage of the approach we propose special algorithm for real-time assignment based on the schedule prepared on the first stage.

The rest of the paper is organized as follows. Section 2 considers the current workload of Web Studio programmers based on the real data presented in the form of Gantt chart. Section 3 sees into two-stage dynamic approach. Subsection 3.1 introduces mathematical model for multi-skilled project scheduling and explains its application with usage of the IBM ILOG CPLEX software. The problem is formulated mathematically as a bi-objective optimization model to minimize total costs of processing the tasks and to minimize reworking risks of the tasks, concurrently. In the subsection 3.2 we propose an algorithm of assigning the tasks in real time. In section 4 we give the performance results of the implemented two-stage approach. Section 5 summarizes accomplished work and gives the conclusion.



## 2. Analysis of Web Studio data

To solve the problem of distribution of Web Studio customer support tasks, a particular set of real data on tasks, performed on a certain period of time within the subscriber service, was obtained and analyzed. The existing method of this Web Studio tasks allocation is the usage of software that works on the principle of priority and the "manual" assigning the tasks for resources (Web Studio programmers) by the project manager. In this case, the tasks are distributed using the priority principle in accordance with which the tasks are assigned to the following values depending on their urgency:

- normal,
- urgent,
- critical.

The consequence of manual assignment of the tasks is "idle periods" in the schedule of programmers, i.e. periods of time in which the programmer does not have tasks to perform. To analyze the workload of the programmers' current schedule, a Gantt chart was constructed on the data received by the Web Studio and presented in figure 1.



Fig. 1. Gantt chart for April, 2016.

This diagram shows the schedule of Web Studio programmers for April 2016. The black colored segments show "idle periods", which are the days when the programmers did not perform any customer support tasks. Vertically the numbers - identifiers of the programmers are shown, horizontally - the day of the month. The diagram shows that for the programmers with identifiers 1283, 2674 and 2662 idle times were noticed two times (idle time is considered the absence of tasks for the whole day). The programmers 1328, 2730 have such a number of idle times goes till 5. Taking into account that in these days programmers could perform project tasks alongside with the implementation of projects, the load of programmers in terms of subscriber service is far from complete. Therefore, we are going to consider two approaches of the dynamic construction of schedules as a way to improve and optimize the current allocation of customer service tasks.

## 3. The two-stage approach to real-time assignment

The schedule is built in two stages. Firstly, a long-term (static) schedule is built using multi – skilled project scheduling. Then, with the help of a dynamic approach, new tasks are built in the existing long-term schedule. Thus, new tasks are distributed without changing the current schedule.

### 3.1. Mathematical model for multi-skilled project scheduling with level-dependent rework risk

Let us use the following optimization problem to get long- term schedule – «Multi-skilled project scheduling with level-dependent rework risk» [13]. A multi-skilled version of the resource constrained project scheduling problem was first proposed by Neron and Baptista in [14]. Such mathematical model supposes that each worker has at least one skill. In this problem setting, a group of workforces with predetermined skills should be assigned to perform all required skills involved in each task. In addition, all the workforces assigned to the skills of each task should start executing the task at the same time and should be available until completing all skills of that task [15, 16].

Let  $i, j$  be indexes of the tasks;  $s$  be index for the skills;  $l$  be index for the levels of the skills;  $m$  be index of manpower;  $t, t'$  be indexes of time;  $P_i$  be processing time of task  $i$ .

Let us denote by  $C_{msl}$  – the cost of performing the skill  $s$  by manpower  $m$  at level  $l$  per unit time, by  $b_{is}$  – the required number of workforces that use skill  $s$  on task  $i$ , by  $Pr_{msl}$  – the risk of reworking if manpower  $m$  performs the skill  $s$  at level  $l$  and by

$$r_{msl} = \begin{cases} 1, & \text{if manpower } m \text{ has skill } s \text{ at level } l; \\ 0, & \text{otherwise} \end{cases};$$

Decision variables that have to be determined have following meanings:

$$z_{it} = \begin{cases} 1, & \text{if task } i \text{ is started at time } t; \\ 0, & \text{otherwise} \end{cases};$$

$$x_{imt} = \begin{cases} 1, & \text{if manpower } m \text{ begins to work on task } i \text{ at time } t; \\ 0, & \text{otherwise} \end{cases};$$

$$y_{imsl} = \begin{cases} 1, & \text{if manpower } m \text{ performs level } l \text{ of skill } s \text{ on task } i; \\ 0, & \text{otherwise} \end{cases};$$

Let us consider the following objectives for our mathematical model:

$$Z_1 = \sum_{m=1}^M \sum_{s=1}^S \sum_{l=1}^L (\text{Pr}_{mssl} \cdot \sum_{i=1}^N Y_{imsl}) \longrightarrow \text{Min}; \quad (1)$$

$$Z_2 = \sum_{m=1}^M \sum_{s=1}^S \sum_{l=1}^L (C_{mssl} \cdot \sum_{i=1}^N Y_{imsl}) \longrightarrow \text{Min}; \quad (2)$$

$$Z_3 = w_1 \cdot Z_1 + w_2 \cdot Z_2 \rightarrow \text{Min}; \quad (3)$$

where  $w_1 + w_2 = 1$

The first objective (1) minimizes the total cost of processing the tasks involved in a project. The second objective (2) minimizes the reworking risk of the processed tasks. Both objectives can be taken into account by using weighted coefficients as represented by formula 3. Values of weighted coefficients are chosen by decision – makers, who figure out which objective is more or less important than another one. In the research, the next combinations were used:  $w_1 = w_2 = 0.5$ ;  $w_1 = 0$  and  $w_2 = 1$ ;  $w_1 = 1$  and  $w_2 = 0$ .

The optimization problem has to be solved under the following constraints:

$$\sum_{t=0}^T t \cdot Z_{it} + P_i \leq \sum_{t=0}^T t' \cdot Z_{jt'}, \quad \forall (i, j) \in E; \quad (4)$$

$$\sum_{t=0}^T Z_{it} \leq 1, \quad \forall i; \quad (5)$$

$$X_{imt} \leq Z_{it}, \quad \forall i, m, t; \quad (6)$$

$$X_{imt} + 1 \geq Z_{it} + \sum_{s=1}^S \sum_{l=1}^L Y_{imsl}, \quad \forall i, m, t; \quad (7)$$

$$\sum_{m=1}^M \sum_{l=1}^L Y_{imsl} = b_{is}, \quad \forall i, s; \quad (8)$$

$$Y_{imsl} \leq r_{mssl}, \quad \forall i, m, s, l; \quad (9)$$

$$\sum_{m=1}^M \sum_{t=0}^T X_{imt} = \sum_{s=1}^S b_{is}, \quad \forall i; \quad (10)$$

$$\sum_{i=1}^N \sum_{t'=t-p_i+1}^t X_{imt'} \leq 1, \quad \forall m, t; \quad (11)$$

$$\sum_{t=0}^T X_{imt} \leq \sum_{s=1}^S \sum_{l=1}^L Y_{imsl}, \quad \forall i, m; \quad (12)$$

$$\sum_{l=1}^L Y_{imsl} \leq 1, \quad \forall i, m, s; \quad (13)$$

$$Z_{it}, X_{imt}, Y_{imsl} \in \{0, 1\}, \quad \forall i, m, s, l; \quad (14)$$

The constraint (4) preserves the precedence relations between the tasks. The constraint (5) ensures that each task must be started once. The constraints (6) and (7) imply that all the workforces assigned to various skills of each task should start their work, concurrently. The constraint (8) implies that the number of workforces assigned to different levels of each skill should be equal to the number of manpower required to perform that skill. The constraint (9) ensures that the manpower assigned to a level of each skill should be able to perform it appropriately. The constraint (10) implies that the number of workforces assigned to each task should be equal to the number of required manpower to accomplish that task. Constraint (11) ensures uninterrupted assignment of workforces to different skills of project tasks. In other word, the manpower assigned to a skill of each task must perform the assigned skill continuously without interruption. The constraint (12) implies that each task should be executed by the workforces assigned to different skills of that task. The constraint (13) ensures that the manpower assigned to each skill of an task should perform that skill in her/his predefined level. Finally, the constraint (14) denotes that all decision variables are binary.

Mathematical model for multi-skilled project scheduling with level-dependent rework risk was solved using an optimization software package IBM ILOG CPLEX Optimization Studio. This software is an enterprise analytical decision support toolkit. It enables rapid development and deployment of decision optimization models using mathematical and constraint programming. ILOG CPLEX combines completely featured integrated development environment (IDE) that supports Optimization

Programming Language (OPL) application development to create high-performance ILOG CPLEX optimizer solvers. ILOG CPLEX enables you to optimize your business decisions, develop and deploy optimization models quickly and create real-world applications. The CPLEX code fragment is presented in figure 2.

```

40 // minimizing the total cost of processing the activities involved in a project
41 minimize
42   sum(mm in m)
43   sum(ss in s)
44
45 (
46   Pr[mm][ss]*
47   sum(ii in i)
48   Y[ii][mm][ss]
49 );
50
51
52 subject to
53 {
54   //1st constraint
55   forall(ii in i)
56     forall(jj in j)
57       ct1:
58       sum(tt in t)
59         tt*Z[ii][tt] + P[ii]<=
60         sum(tti in t)
61           tti*Z[jj][tti];
62
63   //2nd constraint
64   forall(ii in i)
65     ct2:
66     sum(tt in t)
67       Z[ii][tt]<=1;
68
69   //3rd
70   forall(ii in i)
71     forall(mm in m)
72       forall(tt in t)
73         ct3:
74         X[ii][mm][tt]<=Z[ii][tt];
75

```

Fig. 2. CPLEX code fragment.

### 3.2. Real-time assignment

The second stage of the construction of schedule for Web Studio customer support tasks is the integration of incoming new tasks into the schedule that was built at the first stage. To implement such an integration we apply the Real-time assignment algorithm.

The real-time assignment algorithm was proposed due to the spread of the supply chain paradigm. In the supply chain, each project encompasses a whole production cycle, starting from customer requirements to final calculations. In addition, the variety of products involved in the project is limited in terms of the number of tasks types. Today, a project is understood as a continuous stream of tasks. In particular, production systems are gradually moving from small-scale production to conveyor assembly lines, which guarantee not only the performance, but also the flexibility of the system. Thus, the assignment of job (set of tasks) in real time is to assign a job to a set of resources upon the arrival of the order in the production system. In the event of a violation of the performance of any task, previously planned unfinished task is redistributed in the order corresponding to the demand. The goal of this approach is to redistribute tasks in real time.

Consider the following problem of assigning jobs in real time with fixed previous assignments. Let us suppose that task  $i$  can be performed by any of the programmers  $\{m_i^1, m_i^2, m_i^3, \dots, m_i^{K_i}\}$ , and the programmer  $m_i^k, km\{1, 2, \dots, K_i\}$  is idle on the periods

$I_i^k = \left\{ \left[ \alpha_{i,q}^k, \beta_{i,q}^k \right] \right\}_{q=1,2,\dots,Q_{k,i}}$ . Thus,  $K_i$  is the maximum number of programmers who are able to perform the task  $i$ , and  $Q_{k,i}$  is the maximum number of idle periods available for task  $i$  of the programmer  $m_i^k$ .

The initial data for the algorithm are the programmers' initial schedules determined at the first stage, the idle periods of the programmers, and the incoming tasks with their time estimates. For the case of several programmers whose performances are identical, we group the idle periods associated with one group of such programmers as follows.

For  $k_1 \neq k_2$ , where  $k_1, k_2 \in \{1, 2, \dots, K_i\}$ , period  $\left[ \alpha_{i,q}^{k_1}, \beta_{i,q}^{k_1} \right]$ ,

$q \in \{1, 2, \dots, Q_{k_1,i}\}$  precedes  $\left[ \alpha_{i,r}^{k_2}, \beta_{i,r}^{k_2} \right]$ ,  $r \in \{q=1, 2, \dots, Q_{k_2,i}\}$  if:

1.  $\alpha_{i,q}^{k_1} < \alpha_{i,r}^{k_2}$ , или
2.  $\alpha_{i,q}^{k_1} = \alpha_{i,r}^{k_2}$  и  $\beta_{i,q}^{k_1} < \beta_{i,r}^{k_2}$ .

The sequence of such sorted periods is denoted by  $\left[ \alpha_i^s, \beta_i^s \right]$ , where  $s = 1, 2, \dots, \sum_{k=1}^{K_i} Q_{k,i} = Q_i$

Let  $S_i$  be a rank of the idle period assigned to the  $i$ -th operation in the operation sequence;  $t_i$  be a starting time of the operation  $i$ ;  $\theta_i$  be processing time required to complete the operation  $i$ . Let us denote by  $\delta_i$  - maximum overstay permitted for

the operation  $i$ , on the corresponding resource; by  $\alpha_i^{s_i}$  - starting instant of the  $s_i$ -th idle period that could be assigned to the operation  $i$  and by  $\beta_i^{s_i}$  - finishing instant of the  $s_i$ -th idle period.

In this approach, the following Real-time assignment algorithm is used:

1. Set  $s_i = 1$  for  $i = 1, 2, \dots, m$ ;
2. Set  $p_1 = \alpha_1^{s_1}$ ;
3. Set  $p_i = \text{MAX}(\alpha_i^{s_i}, p_{i-1} + \theta_{i-1}), i = 2, \dots, m$ ;
4. Set  $p_{m+1} = p_m + \theta_m$ ;
5. Set  $t_{m+1} = p_{m+1}$ ;
6. Set  $t_i = \text{MAX}(p_i, t_{i+1} - \theta_i - \delta_i)$  for all  $i = m, m-1, \dots, 1$ ;
7. If  $(t_{i+1} > \beta_i^{s_i})$  for all  $i = 1, 2, \dots, m$ , then the optimum is reached;  
 else, for each  $i$  such that  $(t_{i+1} > \beta_i^{s_i})$ , set  $s_i = s_i + 1$  and go to step 2.

**4. Performance results**

Here are the results of solving the optimization problem «Multi-skilled project scheduling with level-dependent rework risk» for the test case with 3 workers (programmers), 2 skills and 4 tasks, see figure 3, figure 4. The optimal value of the decision variable  $Y$  is presented in figure 3 (last column). The variable  $Y$  is equal to 1, if manpower  $m$  performs the skill  $s$  on the task  $i$ , otherwise 0.

For example, in the third row manpower  $m_2$  does perform the skill 1 on the operation 1, but he does not perform the skill 2 on operation 1 (the fourth row). Dimension of the vector  $Y$  is  $M \cdot S \cdot N$ , where  $M$  is the number of manpower,  $S$  is the number of skills and  $N$  is amount of tasks. Dimension of the resulting vector  $Y$  presenting in figure 3 is equal to  $3 \cdot 2 \cdot 4 = 24$ .

IBM ILOG CPLEX found the most optimal solution taking into account all the above-mentioned constraints.

$i$	$m$	$s$	$Y$
1	1	1	0
1	1	2	0
1	2	1	1
1	2	2	0
1	3	1	0
1	3	2	1
2	1	1	0
2	1	2	1
2	2	1	1
2	2	2	0
2	3	1	0
2	3	2	0
3	1	1	0
3	1	2	1
3	2	1	1
3	2	2	0
3	3	1	0
3	3	2	0
4	1	1	0
4	1	2	0
4	2	1	0
4	2	2	0
4	3	1	0
4	3	2	1

Fig. 3. The resulting vector  $Y$ .

The optimal value of decision variable  $Z$  is presented in figure 4. The variable  $Z$  is equal to 1, if the task  $i$  starts at the time  $t$ , and 0 otherwise. For example, operation 1 starts at time 5 (the sixth row). Dimension of the vector  $Z$  is  $T \cdot N$ , where  $T$  is upper bound of index  $t$  and  $N$  is number of tasks. Dimension of the resulting variable  $Z$  presenting in picture 4 is  $6 \cdot 4 = 24$ .

And combination of resulting variables gives us the long-term schedule of Web Studio customer support tasks for programmers.

Obtained resulting variables gives us following long-term schedule: task 1 starts at time 5 executed by manpower 2 performing skill 1; task 2 starts at time 3 executed by manpower 2 performing skill 2; task 3 starts at time 4 executed by manpower 1 performing skill 2 and manpower 2 performing skill 1; and task 4 starts at time 5 executed by manpower 3 using skill 2.

Let us study the dependence of the optimal solution finding time on the number of tasks, workers or skills. That would help us to understand whether this mathematical problem is applicable in real conditions or not. The table 1 shows the dependence of computation time of optimal solution on the number of tasks performed. It can be seen from the table that with the increase in the number of tasks, the execution time also increases substantially. The study was done for the test case with 10 workers and 2 skills.

i	t	Z
1	0	0
1	1	0
1	2	0
1	3	0
1	4	0
1	5	1
2	0	0
2	1	0
2	2	0
2	3	1
2	4	0
2	5	0
3	0	0
3	1	0
3	2	0
3	3	0
3	4	1
3	5	0
4	0	0
4	1	0
4	2	0
4	3	0
4	4	1
4	5	0

Fig. 4. The resulting vector Z.

Table 1. The dependence of the optimal solution finding time on the number of tasks.

The number of tasks	Computation time
15 tasks	4 sec
50 tasks	7 min 1 sec
100 tasks	43 min 34 sec

The table 2 revealed that the dependence of computation time of optimal solution on the number of workers performed is not as high as in the previous case. So from this side, program’s execution time is feasible in the context of Web Studio work. The study was done for the test case with 50 tasks and 2 skills.

Table 2. The dependence of the optimal solution finding time on the number of workers.

The number of workers	Computation time
15 workers	42 sec
30 workers	2 min 37 sec
100 workers	6 min 23 sec

The table 3 also revealed the dependence of computation time of optimal solution on the number of skills. Obviously, the same as in the table 2, this dependence goes up not so quickly, which is positive in and on itself. The study was done for the test case with 50 tasks and 10 workers.

Table 3. The dependence of the optimal solution finding time on the number of skills.

The number of skills	Computation time
2 skills	20 sec
5 skills	2 min 32 sec
10 skills	10 min 37 sec

In all three cases obtained results are feasible in real - life terms, since long-term schedule is calculated once for a long period of time and then it gets corrected with Real-time assignment algorithm.

Now consider the tests of the software implementation of the Real-time assignment algorithm in multi-paradigm numerical computing environment MATLAB.

*Test case with 2 programmers*

Let us consider a test case where the tasks will be executed by two programmers  $m_1$  and  $m_2$ . The initial data necessary for the execution of the algorithm is presented below.

Select the next execution time required to complete the tasks:

- The first task  $i_1$  will be executed during 1 time unit ( $\theta_1 = 1$ )
- The second task  $i_2$  will be executed during 2 time units ( $\theta_2 = 2$ )

The maximum waiting time that a programmer can have is 1 time unit. Let us define the periods of idle time for programmers, that is, the periods in which they can take a customer service task on a performance.

The idle periods  $I_1$  of the first programmer  $m_1$  are  $[0, 3], [5, 10], [15, +\infty)$ .

The idle periods  $I_2$  of the second programmer  $m_2$  are  $[0, 6], [16, 19], [25, +\infty)$ .

$s$  is the result variable which represents the idle period number on which the specific task was assigned.

The results of algorithm execution are shown in Figure 5. In Figure 5 black rectangles represent the periods of business of the programmers, and green rectangles – the periods of time in which tasks will be performed.

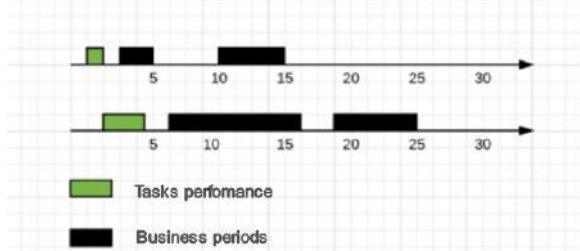


Fig. 5. The solution reached using the algorithm.

In test №1, as we can see two tasks  $i_1$  and  $i_2$  will be executed consistently in the first periods of idle periods of both programmers  $m_1$  and  $m_2$ , respectively. All the idle windows are suitable for performing the related tasks. Furthermore,  $t_i$  is the start time and  $t_{i+1}$  the completion time of the  $i$ -th operation. For each resource, the upper limit of the last window is always  $+\infty$ . As a consequence, applying the above – mentioned algorithm always leads to a solution.

*Test case with 4 programmers*

In this test case, the tasks will be distributed among 4 programmers.

The execution time required to complete the task:

- for the first task  $i_1$ -  $\theta_1=3$
- for the second task  $i_2$ -  $\theta_2=4$
- for the third task  $i_3$ -  $\theta_3=5$
- for the fourth task  $i_4$ -  $\theta_4=5$

The maximum waiting time for a resource is 1.

The idle periods  $I_1$  of the first programmer  $m_1$  are  $[0,2], [5, 15], [25, +\infty)$ .

The idle periods  $I_2$  of the second programmer  $m_2$  are  $[0,15], [25, +\infty)$ .

The idle periods  $I_3$  of the third programmer  $m_3$  are  $[0,5], [10,25], [30, +\infty)$ .

The idle periods  $I_4$  of the fourth programmer  $m_4$  are  $[0,10], [20, +\infty)$ .

The results of the algorithm execution are shown in figure 6.

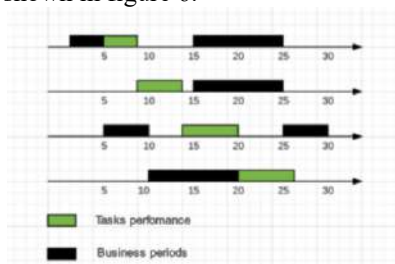


Fig. 6. The solution reached using the algorithm.

For this test, tasks  $i_1, i_2, i_3$  and  $i_4$  will be executed successively in the second idle periods of all programmers except the second one.

*Test case with 8 programmers*

Consider the results of this algorithm in the distribution of tasks for 8 programmers.

The execution time required to complete the task:

- for the first task  $i_1$ -  $\theta_1=3$
- for the second task  $i_2$  -  $\theta_2=2$
- for the third task  $i_3$  -  $\theta_3=3$
- for the fourth task  $i_4$ -  $\theta_4=2$
- for the fifth task  $i_5$  -  $\theta_5=3$
- for the sixth task  $i_6$  -  $\theta_6=5$
- for the seventh task  $i_7$  -  $\theta_7=4$
- for the eighth task  $i_8$  -  $\theta_8=5$

The maximum waiting time for a resource is 1.

The idle periods  $I_1$  of the first programmer  $m_1$  are  $[0,2], [5, 15], [25,30], [35, +\infty)$ .

The idle periods  $I_2$  of the second programmer  $m_2$  are  $[10,15], [25,27], [30, +\infty)$ .

The idle periods  $I_3$  of the third programmer  $m_3$  are  $[0,7], [13, 17], [20,24], [29, +\infty)$ .

The idle periods  $I_4$  for the fourth programmer  $m_4$  are  $[0,5], [10, 12], [16,22], [31, +\infty)$ .

The idle periods  $I_5$  for the fifth programmer  $m_5$  are  $[10,15], [20,27], [30, +\infty)$ .

The idle periods  $I_6$  for the sixth programmer  $m_6$  are  $[7,11], [15,20], [25,30], [35, +\infty)$ .

The idle periods  $I_7$  for the seventh programmer  $m_7$  are  $[6,10], [14,24], [29,35], [43, +\infty)$ .

The idle periods  $I_8$  for the eighth programmer  $m_8$  are  $[0,5], [10,12], [16, 18], [20,22], [31, +\infty)$ .

The results of the algorithm execution are shown at figure 7.

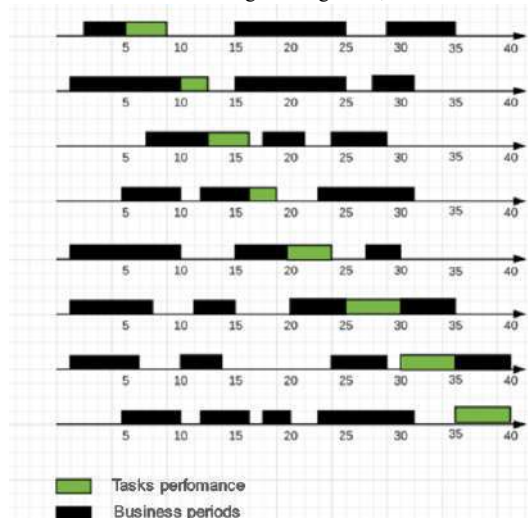


Fig. 7. The solution reached using the algorithm.

In the test cases (figures 5, 6, 7), the tasks were distributed uniformly, taking into account the idle periods of the programmers and their maximum waiting time. The above-given test cases show that the algorithm is looking for the most suitable idle period for the task. If the task does not fit in the idle period, or the maximum waiting time for the resource is exceeded, the algorithm starts looking for another nearest suitable idle period.

Test of the Real-time assignment algorithm revealed the main advantages of this approach. The need to build a new schedule in the event of any breakdown or any other unforeseen circumstance that make the current schedule irrelevant now disappears. The feature of this approach is the fixed previous assignments. Thus, this approach allows you to schedule the execution of incoming tasks without changing the current schedule of programmers. Also, the advantage of this method is the low labor input and high speed, which, undoubtedly, is important for the practical application of the method.

Another advantage of the proposed algorithm is the ability of filling and eliminating idle periods in the programmer schedule. This property allows us to apply this algorithm even when the original schedule is not optimal, as it turned out in the real conditions discussed in Section 2.

## 5. Conclusion

The data analysis of the workload of employees by Web Studio customer support tasks for a certain period showed that the current principle of tasks allocation gives non-optimal workload of the employees. Thus, the conducted analysis revealed the urgency of research and development of methods for more optimal allocation of the customer support tasks. As a result of present research we have proposed the two-stage approach based on dynamic scheduling theory.

The undoubted advantage of this approach is the ability to schedule the work of programmers without changing their current order of execution of the tasks. The Real-time assignment algorithm allows optimally filling the programmers' idle periods and evenly distributing the programmers' workload and reducing the idle periods.

At this stage of the research, the Real-time assignment algorithm considers the process of performing the tasks as a sequential procedure, which does not always suitable for the current principle of executing tasks in Web Studios. Therefore, promising direction for further work is more accurate adaptation of the algorithm to the problem of distribution of the customer support tasks. That is, to take into account the possibility of parallel execution of tasks by programmers [17- 20], the priority of tasks, as well as the priority of the programmer for the task. The latter means the priority of a certain programmer in distribution of the subscriber tasks: the programmer who implemented the web project has a higher priority in the queue for subscriber tasks from this project.

While analyzing the results of the solution of the optimization problem, a significant increase in the time of finding the optimal solution was revealed with an increase in the number of tasks. Therefore, we are going to try reducing this time by integrating the optimization model with different heuristics. This kind of reducing is really important because real – life problems' dimensions are huge. In our case the problem is solved once, and then dynamic approach, which is independent of problems' dimensions, makes its corrections.

## Acknowledgements

The reported study was funded by Russian Ministry of Education and Science, according to the research project No. 2.2327.2017/4.6.

## References

- [1] Avdeenko TV, Petrov RV. On the possibility of applying methods and models of scheduling theory to optimization of working a web-studio. *Sbornik nauchnyh trudov Novosibirsk State Technical University* 2016; 2(84): 7–20. (in Russian)
- [2] Brucker P. *Scheduling Algorithms*. Springer, 2007; 372 p.



- [3] Pinedo M. *Scheduling Theory, Algorithms, and Systems*. Springer, 2008; 672 p.
- [4] Lazarev AA, Gafarov ER. *Scheduling theory. Problems and algorithms*. M.: MSU, 2011; 222 p. (in Russian)
- [5] Pavlov OA, Chernov SK, Misyura OB. Models and algorithms of scheduling theory in planning and project management problems. *Trudy Odessa Polytechnic University* 2006; 1(25): 150–159. (in Russian)
- [6] Dolgui A, Proth J-M. *Supply Chain Engineering - Useful Methods and Techniques*. Springer–Verlag, 2010; 541 p.
- [7] Toroslu I. Personnel assignment problem with hierarchical ordering constraints. *Proc. Personnel assignment Problem with hierarchical ordering constraints* 2003; 493–510.
- [8] Alcaraz J, Maroto C, Ruiz R. Solving the multi-mode resource-constrained project scheduling problem with genetic algorithms. *Journal of the Operational Research Society* 2003; 54(6): 614–626.
- [9] Mika M, Waligora G, Weglarz J. Tabu search for multi-mode resource-constrained project scheduling with schedule-dependent setup times. *European Journal of Operational Research* 2008; 187(3): 1238–1250.
- [10] Jarboui B, Damak N, Siarry P, Rebai A. A combinatorial particle swarm optimization for solving multi-mode resource-constrained project scheduling problems. *Applied Mathematics and Computation* 2008; 195(1): 299–308.
- [11] Abbasi B, Shadrokh S, Arkat J. Bi-objective resource-constrained project scheduling with robustness and makespan criteria. *Applied Mathematics and Computation* 2006; 180(1): 146–152.
- [12] Bouleimen K, Lecocq H. A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version. *European Journal of Operational Research* 2003; 149(2): 268–281.
- [13] Maghsoudlou H, Afshar-Nadjafi B, Akhavan Niaki ST. Multi-skilled project scheduling with level-dependent rework risk; three multi-objective mechanisms based on cuckoo search. *Applied Soft Computing* 2017; 54: 46–61.
- [14] Zhu G, Bard JF, Yu G. Disruption management for resource-constrained project scheduling. *Journal of the Operational Research Society* 2005; 56: 365–381.
- [15] Al-Fawzan M, Haouari M. A bi-objective model for robust resource-constrained project scheduling. *International Journal of Production Economics* 2005; 96(2): 175–187.
- [16] Néron E, Baptista D. Heuristics for multi-skill project scheduling problem. *Int.Symp. Comb. Optim.*2002;
- [17] Avdeenko TV, Mesentsev YA. Efficient approaches to scheduling for unrelated parallel machines with release dates. *IFAC-PapersOnline* 2016; 49(12): 1743–1748.
- [18] Avdeenko TV, Vasiljev MA. Multiagent approach with use of fuzzy modeling in the task multicriterion decision making. *Nauchny vestnik Novosibirsk State Technical University* 2010; 1: 63–74. (in Russian)
- [19] Avdeenko TV. Parameter identification problems in Mathematical modelling of processes. *Obrazovatel'nye resursy i tehnologii* 2014; 1(4): 115–124. (in Russian)
- [20] Mezentsev YuA, Avdeenko TV. Scheduling and Optimization for Parallel-Serial Service Systems. *Proceedings of 8th International Forum On Strategic Technology* 2013; 271–275.



# Digital photoelasticity for calculating coefficients of the Williams series expansion in plate with two collinear cracks under mixed mode loading

L.V. Stepanova<sup>1</sup>, V.S. Dolgikh<sup>1</sup>, V.A. Turkova<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

Photoelasticity method is used to study experimentally the complete Williams series expansion of the stress and displacement fields in the vicinity of the crack tip in isotropic linear elastic plates under Mixed Mode loading. The distribution of the isochromatic fringe patterns is employed for obtaining the stress field near the crack tip by the use of the complete Williams asymptotic expansion for various classes of the experimental specimens (plates with two collinear cracks under tensile loading and under mixed mode loading conditions). The higher order terms of the Williams series expansion are taken into account and the coefficients of the higher order terms are experimentally obtained. The stress field equation of Williams up to fifty terms in each in mode I and mode II has been considered. The comparison of the experimental results and the calculations performed with finite element analysis has shown the importance and significant advantages of photoelastic observations for the multi-parameter description of the stress field in the neighborhood of the crack tip.

*Keywords:* photoelastic stress analysis; elastic behavior; fracture; coefficients of the higher order terms

---

## 1. Introduction

The advent of computers coupled with developments in personal computer digital image processing has had a great influence in the development of modern photoelasticity [1-8]. The term “digital photoelasticity” refers to the automation of the photoelastic data collection and analysis [1]. The classical manual procedure of analysis is usually very tedious and time consuming and requires skilled and experienced personnel. With the advent of digital image processing techniques in photomechanics, digital photoelasticity has become very popular [1-9]. The development of the computer-aided analysis of the experimental data obtained from photoelastic experiment caused rapid growth of the photoelasticity experiments [1-20], especially in fracture mechanics for analysis of the crack tip fields and crack tip parameters. In the past two decades crack-tip mechanics has been studied increasingly using full-field techniques, namely DIC [15] and photoelasticity [1-8].

Thus, in [2] the experimental technique of photoelasticity has been utilized for calculating bi-material notch stress intensities as well as the coefficients of higher order terms. Employing the equations of multi-parameter stress field allows data collection from a larger zone from the notch tip and makes the data collection from experiments more convenient. Moreover, the effects of higher order terms in the region near the notch tip are taken into account. For the photoelasticity experiments, a laboratory specimen known as the Brazilian disk with a central notch, consisting of aluminum and polycarbonate, has been utilized in [2]. Using this specimen, different mode mixities could be easily produced by changing the loading angle. The bi-material notch stress intensities and the first non-singular stress term (called T-stress) were calculated for different test configurations. In order to utilize the advantages of whole-field photoelasticity and minimize the experimental errors, a large number of data points were substituted in the multiparameter stress field equations. Then the resulting system of nonlinear equations was solved by employing an over-deterministic least squares method coupled with the Newton–Raphson algorithm. It has been shown [2] that considering the T-stress term improves, to a large extent, the accuracy of the stress intensities calculated through the photoelasticity technique. Moreover, by reconstructing the isochromatic fringes, the effects of the T-stress term on the shape and size of these fringes around the notch tip were investigated for a 30° notch. The experimental photoelasticity results were also compared with the corresponding values obtained from finite element analysis and a good correlation was observed.

In [3] it is noted that the V-notches are most possible case for initiation of cracks in structure elements. The specifications of cracks on the tip of the notch will be influenced via opening angle, tip radius and depth of V-notch. In [3] the effects of V-notch’s opening angle on stress intensity factor and T-stress of crack on the notch has been investigated. The experiment has been done in different opening angles and various crack length in mode I loading using photoelasticity method. The results illustrate that while angle increases in constant crack’s length, stress intensity factor (SIF) and T-stress will decrease. Beside, the effect of V-notch angle in short crack is more than long crack. These V-notch affects are negligible by increasing the length of crack, and the crack’s behavior can be considered as a single-edge crack specimen.

Guaglianone [4] et al considered the multi-parameter description of the crack tip in isotropic linear elastic materials. The elastic stress field around a crack tip is fully defined through multiparameter equations. A code and program were implemented to evaluate the characteristic parameters of the stress field around a crack tip by photoelastic analysis. The possibility to change the number of parameters makes it possible to adapt the study to different cases, increasing the extension of the analyzed area in order to have a correct modeling of the photoelastic fringes. The performed experimental tests allow emphasizing the importance of using multiparameter equations in the study of the stress field around the crack tip.

Inspired by the Brazilian disk geometry the authors of [5] examine the utility of an edge cracked semicircular disk (ECSD) specimen for rapid assessment of fracture toughness of brittle materials using compressive loading. It is desirable to optimize the geometry towards a constant form factor for evaluating SIF  $K_I$ . In this investigation photoelastic and finite element results for  $K_I$  evaluation highlight the effect of loading modeled using a Hertzian loading. According to authors of [5] a Hertzian loading subtending 4° at the center leads to a surprisingly constant form factor of 1.36. This special case is further analyzed by applying uniform pressure over a chord for facilitating testing.

Lei et al [6] applied photoelasticity method for study of structural imperfection of  $\text{ZnGeP}_2$  crystal. The stresses related to rows and accumulations of dislocations were revealed by photoelastic method for  $\text{ZnGeP}_2$  crystals grown by Vertical Bridgman method. A comparison of information from topographs of photoelastic method and X-Ray topography based on Borrmann method was carried out. It was shown that the strongest contrast is observed on boundaries of dislocation rows and regions of relatively perfect crystals. Photoelastic method gives information about defect structure where X-Ray topography can not be applied because of high density of defect and disorientation of reflection planes. Because of high sensitivity of photoelastic method the images of defects have larger size than in X-Ray topography. That is why in  $\text{ZnGeP}_2$  predominately the total contrast from dislocation rows is fixed. However, in low angle boundaries photoelastic images of separate dislocations were revealed. By comparison with results of simulation it was stated that they are created by edge dislocation of slip system  $\{110\}(110)$  what confirms the data obtained by Borrmann method. Thus, photoelastic method can be, from one side, a simple and express method of analysis of  $\text{ZnGeP}_2$  plates cut along the plane of optical isotropy (001) and, from other side, an analytical method of identification of dislocations and other defects in this material.

In [7] it is noted once again that the photoelastic technique has seen some renewed interest in past few years with digital images and image processing new methods becoming readily available. However, further research is needed to improve the precision, the accuracy and the automation of photoelastic technique. The aim of [7] is to get new numerical equations for the phase-shifting method in digital photoelasticity using a plane polariscope. The model was developed to plane polariscope because of the simplicity and low cost of this equipment. To develop the phase shift and respective intensity equations only the analyzer is rotated. A ring under diametral compression is used for the experimental validation. From these intensity equations the equations for isoclinic and isochromatic parameters are deduced by applying a new numerical technique. This approach can be used to calculate the isoclinic and isochromatic parameters using any number of images. Several analyses are performed with different number of photographic images. The results showed errors reduce when more phase-stepped images are utilized. Hence, one concludes that the uncertainties in results due to effects of errors on photoelastic images can be reduced with a larger amount of phase-stepped images.

In [8] the recent advances in digital photoelasticity have made it possible to use it conveniently for the stress analysis of articles and components made of glass. Depending on the application the retardation levels to be measured range from a few nanometres to several thousand nanometres, which necessitates different techniques and associated equipments. This paper [8] reviews the recent advances in the photoelasticity of glass with a focus on the techniques/methods developed in the last decade.

The aim of the present study is to find coefficients of the multiparameter asymptotic expansion of the stress field in the vicinity of the tips of two collinear cracks in the isotropic linear elastic plate using the photoelasticity method. The main idea of the paper is to keep the higher-order terms of the Williams series expansion of the stress field, to reveal and evaluate the effect of the higher-order terms of the asymptotic expansion. The motivation of this study is twofold. First, we would like to compare theoretical results obtain in [9, 10] with experimental data. According to [9, 10] the higher-order terms of the complete asymptotic expansion of the crack-tip stress field can play a significant role. The more distance from the crack tip the more terms it is necessary to keep in the Williams asymptotic expansion. Thus it follows the second reason when the photoelastic data is processing the number of terms of the asymptotic expansion can't be chosen arbitrary. It depends on the distance from the crack tip. Therefore, the distance from the crack tip should be taken into account when we suppose the structure of the solution in the vicinity of the crack tip.

Note that the coefficients of higher-order terms in the Williams series expansions were computed by different approaches in numerous studies [11-15].

In [11] the digital photoelasticity technique is used to estimate the crack tip fracture parameters for different crack configurations. Conventionally, only isochromatic data surrounding the crack tip is used for SIF estimation, but with the advent of digital photoelasticity, pixel-wise availability of both isoclinic and isochromatic data could be exploited for SIF estimation in a novel way. A linear least square approach is proposed to estimate the mixed-mode crack tip fracture parameters by solving the multi-parameter stress field equation. The stress intensity factor (SIF) is extracted from those estimated fracture parameters. The isochromatic and isoclinic data around the crack tip is estimated using the ten step phase shifting technique. To get the unwrapped data, the adaptive quality guided phase unwrapping algorithm (AQGPU) has been used. The mixed mode fracture parameters, especially SIF are estimated for specimen configurations like single edge notch (SEN), center crack and straight crack ahead of inclusion using the proposed algorithm. The experimental SIF values estimated using the proposed method are compared with analytical/finite element analysis (FEA) results, and are found to be in good agreement.

In [12] the method of photoelasticity is used to study the effects of first non-singular stress term on isochromatic fringe patterns around the tip of a mode I sharp V-notch. Notches are divided into two categories: notches with opening angles a) less than  $45^\circ$ , and b) between two angles  $45^\circ$  and  $152^\circ$ . First, utilizing the mathematical relations of the isochromatic fringes, the effects of the first non-singular stress term on the shape and size of the fringes are studied theoretically. For notch opening angles less than  $45^\circ$ , it is shown that the isochromatic fringes rotate forward and backward when the coefficient of the first non-singular term is negative and positive, respectively. It is also demonstrated that both backward and forward rotations of fringe patterns are possible when the notch angle is between  $45^\circ$  and  $152^\circ$ . For all notch opening angles, as the first non-singular term dominates the notch tip stress field, a new type of fringe appears far from the notch tip. In order to evaluate the analytical findings, a photoelastic test program is also performed on a centrally notched cruciform specimen. Using this specimen, different loading conditions are simulated by changing the lateral load ratio and consequently different effects of the first non-singular term on the shape and size of the fringes are investigated experimentally. Good correlation between the analytical and experimental results is observed.

Harilal et al [13] an experimental study is carried out to estimate the mixed-mode stress intensity factors (SIF) for different cracked specimen configurations using digital image correlation (DIC) technique. For the estimation of mixed-mode SIF's using

DIC, a new algorithm is proposed for the extraction of crack tip location and coefficients in the multi-parameter displacement field equations. From those estimated coefficients, SIF could be extracted. The required displacement data surrounding the crack tip has been obtained using 2D-DIC technique. An open source 2D DIC software Ncorr is used for the displacement field extraction. The presented methodology has been used to extract mixed-mode SIF's for specimen configurations like single edge notch (SEN) specimen and centre slant crack (CSC) specimens made out of Al2014-T6 alloy. The experimental results have been compared with the analytical values and they are found to be in good agreement, there by confirming the accuracy of the algorithm being proposed.

The coefficients of higher-order terms in the Williams series expansions were computed using the DIC method which is a noncontact full-field optical technique [15]. First, the fundamental concepts of DIC method were described and then, this method was proposed to obtain the higher-order terms of the Williams expansion for a CT specimen under pure mode I loading. The displacement field around the crack tip in the CT specimen was determined by the DIC approach. The displacements were utilized in order to obtain the coefficients of Williams expansion. Then, these coefficients were also calculated by using the FE method from the displacement field in the vicinity of the crack tip. The values of stress intensity factor and T-stress obtained from the DIC and FE techniques were compared with the results of previous researches. The efficiency and accuracy of the DIC technique in determining the coefficients of higher order terms in the Williams expansion were demonstrated for the CT specimen. As it is noted in [11] the accuracy of SIF estimate could be improved by improving the accuracy of the isoclinic parameter estimate using the white light photoelasticity thereby eliminating the isochromatic–isoclinic interaction noise. An advanced experimental technique for determination of the stress intensity factor (SIF) and the T-stress is developed in [18] and carefully verified. The approach employs optical interferometric measurements of local deformation response to small crack length increment. Narrow notches are used for crack modeling. Initial experimental data represent inplane displacement component values measured by electronic speckle-pattern interferometry in the vicinity of the crack tip. Determination of the first four coefficients of Williams' series is the main feature of the developed technique. Relationships for transition from measured in-plane displacement components to required fracture mechanics parameters are presented. Availability of high-quality interference fringe patterns, which are free from rigid-body motion, serves as a reliable indicator of real strain state near the crack tip. Experimental verification of the proposed method is performed for non-symmetrical and symmetrical crack in thin rectangular plates subjected to uniaxial tension. The distributions of SIF and T-stress values for cracks of different length in residual stress fields near electronically welded joints of thin plates are presented as an example of practical implementing.

In [19] the common definitions for mode I and mode II are evaluated and improved. For this purpose, the in-plane linear elastic stress field around the crack tip is written as a set of infinite series expansions. Mode I and mode II fields are classically defined as symmetric and anti-symmetric parts of these expansions, respectively. There is also a constant term called "T-stress" in these expansions; parallel to the crack line and independent of the distance from the crack tip. Previous definitions assume that T-stress exists only in pure mode I or combined mode I and mode II conditions. Based on these definitions, T-stress always vanishes in pure mode II. However, the published results of several analytical and experimental researches indicate that the constant stress term can exist in mode II stress field, as well. In this paper, some examples are presented which indicate the presence and importance of T-stress in pure mode II conditions. Then, the classical definition for mode I and mode II is modified to make it consistent with the results presented in the literature.

Thus, the photoelasticity techniques have been extensively used for experimentally determining the state of stress in actual mechanical components. In this research, photoelasticity was employed to assess the singular and higher-order coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip. To utilize the advantages of the whole – field photoelasticity and minimize the experimental errors, the overdeterministic method [18] has been used. The experimental equipment is shown in Fig. 1. The aim of this paper is to obtain the coefficients of the higher-order terms in the Williams expansion and to estimate the influence of these terms on the stress field description taking into account as many as possible terms in the asymptotic presentation of the crack tip fields.

## 2. Elastic stress field around the crack tip

In the development of linear fracture mechanics M. Williams made a major breakthrough in the analysis of the asymptotic stress field at the vicinity of the crack tip in isotropic linear elastic plane media. With the eigenfunction expansion method it is possible to establish the separable variable nature of the solution and to obtain asymptotic expressions for the stress field in a plane medium with a traction-free crack submitted to mode I, mode II and mixed-mode (mode I and mode II) loading conditions:

$$\sigma_{ij}(r, \theta) = \sum_{m=1}^2 \sum_{k=-\infty}^{\infty} a_k^m r^{k/2-1} f_{m,ij}^{(k)}(\theta) \quad (1)$$

with index  $m$  associated to the fracture mode;  $a_k^m$  amplitude coefficients related to the geometric configuration, load and mode;  $f_{m,ij}^{(k)}(\theta)$  angular functions depending on stress component and mode loadings. Analytical expressions for angular eigenfunctions  $f_{m,ij}^{(k)}(\theta)$  are available [9,16]:

$$\begin{aligned} f_{1,11}^{(k)}(\theta) &= \frac{k}{2} \left[ (2+k/2+(-1)^k) \cos(k/2-1)\theta - (k/2-1) \cos(k/2-3)\theta \right], \\ f_{1,22}^{(k)}(\theta) &= \frac{k}{2} \left[ (2-k/2-(-1)^k) \cos(k/2-1)\theta + (k/2-1) \cos(k/2-3)\theta \right], \\ f_{1,12}^{(k)}(\theta) &= \frac{k}{2} \left[ -(k/2+(-1)^k) \sin(k/2-1)\theta + (k/2-1) \sin(k/2-3)\theta \right], \end{aligned} \quad (2)$$

$$\begin{aligned}
 f_{2,11}^{(k)}(\theta) &= -\frac{k}{2} \left[ (2+k/2 - (-1)^k) \sin(k/2-1)\theta - (k/2-1) \sin(k/2-3)\theta \right], \\
 f_{2,22}^{(k)}(\theta) &= -\frac{k}{2} \left[ (2-k/2 + (-1)^k) \sin(k/2-1)\theta + (k/2-1) \sin(k/2-3)\theta \right], \\
 f_{2,12}^{(k)}(\theta) &= \frac{k}{2} \left[ -(k/2 - (-1)^k) \cos(k/2-1)\theta + (k/2-1) \cos(k/2-3)\theta \right].
 \end{aligned} \tag{3}$$

Characteristics of the fracture problems such as the geometry of the specimen and intensity of the load influence neither radial nor angular functions in equation (1) [21]. All the variety of fracture mechanics problems is therefore taken into account in the sole sequence of coefficients  $a_k^m$ . Terms with higher orders have been proven to be influential in some circumstances [9,10,16, 21-34]. In the present paper the experimental technique of photoelasticity was employed to investigate the effects of the higher order terms on both the shape and the size of the near crack tip isochromatic fringes for the wide range of specimens.

### 3. Photoelastic determination of coefficients of the higher order terms in the Williams series expansion

Based on the classical concepts of photoelasticity, the locus of an isochromatic fringe is expressed by the stress optic law

$$2\tau_m = \frac{Nf_\sigma}{h} \tag{4}$$

where  $\tau_m$  is the maximum in-plane shear stress,  $N$  and  $f_\sigma$  represent the fringe order and the material stress-fringe value, respectively,  $h$  is the specimen thickness. On the other hand, the relation between the maximum shear stress  $\tau_m$  and the Cartesian stress components is

$$\tau_m^2 = (\sigma_{11} - \sigma_{22})^2 / 4 + \sigma_{12}^2. \tag{5}$$

The maximum in-plane shear stress around the crack tip can be described by considering  $K$  terms of Mode I and  $M$  terms of Mode II expansion from equation (1). By substituting the truncated series expansion of equation (1) (with  $K$  terms of Mode I and  $M$  terms of Mode II) and equation 5 into equation 4 the mathematical equation for a fringe developed around the crack tip can be written as

$$\begin{aligned}
 \left( \frac{Nf}{2h} \right)^2 &= (\sigma_{11} - \sigma_{22})^2 / 4 + \sigma_{12}^2 = \left( \sum_{k=1}^K a_k^1 r^{k/2-1} f_{1,11}^{(k)}(\theta) + \sum_{k=1}^M a_k^2 r^{k/2-1} f_{2,11}^{(k)}(\theta) - \sum_{k=1}^K a_k^1 r^{k/2-1} f_{1,22}^{(k)}(\theta) - \sum_{k=1}^M a_k^2 r^{k/2-1} f_{2,22}^{(k)}(\theta) \right)^2 / 4 + \\
 &+ \left( \sum_{k=1}^K a_k^1 r^{k/2-1} f_{1,12}^{(k)}(\theta) + \sum_{k=1}^M a_k^2 r^{k/2-1} f_{2,12}^{(k)}(\theta) \right)^2
 \end{aligned} \tag{6}$$

Equation (6) provides  $K$  and  $M$  terms in Mode I and Mode II expansions, respectively to represent the near crack tip stress field. Further the method of evaluation of mixed-mode stress field parameters proposed in [17] is used. The method is called as an over-deterministic method for calculating the stress intensity factor as well as the coefficients of the higher-order terms in the Williams series expansions in cracked bodies.

#### 3.1. Experimental procedure

A set of cracked specimens was used to calculate the stress intensity factor and the coefficients of the higher-order terms of the complete Williams series expansion for the stress field in the vicinity of the crack tip. These specimens are the plate with two collinear crack of the equal length (Fig. 3), the plates with two collinear cracks of different lengths (Fig. 4), the plates with two inclined collinear cracks of the equal and different lengths (Fig. 5).

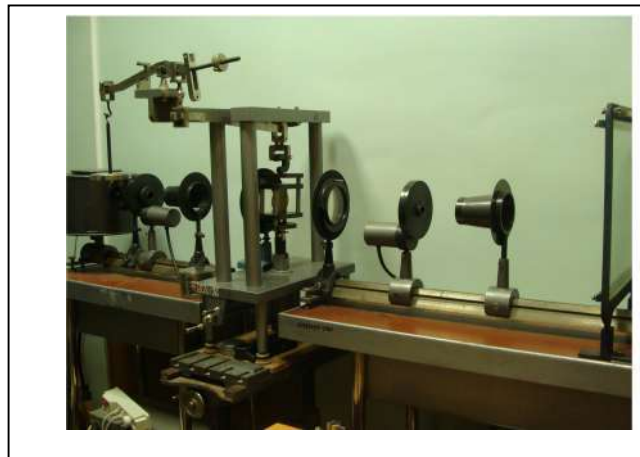


Fig. 1. Experimental setup.

3.2. Calibration of stress optic coefficients and types of experimental specimens

The value of the fringe constant is determined experimentally by inducing a known stress difference  $\sigma_1 - \sigma_2$  in a model that is made of the same material as the specimen of interest by observing the corresponding value of  $N$  and by solving the optic law  $\sigma_1 - \sigma_2 = Nf_\sigma / h$  for  $f_\sigma$ . A common calibration specimen is circular disk (Fig. 2). The isochromatic fringe patten of circular disk under diametral compression is shown in Fig.1. The material fringe constant is  $f_\sigma = 18.33 \text{ Pa cm/fringe}$ .

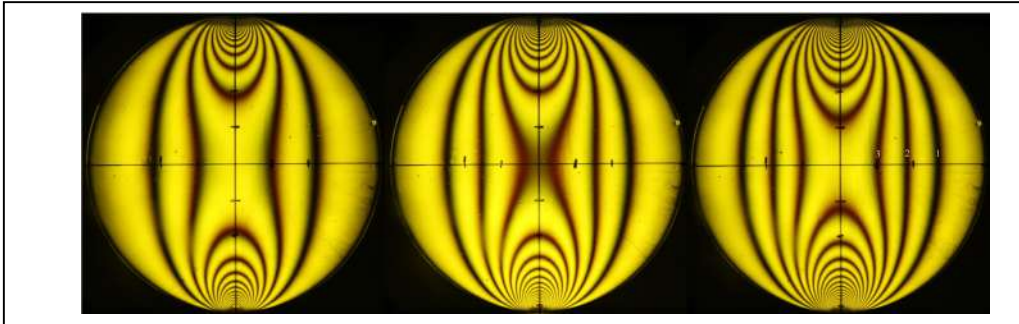


Fig. 2. Photoelastic yellow-light isochromatics for a disk in diametral compression by load 14 N (left), 18 N (center) and 21 N (right).

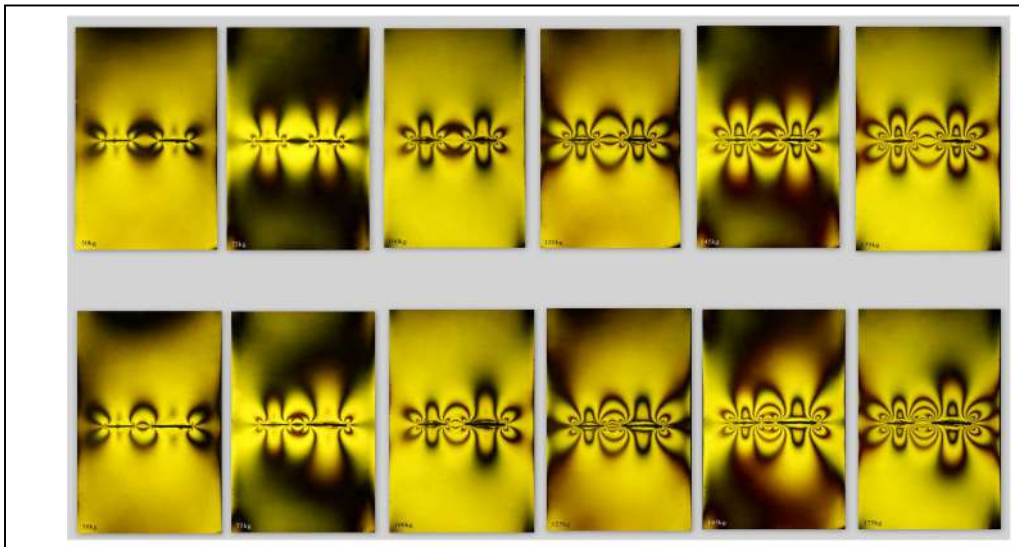


Fig. 3. Photoelastic isochromatic fringe patterns for plates with two collinear cracks under different loads.

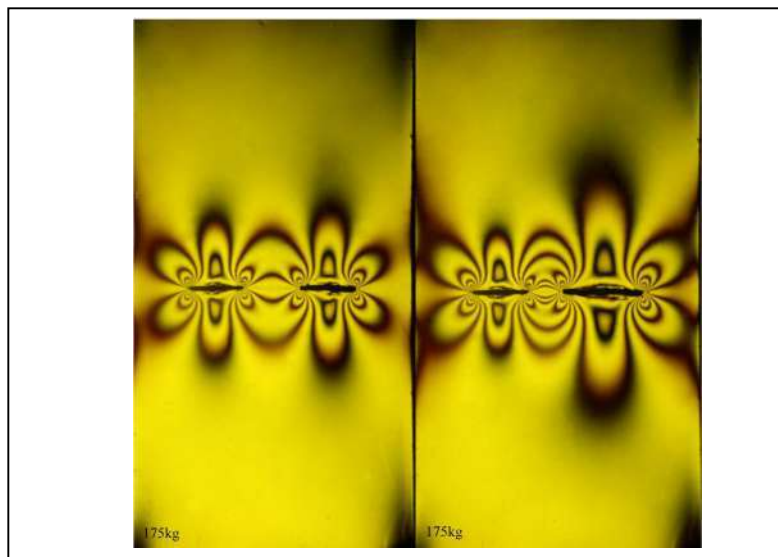


Fig. 4. Isochromatic fringes for a plate with two collinear cracks of different lengths.

After the image acquisition several interactive programs were developed for determining coefficients of higher-order terms. The necessary photoelastic data for a large number of selected points were collected using MATLAB. The algorithm is based on the fact that each pixel of a grayscale photograph has a value or intensity in the range of 0-255, such that a pixel of 0.0 is displayed as black, a pixel value of 255 is displayed as white. Hence, pixels with lower intensities are corresponding to the darker points of grayscale photos. The computer code as it was proposed by M.R. Ayatollahi and M. Nejati in [14] was also developed. The code allows us to find the points of isochromatic fringe pattern. The approach was repeated for several lines in

different directions and the darkest point of every isochromatic fringe in each radial direction were found. The positions of these points were collected and used. The collection of these points is used for determination of the higher order terms of the Williams asymptotic expansion for the stress field in the vicinity of the crack tip. The number of the points depends on the isochromatic fringe. The schematic presentation of the specimen geometry is shown in Fig. 6. The normalized coefficients  $a_k^1 = a_k^1 / \sigma_{22}^\infty$  are obtained for different types of specimens and presented in Tables 1-5.

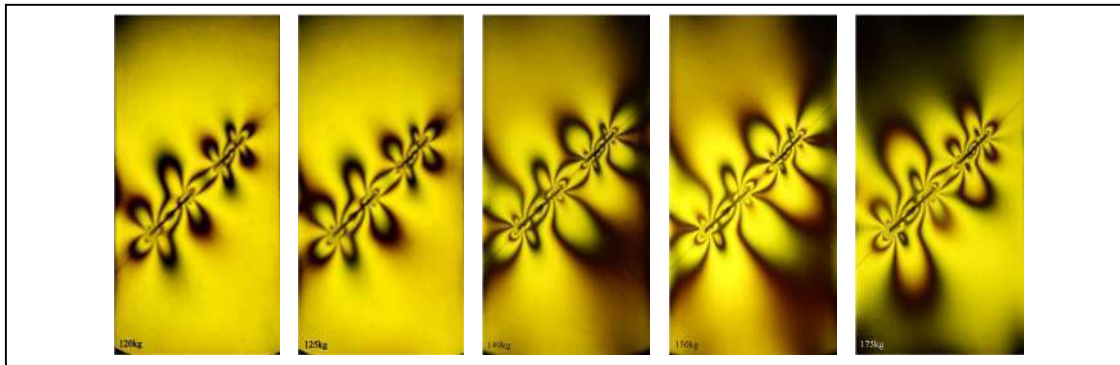


Fig. 5. Isochromatica fringe patterns in plates with two inclined collinear cracks under differents loads.

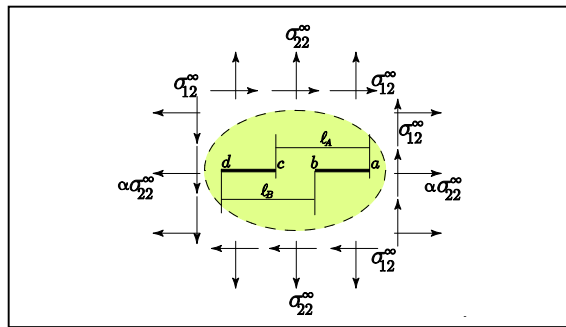


Fig. 6. Schematical presentaion of the specimen geometry considered.

Table 1. The coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip  $z = a$

$$(a = 1.5cm, b = 0.5cm, c = -0.5cm, d = -1.5cm) .$$

$a_1^1 (cm^{-3/2})$	$a_2^1 (cm^{-2})$	$a_3^1 (cm^{-5/2})$	$a_5^1 (cm^{-7/2})$	$a_7^1 (cm^{-9/2})$	$a_9^1 (cm^{-11/2})$	$a_{11}^1 (cm^{-13/2})$	$a_{13}^1 (cm^{-15/2})$	$a_{15}^1 (cm^{-17/2})$
0.5139	-0.2500	0.2512	-0.0646	0.0330	-0.0208	0.0145	-0.0109	0.0086
$a_{17}^1 (cm^{-19/2})$	$a_{19}^1 (cm^{-21/2})$	$a_{21}^1 (cm^{-23/2})$	$a_{23}^1 (cm^{-25/2})$	$a_{25}^1 (cm^{-27/2})$	$a_{27}^1 (cm^{-29/2})$	$a_{29}^1 (cm^{-31/2})$	$a_{31}^1 (cm^{-33/2})$	$a_{33}^1 (cm^{-35/2})$
-0.0069	0.0058	-0.0049	0.00426	-0.0037	0.0033	-0.0029	0.0026	-0.0024
$a_{35}^1 (cm^{-37/2})$	$a_{37}^1 (cm^{-39/2})$	$a_{39}^1 (cm^{-41/2})$	$a_{41}^1 (cm^{-43/2})$	$a_{43}^1 (cm^{-45/2})$	$a_{45}^1 (cm^{-47/2})$	$a_{47}^1 (cm^{-49/2})$	$a_{49}^1 (cm^{-51/2})$	$a_{51}^1 (cm^{-53/2})$
0.0022	-0.0020	0.0018	-0.0017	0.0016	-0.00015	0.0014	-0.0013	0.0012

Table 2. The coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip  $z = \tilde{n}$  (Fig. 6)

$$(a = 1.5cm, b = 0.5cm, c = -0.5cm, d = -1.5cm) .$$

$a_1^1 (cm^{-3/2})$	$a_2^1 (cm^{-2})$	$a_3^1 (cm^{-5/2})$	$a_5^1 (cm^{-7/2})$	$a_7^1 (cm^{-9/2})$	$a_9^1 (cm^{-11/2})$	$a_{11}^1 (cm^{-13/2})$	$a_{13}^1 (cm^{-15/2})$	$a_{15}^1 (cm^{-17/2})$
0.52398	-0.25000	0.27936	-0.04384	0.04700	-0.00820	0.00234	-0.00279	0.001468
$a_{17}^1 (cm^{-19/2})$	$a_{19}^1 (cm^{-21/2})$	$a_{21}^1 (cm^{-23/2})$	$a_{23}^1 (cm^{-25/2})$	$a_{25}^1 (cm^{-27/2})$	$a_{27}^1 (cm^{-29/2})$	$a_{29}^1 (cm^{-31/2})$	$a_{31}^1 (cm^{-33/2})$	$a_{33}^1 (cm^{-35/2})$
-0.00126	0.01027	-0.00068	0.00769	-0.00041	0.00603	-0.00027	0.00489	-0.00018
$a_{35}^1 (cm^{-37/2})$	$a_{37}^1 (cm^{-39/2})$	$a_{39}^1 (cm^{-41/2})$	$a_{41}^1 (cm^{-43/2})$	$a_{43}^1 (cm^{-45/2})$	$a_{45}^1 (cm^{-47/2})$	$a_{47}^1 (cm^{-49/2})$	$a_{49}^1 (cm^{-51/2})$	$a_{51}^1 (cm^{-53/2})$
0.00407	-0.00013	0.00345	-0.00009	0.00298	-0.00007	0.00260	-0.00005	0.00230

Table 3. The coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip  $z = a$  (Fig. 6)

$$(a = 2cm, b = 0.5cm, c = -0.5cm, d = -1.5cm) .$$

$a_1^1 (cm^{-3/2})$	$a_2^1 (cm^{-2})$	$a_3^1 (cm^{-5/2})$	$a_5^1 (cm^{-7/2})$	$a_7^1 (cm^{-9/2})$	$a_9^1 (cm^{-11/2})$	$a_{11}^1 (cm^{-13/2})$	$a_{13}^1 (cm^{-15/2})$	$a_{15}^1 (cm^{-17/2})$
0.62576	-0.25000	0.20400	-0.03470	0.01181	-0.00499	0.00234	-0.00117	0.00061
$a_{17}^1 (cm^{-19/2})$	$a_{19}^1 (cm^{-21/2})$	$a_{21}^1 (cm^{-23/2})$	$a_{23}^1 (cm^{-25/2})$	$a_{25}^1 (cm^{-27/2})$	$a_{27}^1 (cm^{-29/2})$	$a_{29}^1 (cm^{-31/2})$	$a_{31}^1 (cm^{-33/2})$	$a_{33}^1 (cm^{-35/2})$
-0.00033	0.00018	-0.00010	0.00005	-0.00003	0.00002	-0.00001	$0.733 \cdot 10^{-5}$	$-0.442 \cdot 10^{-5}$



Table 4. The coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip  $z = d$

$$(a = 1.5cm, b = 0.5cm, c = -0.5cm, d = -1.5cm).$$

$a_1^1 (cm^{-3/2})$	$a_2^1 (cm^{-2})$	$a_3^1 (cm^{-5/2})$	$a_5^1 (cm^{-7/2})$	$a_7^1 (cm^{-9/2})$	$a_9^1 (cm^{-11/2})$	$a_{11}^1 (cm^{-13/2})$	$a_{13}^1 (cm^{-15/2})$	$a_{15}^1 (cm^{-17/2})$
0.52397	-0.25000	0.27937	-0.04384	0.04700	-0.00820	0.02343	-0.00279	0.014687
$a_{17}^1 (cm^{-19/2})$	$a_{19}^1 (cm^{-21/2})$	$a_{21}^1 (cm^{-23/2})$	$a_{23}^1 (cm^{-25/2})$	$a_{25}^1 (cm^{-27/2})$	$a_{27}^1 (cm^{-29/2})$	$a_{29}^1 (cm^{-31/2})$	$a_{31}^1 (cm^{-33/2})$	$a_{33}^1 (cm^{-35/2})$
-0.00126	0.01027	-0.000068	0.00769	-0.00041	0.00603	-0.00026	0.00489	-0.00018
$a_{35}^1 (cm^{-37/2})$	$a_{37}^1 (cm^{-39/2})$	$a_{39}^1 (cm^{-41/2})$	$a_{41}^1 (cm^{-43/2})$	$a_{43}^1 (cm^{-45/2})$	$a_{45}^1 (cm^{-47/2})$	$a_{47}^1 (cm^{-49/2})$	$a_{49}^1 (cm^{-51/2})$	$a_{51}^1 (cm^{-53/2})$
0.00407	-0.00013	0.00345	-0.00009	0.00298	-0.00007	0.00260	-0.00005	0.00230

Table 5. The coefficients of the Williams series expansion for the stress field in the vicinity of the crack tip  $z = d$

$$(a = 2.5cm, b = 0.5cm, c = -0.5cm, d = -1.5cm).$$

$a_1^1 (cm^{-3/2})$	$a_2^1 (cm^{-2})$	$a_3^1 (cm^{-5/2})$	$a_5^1 (cm^{-7/2})$	$a_7^1 (cm^{-9/2})$	$a_9^1 (cm^{-11/2})$	$a_{11}^1 (cm^{-13/2})$	$a_{13}^1 (cm^{-15/2})$	$a_{15}^1 (cm^{-17/2})$
0.56275	-0.25000	0.31993	-0.02563	0.06089	-0.00065	0.02999	0.00117	0.01853
$a_{17}^1 (cm^{-19/2})$	$a_{19}^1 (cm^{-21/2})$	$a_{21}^1 (cm^{-23/2})$	$a_{23}^1 (cm^{-25/2})$	$a_{25}^1 (cm^{-27/2})$	$a_{27}^1 (cm^{-29/2})$	$a_{29}^1 (cm^{-31/2})$	$a_{31}^1 (cm^{-33/2})$	$a_{33}^1 (cm^{-35/2})$
0.00122	0.01286	0.00106	0.00958	0.00901	0.00749	0.00766	0.00060	0.00065
$a_{35}^1 (cm^{-37/2})$	$a_{37}^1 (cm^{-39/2})$	$a_{39}^1 (cm^{-41/2})$	$a_{41}^1 (cm^{-43/2})$	$a_{43}^1 (cm^{-45/2})$	$a_{45}^1 (cm^{-47/2})$	$a_{47}^1 (cm^{-49/2})$	$a_{49}^1 (cm^{-51/2})$	$a_{51}^1 (cm^{-53/2})$
0.00504	0.00057	0.00027	0.00050	0.00368	0.00044	0.00321	0.00039	0.00284

The theoretically reconstructed isochromatic fringes calculated by the use of the Williams series expansion with the coefficients presented in Tables 1-5 are shown in Fig. 7. Here blue points are experimentally chosen points and the red lines are theoretically obtained contours of the differences of the principal stresses. One can see good consistency the theoretical results with the experimental points. Our analysis shows that the higher order terms of the Williams series expansion play significant role in the description of the stress field in the vicinity of the crack tip.

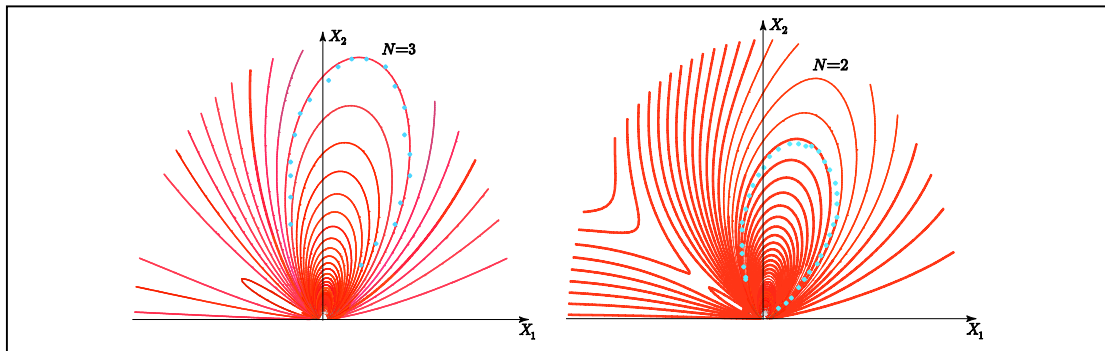


Fig. 7. Experimental points and theoretically constructed isochromatic fringe patterns in the vicinity of the crack tip in the plate with two collinear cracks of different length.

A bitmap image can be constructed by calculating the intensity numerically at each point in a large array of regularly spaced points and observing the resulting pattern. An example is shown in Fig. 10 for dark-field isochromatics in multi-parameter stress field based on Eq. 1.

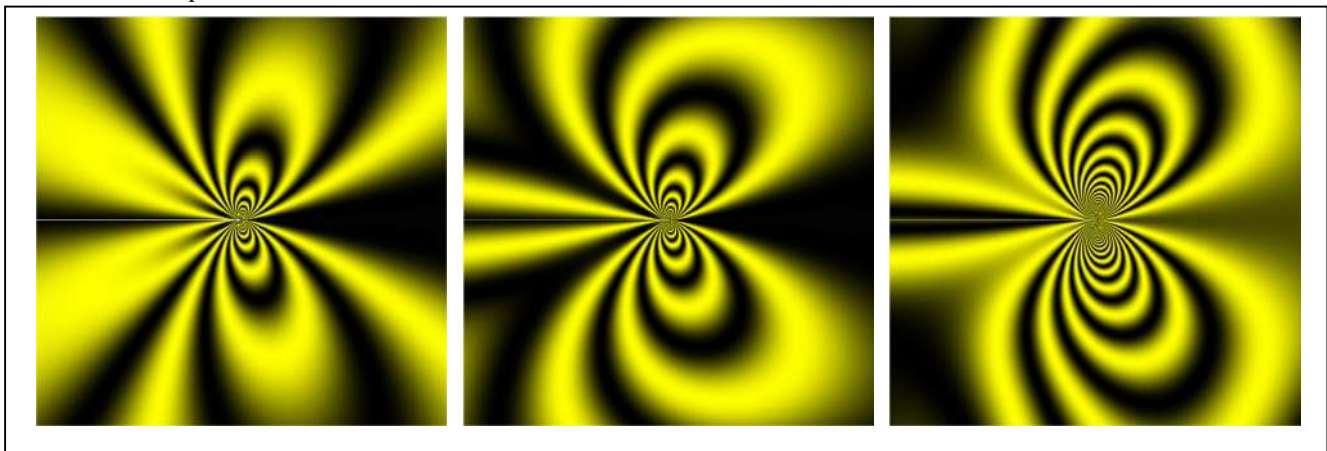


Fig. 8. Simulated multi-parameter stress field in the vicinity of the crack tip with 2,5 and 9 terms of the Williams series expansion of the stress field.

The contours of the Mises equivalent stress in the vicinity of the Mode I crack tip and Mode II crack tip are shown in Fig. 9 and Fig. 10 respectively. From Fig. 9 (left) one can see the contours of the Mises equivalent stress in the vicinity of the Mode I crack tip obtained by the leading order term of the Williams series expansion shown by red lines and the exact solution of the

problem and multi-parameter description of the stress field [16] shown by blue lines. One can see that the asymptotic solution and the exact solution coincide only in the vicinity of the crack tip at very small distances from the crack tip. Increasing the number of terms of the Williams expansion results in extension of the domain where the Williams asymptotic expansion is valid. Finally, Figure 9 (right) shows the contours of the Mises equivalent stress obtained by the Williams expansion keeping 100 terms (red lines). The red lines coincide completely with the exact solution for the infinite elastic medium [16]. The similar results have been obtained for mode II (Fig. 10) and mixed mode crack problems.

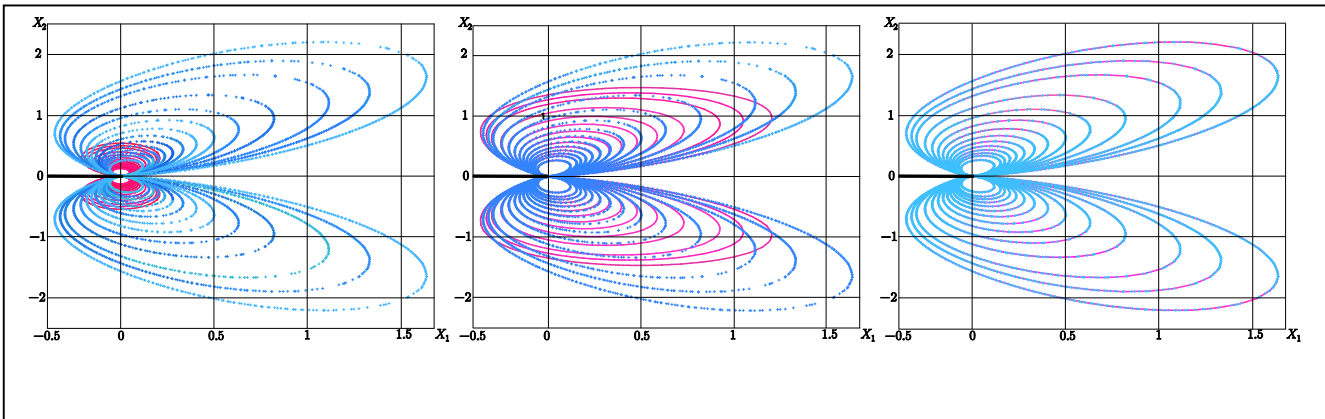


Fig. 9. Influence of the higher order terms of the Williams asymptotic expansion.

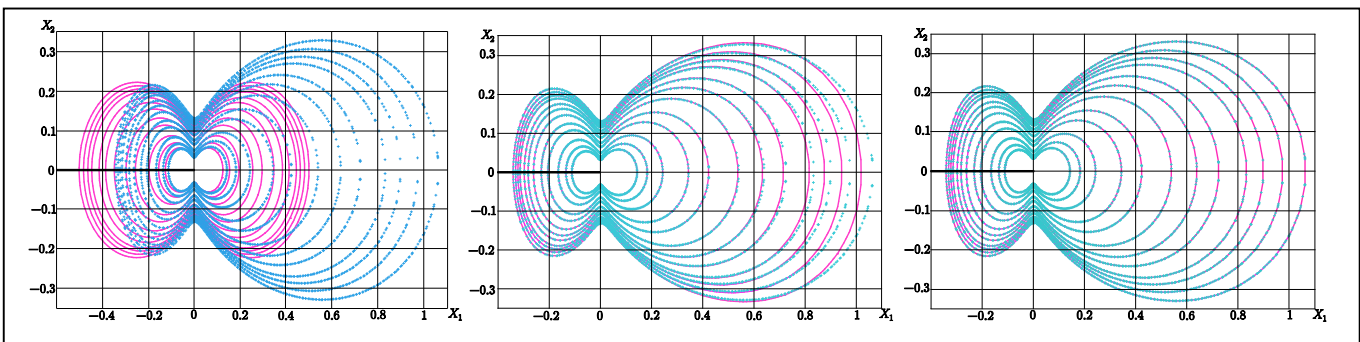


Fig. 10. Influence of the higher order terms of the Williams asymptotic expansion.

The first coefficients of the Williams series expansions  $a_1^I$  and  $a_1^{II}$  were compared with the theoretical results known in literature for stress intensity factors for an infinite plane medium with two collinear cracks [35]:

$$\begin{Bmatrix} K_I \\ K_{II} \end{Bmatrix}_b = \begin{Bmatrix} \sigma_{22}^\infty \\ \sigma_{12}^\infty \end{Bmatrix} \sqrt{\pi(a-b)/2} \frac{1}{\sqrt{1-\alpha_b}} \left\{ 1 - \frac{1}{\alpha_a} \left[ 1 - \frac{E(k)}{K(k)} \right] \right\},$$

$$\alpha_a = (a-b)/l_A, \quad \alpha_b = (c-d)/l_B, \quad k = \sqrt{\alpha_a \alpha_b}.$$

The experimentally obtained coefficients and the theoretical results are in a good agreement. The exact expressions for the coefficients of the higher order terms in the Williams series expansions known for the plate with two collinear cracks have been used either for checking the results of the program developed. The accuracy of the method has been tested for several configuration subjected to pure Mode I and Mixed-Mode loadings.

#### 4. Results and Discussion

In this research, photoelasticity is employed to assess coefficients of the complete Williams series expansion of the linear elastic stress field in the vicinity of the crack tip for a wide class of experimental specimens subject to mixed mode loading: plates with two collinear cracks of equal and different lengths under tensile loading and mixed mode loading. The study has showed that the coefficients of higher order terms can play an important role in fracture process in notched and cracked structures. The approach developed allows us to construct all the higher order terms in the asymptotic expansions in order to better approximate stress field. We use the plates with two collinear cracks under mixed mode loading to construct the complete multi-parameter asymptotic expansion of the stress field in the vicinity of the crack tip.

By means of photoelasticity the distribution of the isochromatic fringe patterns and the stress field near the crack tip based on the complete Williams asymptotic expansion for various classes of the experimental specimens under mixed mode loading are obtained. In the present contribution fracture mechanics problems regarding the study of the stress and displacement fields around a crack tip under mixed mode loading are discussed in the framework of photoelastic techniques and an over-deterministic method for calculation of the coefficients of crack tip asymptotic field. The comparison of the experimental results and the calculations performed with finite element analysis has shown the importance and significant advantages of photoelastic observations for the multiparametric description of the stress field in the neighborhood of the crack tip.



## 5. Conclusion

The study is aimed at experimental and computational determination of the coefficients in the crack tip stress asymptotic expansions for a wide class of specimens under mixed mode loading conditions. In the paper multiparametric presentation of the stress field near the crack tips for a wide class of specimens is obtained. Theoretical, experimental and computational results obtained in this research show that the isochromatic fringes in the vicinity of the crack tip are described more accurately when we consider the complete Williams asymptotic expansion of the stress field and we have to keep the higher order stress terms in the asymptotic expansion since the contribution of the higher order stress terms (besides the stress intensity factors and the T-stress) is not negligible in the crack tip stress field. The example problems revealed the advantage of using a multi-parameter solution in terms of collecting data from a larger zone. A good correlation was observed between the experimental results and the numerical results obtained from finite element analysis.

## References

- [1] Asundi AK. *Matlab for Photomechanics*. Elsevier, London, 2002.
- [2] Aytollahi MR, Mirsayar MM, Dehghany M. Experimental determination of stress field parameters in bi-material notches using photoelasticity. *Materials and Design* 2011; 32: 4901–4908.
- [3] Saravani M, Azizi M. The investigation of crack's parameters on the V-notch using photoelasticity method. *International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering* 2011; 5(1): 152–157.
- [4] Guagliano M, Sangirardi M, Sciuccati A, Zakeri M. Multiparameter Analysis of the stress field around a crack tip. *Procedia Engineering* 2011; 10: 2931–2936.
- [5] Surendra KVN, Simha KRY. Design and analysis of novel compression fracture specimen with constant form factor: Edge cracked semicircular disk (ECS). *Engineering Fracture Mechanics* 2013; 102: 235–248.
- [6] Lei Z, Okunev AO, Zhu C, Verozubova GA, Ma T. Photoelasticity method for study of structural imperfection of ZnGeP<sub>2</sub> crystal. *Journal of Crystal Growth*. 2016; 450: 34–38.
- [7] Magalhaes junior PAA, Magalhaes CA, Magalhaes ALMA. Computational methods of phase shifting to stress measurement with photoelasticity using plane polariscope. *Optik* 2017; 130: 213–226.
- [8] Ramesh K, Ramakrishnan V. Digital photoelasticity of glass: A comprehensive review 2016; 87: 59–74.
- [9] Stepanova LV, Roslyakov PS. Complete asymptotic expansion m. Williams near the crack tips of collinear cracks of equal lengths in an infinite plane medium. *PNRPU Mechanics Bulletin* 2015; 4: 188–225.
- [10] Stepanova LV, Igonin SA. Asymptotics of the near-crack-tip stress field of a growing fatigue crack in damaged materials: Numerical experiment and analytical solution. *Numerical Analysis and Applications* 2015; 8(2): 168–181.
- [11] Patil P, Vyasrayani CP, Ramji M. Linear least squares approach for evaluating crack tip fracture parameters using isochromatic and isoclinic data from digital photoelasticity. *Optic and Lasers in Engineering* 2017; 93: 182–194.
- [12] Ayatollahi MR, Dehghany M, Mirsayar MM. A comprehensive study for mode I sharp V-notches. *European Journal of Mechanics A/Solids* 2013; 37: 216–230.
- [13] Harilal R, Vyasrayani CP, Ramji M. A linear least squares approach for evaluation of crack tip stress field parameters using DIC. *Optic and lasers in Engineering* 2015; 75: 95–102.
- [14] Ayatollahi MR, Nejati M. Experimental evaluation of stress field around the sharp notches using photoelasticity. *Materials and Design* 2011; 32: 561–569.
- [15] Ayatollahi MR, Moazzami M. Digital image correlation method for calculating coefficients of Williams expansion in compact tension specimen. *Optic and Lasers in Engineering* 2017; 90: 26–33.
- [16] Stepanova L, Roslyakov P. Multi-parameter description of the crack-tip stress field: Analytic determination of coefficients of crack-tip stress expansions in the vicinity of the crack tips of two finite cracks in an infinite plane medium. *International Journal of Solids and Structures* 2016; 100–101: 11–28.
- [17] Ramesh K, Gupta S, Kelkar AA. Evaluation of stress field parameters in fracture mechanics by photoelasticity – revisited. *Engineering fracture mechanics* 1997; 56(1): 25–45.
- [18] Pisarev VS, Matvienko YG, Eleonsky SI, Odintsev IN. Combining the crack compliance method and speckle interferometry data for determination of stress intensity factors and T-stress. *Engineering fracture mechanics* 2017; 179: 348–374.
- [19] Ayatollahi MR, Zakeri M. An improved definition for mode I and mode II crack problems. *Engineering fracture mechanics* 2017; 175: 235–246.
- [20] Sestakova L. Using the multi-parameter fracture mechanics for accurate description of stress/displacement crack-tip fields. *Key Engineering Materials* 2014; 586: 237–240.
- [21] Hello G, Tahar M-B, Roeland J-M. Analytical determination of coefficients in crack-tip stress expansions for a finite crack in an infinite plane medium. *International Journal of Solids and Structures* 2012; 49: 556–566.
- [22] Vesely V, Sobek J, Frantik P, Seitl S. Multi-parameter approximation of the stress field in a cracked body in the more distant surroundings of the crack tip. *International Journal of Fatigue* 2016; 89: 20–35.
- [23] Stepanova LV, Yakovleva EM. Mixed-mode loading of the cracked plate under plane stress conditions. *PNRPU Mechanics Bulletin* 2014; 3: 129–162.
- [24] Malikova L, Vesely V. Influence of the elastic mismatch on crack propagation in a silicate-based composite. *Theoretical and Applied Fracture Mechanics*, 2017.
- [25] Malikova L, Vesely V, Seitl S. Estimation of the crack propagation direction in a mixed-mode geometry via multi-parameter fracture criteria. *Frattura ed Integrita Strutturale* 2015; 33: 25–32.
- [26] Akbaridoost J, Rastin A. Comprehensive data for calculating the higher order terms of crack tip stress field in disk type specimens under mixed-mode loading. *Theoretical and Applied Fracture Mechanics* 2015; 76: 75–90.
- [27] Hello G, Tahar M-B. On the exactness of truncated crack-tip stress expansions. *Procedia Materials Science* 2014; 3: 750–755.
- [28] Holynskiy IS. Influence of errors of determination of stresses near a crack tip on the accuracy of computation of the coefficients of the Williams series under Mode II loading. *Materials Science* 2013; 48(4): 438–443.
- [29] Stepanova LV, Adylina EM. Stress-strain state in the vicinity of a crack tip under mixed loading. *Journal of Applied Mechanics and Technical Physics* 2014; 55(5): 885–895.
- [30] Lychak O, Holynskiy IS. Evaluation of random errors in Williams' series coefficients obtained with digital image correlation. *Measurement Science and Technology* 2016; 27(3): 035203.
- [31] Stepanova LV, Igonin SA. Rabotnov damage parameter and description of delayed fracture: Results, current status, application to fracture mechanics, and prospects. *Journal of Applied Mechanics and Technical Physics* 2015; 56(2): 282–292.
- [32] Berto F, Lazzarin P. Multiparametric full-field representations of the in-plane stress fields ahead of cracked components under mixed mode loading. *International Journal of Fatigue* 2013; 46: 16–26.
- [33] Stepanova LV, Yakovleva EM. Asymptotic stress field in the vicinity of a mixed-mode crack under plane stress conditions for a power-law hardening material. *Journal of Mechanics of Materials and Structures* 2015; 10(3): 367–393.
- [34] Tada H, Paris PC, Irwin GR. *The stress analysis of cracks*. Handbook, ASME Press, New York, 2000.

# Methods of bipolar microcircuits learning experiment

S.V. Tyulevin<sup>1</sup>, M.N. Piganov<sup>2</sup>, E.S. Erantseva<sup>2</sup>

<sup>1</sup>JSC Progress Rocket and Space Centre "Progress", Zemets St., 18, 443009, Samara, Russia

<sup>2</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The analysis of learning experiment methods of semiconductor microcircuits for individual prediction of their quality is carried out. A choice of the informative parameters and means of their control is made. The schemes of insertion the microcircuits in the course of investigation tests and modes of their control are justified. The analysis of microcircuits constructive and technological options is carried out. The program of investigation tests is developed. Results of learning experiment for 522 series microcircuits are given. The analysis of experimental data is carried out. It is recommended to use results of learning experiment for creation the mathematical prediction models of microcircuits quality.

*Keywords:* learning experiment; bipolar microcircuit; method; investigation tests control; prediction model

---

## 1. Introduction

The industry way out from crisis in the case of open market economy is almost impossible without solution the problem of improvement the quality and competitiveness of products. The problem of improvement the quality is particularly acute, first of all, before the knowledge-intensive branches of engineering to which also the microelectronics belongs. The main products of microelectronics is integrated microcircuits (IMS). It is taking into account that quality of IMS is defined by their construction, the initial materials, complexity and stability of technological processes. At the same time the main link is the manufacturing technology [1].

One of the conceptual principles of microcircuit quality management can become principles of open quality management. At the same time it is expedient to select the following contours of formation the quality: quality establishment; quality support; quality maintenance; quality prediction; guaranteeing quality; quality improvement [2].

Thus, one of the main stages in formation the microcircuits quality is prediction of their quality indices. For IMS, using in the responsible equipment, the most effective is the individual prediction [3, 4]. The most important stage of the individual prediction (IP) is the learning experiment.

The learning experiment is a test in the given mode of a certain quantity of the researched products during the required time, usually equal time of the subsequent prediction of  $t_{pr}$ , and determination the actual state of each specimen of selection to the time of the test end. The purpose of learning experiment consists in receiving the necessary array of initial data, i.e. such array which is required for the subsequent training. The maintenance of initial data array is defined by a type of IP. For example, values of the informative  $x_i$  parameters (signs) and the predicted  $y_0$  parameter for all specimen in the initial timepoint, values of the predicted parameter in finite timepoint of  $y_k$ , i.e. in case of  $t = t_{pr}$ , the intermediate values of the predicted parameter. Sometimes it is required to know the intermediate and finite values of the informative parameters.

Basis of a learning experiment are investigation tests. They allow to reveal, except obtaining the above-mentioned information, processes and schemes of elements degradation, to set types and mechanisms of failures, types of defects, load ranges which accelerate failures. It allows to set up a level of signs informtiveness, criteria of rejection and classification for each constructive and technological option (CTO), to select the most informative parameters, and also to define the modes of technological tests, to optimize the researched CTO, to improve the methods of carrying out the investigation tests, learning experiment and rejection [5].

The most difficult question is determination the types and modes of test influences. They depend on a risk degree or a measure of damage caused in the case of equipment operation. The great problems arise in case of choice the informative parameters, methods and control means.

Imperfection of a stage of learning experiment carries to lowering the accuracy of prediction model, increase in the conditional risks of the supplier and a customer. The work purpose – is a choice of informative parameters and schemes of insertion the 522 series microcircuit when carrying out a learning experiment.

## 2. The analysis of learning experiment methods

The common method of learning experiment in the case of individual prediction the quality indices of space radio-electronic means (REM) is given in [6]. It includes seven main stages:

1. Analysis of constructive and technological features of electric radio products (ERP) and REM.
2. Development or choice the schemes of insertion for control their working capacity and measurement the key parameters.
3. Choice of methods and control means and informative parameters.
4. Determine the selection volume.
5. Development the program of investigation tests.
6. Carrying out investigation tests and experiments.
7. Analysis the test and experiments results.

This method was approved on 286 series microcircuit and showed good results.

Authors [5] within this common approach offered the particular method of carrying out the learning experiment for CMOS type microcircuits. This method provides control of mismatch pulse duration.

For control the informative parameters the installation, containing a dialup field, 2 square waveform oscillator, adapters, comparator, averaging circuit, indication device is offered. As the informative parameter for IMS rejection the propagation delay

time of a signal in case of turning power on and off is used. At the same time on testing microcircuit they give supply voltage near-critical.

In a number of works the types of test influences when carrying out the diagnostic predicting check are specified.

Many works are devoted to choice and analysis of informative parameters. So, in [7] for prediction of chips quality it is offered to use the level of internal stresses. It is shown that the level of internal stresses depends on the operation modes of diffusion, epitaxy, oxidation, etc. In [8] it was used the distributions of thermal and physical parameters of high-power transistors. Authors in [9] as the informative parameters of semiconductor products on a plate use volt-ampere characteristics (VAC), volt-farad characteristic (VFC) or ampere-noise characteristic (ANC). At the same time they estimate charge stability in case of corona discharge influence. Information of semiconductor devices quality is performed also by m-parameter, which characterizes VAC not ideality level. It was used in case of prediction the transistors durability by image identification methods [10]. High informativeness in case of electrophysical diagnosing of CMOS types microcircuits was shown by critical power voltage [11]. For quality control of digital integrated microcircuits assembly it is expedient to use matrix parameters of thermal communication [12]. Authors [13] for quality control of light-emitting diodes and semiconductor lasers suggest to use thermomechanical stresses. They arise both as a result of change the ambient temperature, and as a result of sharply heterogeneous self-heating of the instrumental structures by dissipation power. In [14] it is set that it is possible to estimate the semiconductor devices reliability by the value of mechanical stresses. They arise because of distinction the thermal extension coefficients of the applied materials. At the same time the concentration of minority carriers of a charge, their mobility and lifetime change, the energy levels displace.

To predict the drift of semiconductor devices parameters in time and their durability alloys m-parameter [15].

For a kind of transistors with a small area of emitter after passage the several pulses of current through direct switches emitter junction in the active standard mode there can be considerable leakage currents of emitter junction, increasing basis current in the micromode and reducing the current amplification factor [16]. On value of a leakage current they estimate quality of a product.

Authors [17] suggest to estimate CMOS type microcircuits quality on value of critical power voltage after influence of electric discharge. Rather informative parameter for many types of semiconductor devices quality is low-quality noise.

### 3. The analysis of the researched microcircuits CTV

For research and analysis of control and testing processes of 522 series microcircuits were selected, as at them failures were watched earlier. Key parameters of 522KH microcircuits are specified below.

Residual voltage at the output of microcircuit, with  $I=I_{out.max.}$ ,

$U_{res.}, V$	0,8
Failure voltage, $U_{fail.}, V$ :	
Input #4 - #3	0,35
Input #4 - #2	0,7
Disruptive voltage across power circuits, $U_{disr.}, V$	47
Current consumption in closed condition, $I_{cons.}, \mu A$	75
Coefficient of return, K	
no more	0,35
no less	0,85
Supply voltage, $U_{sv}, V$	36
Output amperage ( $-60...+85^{\circ}C$ ), $I_{out.}, mA$	120
The power dissipated in the chip package ( $-60^{\circ}C...+25^{\circ}C$ ), $P_{diss.}, W$	0,4
Active load value, $P_{load.}, \Omega$ , no less	280
Inductive load value, $L_{load.}, H$ , no more	0,22
Maximum permissible voltage between terminals, $U_{max.}, V$	
1-3, 14-3, 1-2, 14-2,	4
4-3, 2-3, 2-1, 2-14	36
Residual voltage at the output of microcircuit in the absence of loading, V	0,3
Maximum switching time, $\mu s$	100
Maximum operating frequency, Hz	100
Resistance of circuit in the open state (with $I=I_{out.max.}$ ), $\Omega$	10
Emission amplitude at the top of the output pulse, U, V	3

The researches carried out by the authors showed the expediency of using the residual voltage of the microcircuits as an informative parameter.

### 4. The switching on schemes

For measuring the electrical parameters the switching on schemes, adducing on fig. 1-5, were offered. At measuring the  $U_{srb}$  and K parameters the supply voltage is applied to the chip before submission the signals on the microcircuit inputs, switch-off is made upside-down. In remaining cases the signals first of all are applied on the microcircuit inputs, then – supply voltages as their increase are given, switch-off is made upside-down.

In the case of measuring the electrical parameters and testing the microcircuits it is allowed the simultaneous submission and switch-off supply voltage and signals on inputs and outputs of microcircuits. An error of setting up the test voltage and supply voltages must not exceed  $\pm 2\%$ , test currents –  $\pm 4\%$ .

Measuring instruments of a direct current and voltage shall have an accuracy class not worse than 1,0. The accuracy class of voltmeters – indicators of voltage level meters isn't regulated. Checking the values of supply voltage is made by the accuracy class voltmeter not worse 1,0 with input resistance at least  $10k\Omega / V$ .

Metrological aspects of control are given in [18].

Measuring the residual voltage on the output of integrated microcircuit in the case of  $I=I_{out.max.}$  is carried out according to the measurements scheme provided on fig. 1.

Value of the output current of  $I=I_{out.max.}$  is set ( $R_3$  resistor) by active component of equivalent loading and power supply  $E_2$ . In case of connection the output  $I_3$  over the  $R_2$  resistor with  $E_2$  minus, the microcircuit connects loading to the power supply, through  $R_3$  run current of  $I=I_{out.max.}$ , causing existence of residual voltage  $U_{res}$ .

Value of  $U_{res}$  is controlled by the PV voltmeter.

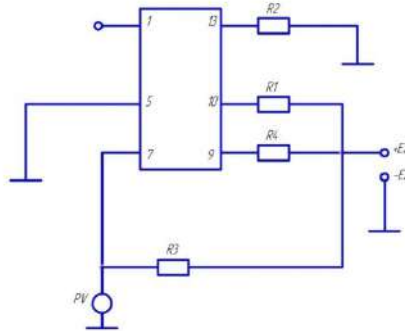


Fig. 1. Scheme of measurement the residual voltage.

Measurement of the operating voltage  $U_{oper}$  is carried out according to the scheme shown on fig. 2.

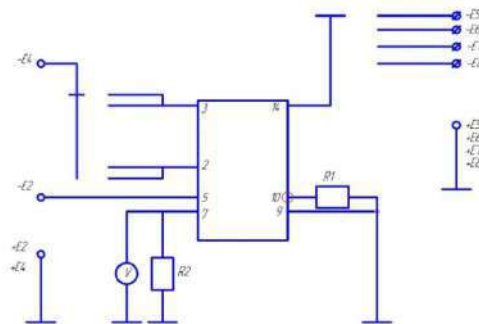


Fig. 2. Scheme of measurement the operating voltage.

The operation voltage  $U_{oper}$  is called the minimum voltage, applied to the microcircuit input, that calls the connection of load to the power supply.

The microcircuit has two control inputs: input # 4 - # 3 and input # 4 - # 2.

Value of voltage  $U_{oper}$  is set discretely by power supplies  $E_4 - E_8$ . The presence or absence of the loading connection to the power supply registers the PV voltmeter – indicator of voltage level.

Measurement of the failure voltage  $U_{fail}$  are carried out according to the same scheme of measurement (fig. 2).

The failure voltage  $U_{fail}$  is called the maximum voltage applies to the microcircuit input which doesn't yet cause connection of loading to the power supply.

Value of voltage  $U_{fail}$  is set discretely by power supplies  $E_4 - E_8$ .

In the voltage range between  $U_{oper}$  and  $U_{fail}$  for the same control input, the microcircuit can either connect the load to the power supply or not connect it.

Measuring the consuming current in the closed status  $I_{cons.}$  of the microcircuit and disruptive voltage in the supply circuits is carried out simultaneously according to the measurement scheme shown in Fig. 3.

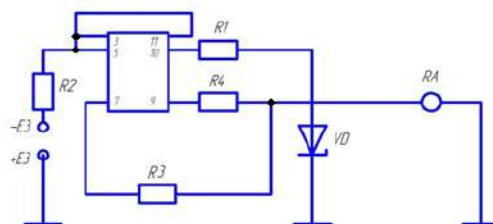


Fig. 3. Scheme of measurement the consumption current and disruptive voltage.

Consumption current in the close state is the current flowing through the circuits of the microcircuit power supply, when the voltage is not supplied to the load by the microcircuit.

The disruptive voltage through the supply circuits is the maximum voltage applied to the microchip over the power circuits, which does not yet cause the load to be connected to the power supply without input signal.

Current consumption  $I_{\text{cons.}}$  is controlled by microammeter RA, disruptive voltage  $U_{\text{disr}}$  is set by the power supply  $E_3$ . If the microcircuit does not comply with technical requirements for the disruptive voltage  $U_{\text{disr}}$ , the current consumption  $I_{\text{cons.}}$  will be more than values specified in this requirements.

Control of the operation at maximum operating frequency of 100 Hz is carried out according to the control scheme shown in Fig. 4.

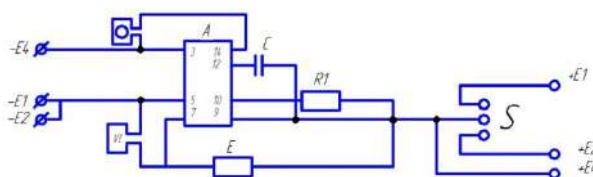


Fig. 4. Scheme for control the microcircuit operation.

Test for faultless is carried out by method 700-1 OST 11 073.013-83 at the temperature + 850C. The scheme of switching on in case of tests is provided on fig. 5.

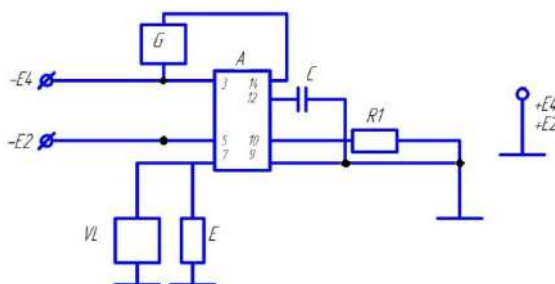


Fig. 5. Scheme of microcircuits testing for faultless and durability.

The holding time in normal conditions before measurement the parameters – 2 h.  
Some questions of increasing the efficiency of ERP control are given in [19-24].

## 5. Conclusion

The made analysis revealed a row of problems when carrying out a learning experiment with bipolar microcircuits: absence of the approved schemes of switching on when carrying out investigation tests and low informativeness of parameters by development the expected models. For 522 series microcircuits as the informative parameter it is recommended to use the value of residual voltage. Schemes of measurement of residual voltage, actuation voltages, consuming current, operating control, test for faultless are offered.

## References

- [1] Piganov MN. Technological bases of support the microassemblies quality. Samara: SGAU, 1999; 231 p.
- [2] Piganov MN. Individual prediction the quality indices of elements and components of microassemblies. M.: New technologies, 2002; 267 p.
- [3] Zhadnov VV. Prediction of reliability the electronic means with mechanical elements. Ekaterinburg: LLC Fort Dialog, 2014; 172 p.
- [4] Keydzhyan GA. Prediction the reliability of microelectronic equipment on the basis of the large-scale integrated circuit. M.: Radio and communication, 1987; 152 p.
- [5] Tyulevin SV, Piganov MN, Erantseva ES. To the problem of prediction the quality indices space equipment elements. Reliability and quality of multiple systems 2014; 1(5): 9–17.
- [6] Tyulevin SV, Piganov MN. Methods of learning experiment in the case of individual prediction the quality indices of space RES. Actual problems of radioelectronics and telecommunications: mater. of all-russian STC. Samara: SGAU, 2008; 239–253.
- [7] Berenstein GV, Dyachenko AM. Prediction of IC quality on the basis of internal voltages analysis. Physical bases of reliability and degradation of semiconductor devices: theses of report at All-Union conf. Chisinau 1991; II: 36.
- [8] Sergeyev VA. Characteristics and features of selective distributions of powerful bipolar transistors on thermophysical parameters. News of the Samara scientific center of RAS 2004; 1: 154–160.
- [9] Gorlov MI, Zharkikh AP. Device for control the charge stability of semiconductor products using corona discharge. Russian Federation Patent 2312424. Publ. 10.12.2007. Bulletin No. 34.
- [10] Luchino AI, Savin AS. Researches the possibility of individual prediction the durability of transistors by image identification. Electronic engineering 1976; 8(10): 3–9.
- [11] Aladinskiy VK, Gavrilov VYu, Gorelkina EN. The critical supply voltage as the informative parameter in case of CMOS IC electrophysical diagnosing. Electronic engineering 1990; 2(4): 87–90.
- [12] Sergeyev VA, Yudin VV. Quality control of digital integrated microcircuits by parameters of the thermal communications matrix. News of higher education institutions. Electronics 2009; 6: 72–78.
- [13] Chang MH, Das D, Varde PV, Pecht M. Light emitting diodes reliability review. Microelectronics Reliability 2012; 5: 762–782.
- [14] Kuba J. Application of low temperature infailure diagnostics of semiconductor devices. Power Semic. Hybrid Device – 8-th Int. Spring Semin. Electrotechnol. Prenet 1985; 31–34.
- [15] Kleshko VM, Semenov AS. Quality control and reliability of semiconductor devices using m-characteristics. Electronic equipment 1974; 8(12): 17–21.
- [16] Watchik R, Bucelot T, Li G. J. Appl. Phys. 1998; 9: 4734–4740.

- [17] Gorlov MI, Vinokurov AA. The influence of electrostatic discharges on the critical power supply voltage of K561LNZ type. Solid state electronics, microelectronics and nanoelectronics: collection of scientific papers. Voronezh, 2011; 10: 96–98.
- [18] Piganov MN. Metrological aspects of microassemblies quality assurance. Modern information and electronic technologies: Proceedings of the 3rd international scientific practical conference. Ukraine, Odessa, 2002: 140.
- [19] Mishanov R, Piganov M. Individual forecasting of quality characteristics by an extrapolation method for the stabilitrons and the integrated circuits. The Experience of Designing and Application of CAD Systems in Microelectronics: proceeding XIII international conference. Ukraine, Lviv, 2015: 242–244.
- [20] Piganov M, Tyulevin S, Erantseva E. Individual prognosis of quality indicators of space equipment elements. The Experience of Designing and Application of CAD Systems in Microelectronics: proceeding XIII international conference. Ukraine, Lviv, 2015: 367–371.
- [21] Piganov MN, Tyulevin SV, Erantseva ES, Mishanov RO. Apparatus diagnostic for non-destructive control chip CMOS-type. European Science and Technology: materials of the VII international research and practice conference. Germany, Munich, 2014: 398–401.
- [22] Jonson JB. The Sholiky effect in low frequency circuits. Phys. Rev. 1925; 26: 71–85.
- [23] Mishanov RO. The installation of diagnostic non-destructive control for the bipolar IC. Science and Education: materials of the VII international research and practice conference. Germany, Munich, 2014: 227–232.
- [24] Mishanov RO, Piganov MN. Technology of diagnostic for non-destructive control of the bipolar integrated circuits. Sense. Enable. Spitse: proceedings 2-nd international scientific symposium. Russia, St. Petersburg, 2015: 38–41.

# Frames and subspaces for phaseless reconstruction

S.Ya. Novikov,<sup>1</sup> M.E. Fedina<sup>1</sup>

<sup>1</sup>Samara National Research University, Moskovskoe shosse, 34, Samara, 443086, Russia

---

## Abstract

Frames and subspaces, that are used to the reconstruction of the vector signal without phase measurements, represented. The new concept of equidistributed frames is considered. The possibility of reconstruction of the vector by the norms of the projections on the subspaces is asserted. Particular attention is paid to systems of subspaces for which there is the possibility of reconstruction by the norms of the projections on them and on their orthogonal complements.

*Keywords:* equidistributed frames; phaseless reconstruction; complement property; full spark set; norm retrieval

---

## 1. Basic facts. Phaseless recovery

Let  $\mathbb{H}^M$  denotes  $M$ -dimensional space with the scalar product.

**Definition 1.** A set of vectors  $\Phi = \{\varphi_k\}_{k=1}^N$  is called a *frame* for the  $\mathbb{H}^M$ , if there are positive constants  $A, B$  such that for all  $x \in \mathbb{H}^M$

$$A\|x\|^2 \leq \sum_{k=1}^N |\langle x, \varphi_k \rangle|^2 \leq B\|x\|^2.$$

Numbers  $A$  and  $B$  are called the lower and upper frame bounds respectively. If we can choose  $A = B$ , then the frame is called *tight*, and if  $A = B = 1$ , it is called a *Parseval-Steklov frame* (This name was proposed by Acad. V.S. Vladimirov during a report of the second author in Math. Steklov Institute in 2008 instead of usual Parseval Frame).

Note that in the finite dimensional setting, a frame is simply a spanning set of vectors in the Hilbert space ( $\text{span}\{\varphi_k\}_{k=1}^N = \mathbb{H}^M$ ) [1, 2].

There are three operators connected with a frame  $\Phi$ :

*analysis operator*  $T : \mathbb{H}^M \rightarrow \ell_2^N$ , defined by

$$T(x) = \{\langle x, \varphi_k \rangle\}_{k=1}^N,$$

*adjoint synthesis operator*

$$T^* \left( \{a_k\}_{k=1}^N \right) = \sum_{k=1}^N a_k \varphi_k$$

and *frame operator*  $S := T^*T$  on  $\mathbb{H}^M$ , defined by

$$S(x) = T^*T(x) = \sum_{k=1}^N \langle x, \varphi_k \rangle \varphi_k.$$

The frame operator is positive, self-adjoint and invertible. Besides, we have

$$AI \leq S \leq BI,$$

where  $I$  is identity operator in  $\mathbb{H}^M$ .

In particular, for the Parseval-Steklov frame the frame operator is the identity operator, so this frame is the most useful for the reconstruction of signals. In fact, in this case for every  $x \in \mathbb{H}^M$  the following equality is true

$$x = \sum_{k=1}^N \langle x, \varphi_k \rangle \varphi_k.$$

The operator  $G = TT^*$  is Gram operator with the matrix

$$\begin{pmatrix} \|\varphi_1\|^2 & \langle \varphi_2, \varphi_1 \rangle, & \dots & \langle \varphi_N, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_2 \rangle & \|\varphi_2\|^2 & \dots & \langle \varphi_N, \varphi_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_1, \varphi_N \rangle & \langle \varphi_2, \varphi_N \rangle & \dots & \|\varphi_N\|^2 \end{pmatrix}$$

and for the Parseval-Steklov frame coincides with the projection  $P : \ell_N^2 \rightarrow \ell_N^2$  to the image of the analysis operator [1, 2].

An easy way is known to construct Parseval-Steklov frames. It is based on the following proposition.

**Proposition 1.** *Let  $\{\varphi_k\}_{k=1}^N$  be a frame for  $\mathbb{H}^M$  with bounds  $A$  and  $B$ , and let  $P$  be the orthogonal projection in  $\mathbb{H}^M$  on the subspace  $W$ . Then  $\{P\varphi_k\}_{k=1}^N$  is a frame for  $W$  with bounds  $A$  and  $B$ . In particular, if  $\{\varphi_k\}_{k=1}^N$  is Parseval-Steklov frame for  $\mathbb{H}^M$  and  $P$  is the orthogonal projection on  $W$ , then  $\{P\varphi_k\}_{k=1}^N$  is Parseval-Steklov frame for  $W$ .*

Proof. We have for  $x \in W$

$$\begin{aligned} A\|x\|^2 &= A\|Px\|^2 \leq \sum_{k=1}^N |\langle Px, \varphi_k \rangle|^2 = \\ &= \sum_{k=1}^N |\langle x, P\varphi_k \rangle|^2 \leq B\|Px\|^2 = B\|x\|^2. \end{aligned}$$

**Corollary 1.** *Let  $\{e_k\}_{k=1}^M$  be an orthonormal basis (ONB) in  $\mathbb{H}^M$ , and let  $P$  be the orthogonal projection on the subspace  $W$ . Then  $\{Pe_k\}_{k=1}^M$  is Parseval-Steklov frame for  $W$ .*

Corollary 1 is the foundation of the following algorithm for construction of Parseval-Steklov frame. We construct  $N \times N$  unitary matrix for  $N \geq M$ , then we choose any  $M$  rows, columns of thus obtaining  $M \times N$ -matrix form Parseval-Steklov frame in  $\mathbb{H}^M$ . If we construct from the remaining  $N - M$  rows  $(N - M) \times N$ -matrix, then its columns are Parseval-Steklov frame in  $\mathbb{H}^{N-M}$ .

The following theorem, actually proved by Naimark, shows that such process is essentially the only one for constructing Parseval-Steklov frame [3].

**Theorem 1.**

*Let  $\Phi = \{\varphi_k\}_{k=1}^N$  be a frame in  $\mathbb{H}^M$  with the analysis operator  $T$ , let  $\{e_k\}_{k=1}^N$  be the standard basis in  $\ell_N^2$ , let  $P : \ell_N^2 \rightarrow \ell_N^2$  be the orthogonal projection on  $\text{Im}(T)$ .*

*The following assertions are equivalent:*

1.  $\Phi$  is Parseval-Steklov frame for  $\mathbb{H}^M$ .
2. For all  $k = 1, \dots, N$  we have  $Pe_k = T\varphi_k$ .
3. There are vectors  $\{\psi_k\}_{k=1}^N \subset \mathbb{H}^{N-M}$  such that  $\{\varphi_k \oplus \psi_k\}_{k=1}^N$  form ONB in  $\mathbb{H}^N$ .

*Besides,  $\{\psi_k\}_{k=1}^N$  are Parseval-Steklov frame in  $\mathbb{H}^{N-M}$ .*

Proof.

(1)  $\Leftrightarrow$  (2). As noted, the system  $\{\varphi_k\}_{k=1}^N$  forms Parseval-Steklov frame iff Gram operator  $TT^*$  coincides with the projection  $P$ . So (1) and (2) are equivalent according to equality  $T^*e_k = \varphi_k$  for  $k = 1, \dots, N$ .

(1)  $\Rightarrow$  (3). Let's put  $d_k = e_k - T\varphi_k$ ,  $k = 1, \dots, N$ . According to (2),  $d_k \in (\text{Im}(T))^\perp$  for all  $k$ . For a unitary operator

$$\Phi : (\text{Im}(T))^\perp \rightarrow \mathbb{H}^{N-M}$$

let's put

$$\psi_k := \Phi d_k, k = 1, \dots, N.$$

We have using the isometry of the operator  $T$ ,

$$\langle \varphi_i \oplus \psi_i, \varphi_k \oplus \psi_k \rangle = \langle \varphi_i, \varphi_k \rangle + \langle \psi_i, \psi_k \rangle =$$



$$= \langle T\varphi_i, T\varphi_k \rangle + \langle d_i, d_k \rangle = \delta_{ik}.$$

(3)  $\Rightarrow$  (1). Let's apply corollary 1.

As in [4], we call vectors  $\{\psi_k\}_{k=1}^N$  *Naimark complement of the frame  $\Phi$* .

For Parseval-Steklov frame, written as  $\{Pe_k\}_{k=1}^N$ , Naimark complement is the system of vectors  $\{(I - P)e_k\}_{k=1}^N$ .

Naimark complements are defined only for Parseval-Steklov frames, and they are defined up to unitary equivalence. If  $\{\varphi_k\}_{k=1}^N \subset \mathbb{H}^M$  and  $\{\psi_k\}_{k=1}^N \subset \mathbb{H}^{N-M}$  complement each other,  $U$  and  $V$  are unitary operators ( $U^*U = UU^* = I$ ), then  $\{U\varphi_k\}_{k=1}^N, \{V\psi_k\}_{k=1}^N$  also complement each other.

An important application of frames is the reconstruction of a signal with incomplete data. In particular, much attention is attracted to the problem of the reconstruction phase information. In recent papers on this topic two aspects of the problem were emphasized: phaseless reconstruction and phase retrieval [7]. This paper focuses on the first aspect.

**Definition 2.** The set of vectors  $\Phi = \{\varphi_i\}_{i=1}^N$  in  $\mathbb{R}^M$  (or  $\mathbb{C}^M$ ) provides *phaseless reconstruction (PLR)*, if equalities of measurement modules

$$|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|, \quad x, y \in \mathbb{R}^M \text{ (} \mathbb{C}^M \text{)}, \quad i = 1, \dots, N,$$

imply the equality of vectors-signals up to unimodular factor, i.e.  $x = cy$  with some  $c = \pm 1$  for  $\mathbb{R}^M$  or  $c \in \mathbf{T}$  for  $\mathbb{C}^M$ , where  $\mathbf{T}$  is the unit circle in  $\mathbb{C}$ .

In the rest of the text sets, which are satisfied the definition of 2, is called PLR-systems or PLR-sets. The next property is important in these questions.

**Definition 3** [4, 5]. The set  $\Phi = \{\varphi_n\}_{n=1}^N$  in  $\mathbb{H}^M$  has *complement property (CP)*, if for any  $S \subseteq \{1, \dots, N\}$   $\{\varphi_n\}_{n \in S}$  or  $\{\varphi_n\}_{n \in S^c}$  is complete in  $\mathbb{H}^M$ . Complement property in  $\mathbb{R}^M$  is equivalent to PLR (theorem 2 below).

**Definition 4** [4, 5, 6]. The *spark* of the set  $\Phi = \{\varphi_n\}_{n=1}^N \subset \mathbb{H}^M$  is the cardinality of the smallest linear dependent subset of  $\Phi$ . If  $\text{spark}(\Phi) = M + 1$ , then any subset with  $M$  vectors linear independent, in this case  $\Phi$  is called *full spark set*.

In earlier works the term "girth" was used instead of the term "spark". Spark of the linear independent system, for example, basic, is assumed to be zero.

**Theorem 2** [5, 8].

*Frame  $\{\varphi_n\}_{n=1}^N$  in  $\mathbb{R}^M$  is the PLR-system iff it has complement property. In particular, full spark frame with at least  $2M - 1$  vectors is PLR-system. If  $\{\varphi_n\}_{n=1}^N$  is PLR-system in  $\mathbb{R}^M$ , then  $N \geq 2M - 1$ , any subset with  $2M - 2$  vectors can't be PLR-system.*

Generally speaking, the recovery without phases is possible not only by full spark frames. Each frame, containing  $(2M - 1)$  full spark frame, will also provide recovery without phases. However, if the frame contains exactly  $2M - 1$  elements, it is a PLR-system only for full spark frame [5, 8].

If  $\Phi$  is the Parseval-Steklov frame for  $\mathbb{H}^M$  with  $N$  elements, the analysis operator is isometric according to

$$\|Tx\|^2 = \sum_{n=1}^N |\langle x, \varphi_n \rangle|^2 = \|x\|^2, \quad x \in \mathbb{H}^M.$$

In this case, we obtain the reconstruction identity  $x = \sum_{n=1}^N |\langle x, \varphi_n \rangle| \varphi_n$ , or  $x = T^*Tx$ . In this case, we also have that the Gramian  $G := TT^*$  is a rank- $M$  orthogonal projection, because  $G^*G = TT^*TT^* = TT^* = G$  and the rank of  $G$  equals the trace,  $\text{tr}G = M$ .

**Definition 5.** Two frames  $\Phi = \{\varphi_n\}_{n=1}^N$  and  $\Phi' = \{\varphi'_n\}_{n=1}^N$  for a finite dimensional space  $\mathbb{H}^M$  are called *unitarily equivalent* if there exists an orthogonal or unitary operator  $U$  on  $\mathbb{H}^M$  such that  $\varphi_n = U\varphi'_n$  for  $n = 1, \dots, N$ .

Each equivalence class of frames is characterized by the corresponding Gram matrix.

**Proposition 2** [13]. *The Gramians of two frames  $\Phi = \{\varphi_n\}_{n=1}^N$  and  $\Phi' = \{\varphi'_n\}_{n=1}^N$  for  $\mathbb{H}^M$  are identical if and only if the frames are unitarily equivalent.*

In [13] the new class of frames is introduced.

**Definition 6.** Let  $\Phi = \{\varphi_n\}_{n=1}^N$  be Parseval-Steklov frame for  $\mathbb{H}^M$ , let  $G$  be its Gramian. The frame  $\Phi$  is called *equidistributed* if for each pair  $p, q \in \mathbb{Z}_N$  there exists a permutation  $\pi$  on  $\mathbb{Z}_N$  such that  $|G_{j,p}| = |G_{\pi(j),q}|$  for all  $j \in \mathbb{Z}_N$ .

In other words,  $\Phi$  is equidistributed if and only if the magnitudes in any column of the Gram matrix repeat in any other column, up to a permutation of their position.

**Proposition 3.** *If  $\Phi = \{\varphi_n\}_{n=1}^N$  is an equidistributed Parseval-Steklov frame in  $\mathbb{H}^M$ , then  $\|\varphi_n\|^2 = M/N, n = 1, 2, \dots, N$ .*

Proof. By assumption, for each  $n$  there exists  $\pi$  such that  $|G_{n,p}| = |G_{\pi(n),1}|$  holds for the entries of the associated Gram matrix  $G$  for all  $n$ . By the Parseval-Steklov identity

$$\|\varphi_p\|^2 = \sum_{n=1}^N |\langle \varphi_p, \varphi_n \rangle|^2 = \sum_{n=1}^N |G_{n,p}|^2 = \sum_{n=1}^N |G_{\pi(n),1}|^2 = \|\varphi_1\|^2.$$

The trace condition  $\sum_{n=1}^N G_{n,n} = \sum_{n=1}^N \|\varphi_p\|^2 = M$  for the Gram matrices of Parseval-Steklov frames implies that  $\|\varphi_n\|^2 = M/N, n = 1, 2, \dots, N$ .

**Examples:**

1. *Equiangular Parseval-Steklov frames.*

Let  $\Phi = \{\varphi_n\}_{n=1}^N$  be an equal-norm frame and there exists  $C \geq 0$  such that  $|\langle \varphi_n, \varphi_{n'} \rangle| = C$  for all  $n, n' \in \mathbb{Z}_N$  with  $n \neq n'$ . Such frames are called equiangular. Such Parseval-Steklov frames exist only with some restrictions on  $N$  and  $M$  [13]. The simplest example of the equiangular Parseval-Steklov frame in  $\mathbb{R}^2$  is a well-known "Mercedes-Benz frame".

Magnitudes of the entries of any column of  $G$  for such frame consist of  $N - 1$  instances of  $C$  and one instance of  $M/N$ , so  $\Phi$  is equidistributed.

2. *Mutually unbiased bases.*

Such frame is union of orthonormal bases such that the modulus of the inner product between any two vectors from distinct bases is constant. Such examples are widely used in quantum information theory. The simplest example of mutually unbiased bases in  $\mathbb{C}^2$  is given by the following three bases:

$$M_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad M_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad M_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}.$$

To get the Parseval-Steklov frames one should renorm vectors to  $\Phi$  because of these 3 matrices must be multiplied to  $1/\sqrt{3}$ . We get the Gram matrix

$$G = \begin{pmatrix} \frac{1}{3} & 0 & \lambda & \lambda & \lambda & \lambda \\ 0 & \frac{1}{3} & \lambda & -\lambda & -i\lambda & i\lambda \\ \lambda & \lambda & \frac{1}{3} & 0 & \frac{1-i}{6} & \frac{1+i}{6} \\ \lambda & -\lambda & 0 & \frac{1}{3} & \frac{1+i}{6} & \frac{1-i}{6} \\ \lambda & i\lambda & \frac{1+i}{6} & \frac{1-i}{6} & \frac{1}{3} & 0 \\ \lambda & -i\lambda & \frac{1-i}{6} & \frac{1+i}{6} & 0 & \frac{1}{3} \end{pmatrix},$$

where  $\lambda = \sqrt{2}/6$ .

3. *Group frames.*

Let  $\Gamma$  be a finite group of size  $N = |\Gamma|$  and  $\pi : \Gamma \rightarrow B(\mathbb{H}^M)$  be an orthogonal or unitary representation of  $\Gamma$  on the real or complex space  $\mathbb{H}^M$  respectively.

The orbit  $\Phi = \{f_g = \pi(g)f_e\}_{g \in \Gamma}$ , generated by a vector  $f_e$  of norm  $\sqrt{N/M}$ , indexed by the unit  $e$  of the group, forms the Parseval-Steklov frame, if the representation is irreducible [14]. In this case  $\Phi$  is equidistributed, because  $\langle f_g, f_h \rangle = \langle \pi(h^{-1}g)f_e, f_e \rangle$ , and left multiplication  $h, h^{-1}$  acts as a permutation on the group elements. So the entries of Gram matrix has equal modules, up to a permutation in rows (columns).

4. *Cycle frames.*

Consider the Discrete Fourier Transform matrix

$$F = \frac{1}{\sqrt{N}} (\omega^{jl})_{j,l=1}^N, \quad \text{where } \omega = e^{\frac{2\pi i}{N}}.$$

Its columns form an orthonormal basis for  $\mathbb{C}^N$ . If  $A$  is a  $M \times N$  matrix obtained by deleting any choice of  $N - M$  rows from  $F$ , then its columns form a Parseval-Steklov frame for  $\mathbb{C}^M$ .

**Definition 7.** Let  $b_1, \dots, b_M \in \{1, 2, \dots, N\}$  be any choice of distinct integers. F frame  $\Phi = \{\varphi_n\}_n \in \mathbb{Z}_N$ , where

$$\varphi_n = \frac{1}{\sqrt{N}} (\omega^{nm})_{m=1}^N, \quad \text{for all } n \in \mathbb{Z}_N,$$

is called a *cycle frame*.

Every cycle frame is equidistributed, and, because the construction described above works for every pair of positive integers  $M$  and  $N$  with  $M < N$ , the existence of equidistributed frames is ensured in the complex setting.

**Theorem 3.**

For every  $N > M$  there exists equidistributed Parseval-Steklov frame in  $\mathbb{C}^M$  with  $N$  vectors.

Let's see if the possibility of recovery without phases is transferred to the Naimark complements. We require the following theorem for this.

**Theorem 4 [9].**

Let  $P$  be an projection in  $\mathbb{H}^N$  with ONB  $\{e_n\}_{n=1}^N$  and  $S \subset \{1, 2, \dots, N\}$ .

The following assertions are equivalent:

1.  $\{Pe_i\}_{i \in S}$  linear independent.
2.  $\text{span}\{(I - P)e_i\}_{i \in S^c} = (I - P)(\mathbb{H}^N)$ .

Proof.

(1)  $\Rightarrow$  (2). Let's suppose, that

$$\text{span}\{(I - P)e_i\}_{i \in S^c} \neq (I - P)(\mathbb{H}^N).$$

It means, that there exists  $0 \neq x \in (I - P)(\mathbb{H}^N)$  such that  $x \perp \text{span}\{(I - P)e_i\}_{i \in S^c}$ . As  $x = \sum_{i=1}^N \langle x, e_i \rangle (I - P)e_i$ , then

$$\langle x, (I - P)e_i \rangle = \langle (I - P)x, e_i \rangle = \langle x, e_i \rangle = 0$$

for any  $i \in S^c$ . Hence,  $x = \sum_{i \in S} \langle x, e_i \rangle e_i$ , so

$$\sum_{i \in S} \langle x, e_i \rangle e_i = x = (I - P)x = \sum_{i \in S} \langle x, e_i \rangle (I - P)e_i,$$

i.e.  $\sum_{i \in S} \langle x, e_i \rangle Pe_i = 0$ , and, thus,  $\{Pe_i\}_{i \in S}$  are linearly dependent.

(2)  $\Rightarrow$  (1). Let's suppose, that  $\{Pe_i\}_{i \in S}$  are linearly dependent: there exist numbers  $\{b_i\}_{i \in S}$ , among which there are nonzero, and  $\sum_{i \in S} b_i Pe_i = 0$ . Then

$$x := \sum_{i \in S} b_i (I - P)e_i = \sum_{i \in S} b_i e_i \in (I - P)(\mathbb{H}^N).$$

Let's consider

$$\langle x, (I - P)e_j \rangle = \langle (I - P)x, e_j \rangle = \left\langle \sum_{i \in S} b_i e_i, e_j \right\rangle = \sum_{i \in S} b_i \langle e_i, e_j \rangle = 0,$$

if  $j \in S^c$ . Thus,  $x \perp \text{span} \{(I - P)e_i\}_{i \in S^c}$ , and hence,

$$\text{span} \{(I - P)e_i\}_{i \in S^c} \neq (I - P)(\mathbb{H}^N).$$

**Proposition 4.** *Parseval-Steklov frame is a full spark frame iff Naimark complement of this frame is a full spark frame also.*

Proof. By theorem 1, Parseval-Steklov frame can be written as  $\{Pe_i\}_{i=1}^N$ , where  $\{e_i\}_{i=1}^N$  is an ONB in  $\mathbb{H}^N$  and  $P$  is the orthogonal projection in  $\mathbb{H}^N$ . Naimark complement for Parseval-Steklov frame looks as  $\{(I - P)e_i\}_{i=1}^N$ . By definition  $\{Pe_i\}_{i=1}^N$  is a full spark frame, if for any  $S \subseteq \{1, \dots, N\}$  with  $|S| = M$   $\{Pe_i\}_{i \in S}$  is a basis in the range of the projection  $P$ . By theorem 3, we have that  $\{(I - P)e_i\}_{i \in S^c}$  is a basis in the range of the projection  $I - P$ , so  $\{(I - P)e_i\}_{i=1}^N$  is a full spark frame also. The reverse assertion is proved similarly.

If Parseval-Steklov frame ensures recovery without phases, Naimark complement can not provide recovery without phases. The thing is including, in particular, that in Naimark complement may be insufficient number of vectors.

**Proposition 5.** *If Parseval-Steklov frame  $\{\varphi_n\}_{n=1}^N$  ensures recovery without phases in  $\mathbb{R}^M$ , and Naimark complement to this frame also ensures recovery without phases in  $\mathbb{R}^{N-M}$ , then*

$$2M - 1 \leq N \leq 2M + 1.$$

Proof. If  $\{\varphi_n\}_{n=1}^N$  ensures recovery without phases in  $\mathbb{R}^M$ , then  $N \geq 2M - 1$  (theorem 2). If Naimark complement ensures recovery without phases in  $\mathbb{R}^{N-M}$ , then  $N \geq 2(N - M) - 1$ , or  $N \leq 2M + 1$ .

But Naimark complement can fail to ensure recovery without phases even under conditions of proposition 3.

**Example.** Let  $\{\varphi_m\}_{m=2}^{2M}$  be the full spark frame in  $\mathbb{R}^M$ ,  $M \geq 3$ . Let's put  $\varphi_1 = \varphi_2$ , and let  $S$  be the frame operator for  $\{\varphi_m\}_{m=1}^{2M}$ . Note that  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=2}^{2M}$  is full spark frame, and ensures recovery without phases. For any partition  $S, S^c \subset \{1, \dots, 2M\}$  one of the sets  $S$  or  $S^c$  has at least  $M$  elements from the full spark frame  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=2}^{2M}$  and hence complete in  $\mathbb{R}^M$ .

Now let's show, that Naimark complement for  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=1}^{2M}$  does not ensure recovery without phases. Let's break  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=1}^{2M}$  on  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=1}^2$  and  $\{S^{-\frac{1}{2}}\varphi_m\}_{m=3}^{2M}$ . None of them is linear independent, as  $\varphi_1 = \varphi_2$ , and  $M \geq 3$ . According to theorem 3, Naimark complements for each of these sets are not complete in  $\mathbb{R}^{2M-M} = \mathbb{R}^M$ . Thus, there is a partition of Naimark complement which contradicts the complement property and does not ensure phaseless recovery.

If Parseval-Steklov frame is full spark frame, then phaseless recovery is inherited by Naimark complement.

**Proposition 6.** *If  $\Phi = \{\varphi_n\}_{n=1}^N$  is full spark Parseval-Steklov frame,  $2M - 1 \leq N \leq 2M + 1$ , then  $\Phi$  ensures phaseless recovery in  $\mathbb{R}^M$ , and Naimark complement for  $\Phi$  ensures phaseless recovery in  $\mathbb{R}^{N-M}$ .*

Proof. By proposition 2 Naimark complement for  $\Phi$  is full spark frame in  $\mathbb{R}^{N-M}$ . We have  $2M - 1 \leq N$  and  $2(N - M) - 1 \leq N$ , then, by theorem 2, both  $\Phi$  and its Naimark complement have complement property in relevant spaces.

## 2. Recovery by the norms of projections

Following [4, 10] we define the recovery of a vector-signal by the norms of projections on subspaces.

**Definition 8.** Let  $\{W_n\}_{n=1}^N$  be the set of subspaces in  $\mathbb{H}^M$ , let  $\{P_n\}_{n=1}^N$  be orthogonal projections on these subspaces.

We say, that  $\{W_n\}_{n=1}^N$  (or  $\{P_n\}_{n=1}^N$ ) ensures recovery by the norms of projections, if for any  $x, y \in \mathbb{H}^M$  equalities  $\|P_n x\| = \|P_n y\|$  for  $n = 1, \dots, N$  imply  $x = cy$  for some  $c$  with  $|c| = 1$ .

Further such sets of subspaces will be called *RNP-sets*.

A lot of attention to such recovery is paid in [10]. For one-dimensional subspace  $W_n$  the number  $\|P_n x\|$  can be received only from two vectors  $\pm P_n x$ . For subspaces  $W_n$  with higher dimensions we have continuum of vectors with  $\|P_n x\|$ .

Nevertheless the map

$$\mathcal{A}(x)(n) = \|P_n x\|$$

can be injective for subspaces with higher dimensions. The proof of this result uses the scheme of [10], we need some auxiliary assertions.

**Lemma 1.**

Let  $\{\varphi_n\}_{n=1}^N$  be full spark frame in  $\mathbb{R}^M$ . Let's define ONB in  $\mathbb{R}^M$  using the following algorithm:  $\psi_1$  is a random vector;  $\psi_2$  is a random vector from  $[\text{span}(\psi_1)]^\perp$ , ...,  $\psi_k$  is a random vector from  $[\text{span}(\{\psi_n\}_{n=1}^{k-1})]^\perp$ . Then  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^M$  is the full spark frame with the probability 1.

Proof.

Let  $1 \leq k < M$ . We suppose, that  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^k$  is full spark frame, we need to check, that  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^{k+1}$  is full spark frame too. For this we have to show that  $\psi_{k+1}$  does not lie in the span of any  $M - 1$  vectors from  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^k$ . Choose any  $M - 1$  such vectors and denote them by  $A$ . Put  $W_k := [\text{span}(\{\psi_m\}_{m=1}^k)]^\perp$  and pick  $\psi_{k+1}$  as a random unit norm vector from this  $(M - k)$ -dimensional space. Then  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^{k+1}$  is full spark system  $\Leftrightarrow \psi_{k+1} \notin \text{span}(A)$ . The last is truly with probability 1 iff

$$\dim(\text{span}(A) \cap W_k) \leq (M - k) - 1. \quad (1)$$

In fact,  $\text{span}(A) \cap W_k$  is a subset in  $(M - k)$ -dimensional space  $W_k$ , and so inequality (1) implies that this intersection has zero measure. Hence, we have with probability 1  $\psi_{k+1} \notin \text{span}(A) \cap W_k$  and  $\psi_{k+1} \in W_k$ . Now we are going to the proof of inequality (1).

Let's apply the method of mathematical induction. A vector  $\psi_1$  is chosen randomly from  $W_0 = \mathbb{R}^M$ . If  $A$  any  $M - 1$  vectors from  $\{\varphi_n\}_{n=1}^N$ , then

$$\dim(\text{span}(A) \cap W_0) = M - 1,$$

and  $\{\varphi_n\}_{n=1}^N \cup \psi_1$  is full spark frame with probability 1.

Let's suppose that  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^k$  is full spark frame. We denote by  $A$  any  $M - 1$  vectors from  $\{\varphi_n\}_{n=1}^N \cup \{\psi_m\}_{m=1}^k$ .

Let's consider two possible cases.

1.  $\psi_k \notin A$ . We have  $W_k \subset W_{k-1}$  and

$$\text{span}(A) \cap W_k = (\text{span}(A) \cap W_{k-1}) \cap W_k.$$

Note that  $\dim W_k = M - k$ ,  $\dim(\text{span}(A) \cap W_{k-1}) \leq M - k$ , because  $\psi_k \notin A$ . So for the proof (1) it's suffice to check that these subspaces do not match. Let's suppose that  $\text{span}(A) \cap W_{k-1} = W_k$ . We remember that  $\psi_k \in W_k^\perp$ , and hence,  $\psi_k \in [\text{span}(A) \cap W_{k-1}]^\perp$ , this subspace has dimension  $k$ . As  $\psi_k \notin W_{k-1}^\perp$ ,  $\dim W_{k-1} = k - 1$ , and

$$W_k^\perp \subset [\text{span}(A) \cap W_{k-1}]^\perp,$$

it turns to be that  $\psi_k$  lies in one-dimensional subspace, determined by  $\text{span}(A)$  and  $W_{k-1}$ . It's possible only with zero probability for randomly chosen vector from  $M - (k - 1)$ -dimensional subspace  $W_{k-1}$ .

2.  $\psi_k \in A$ . Let's note that

$$\dim(\text{span}(A) \cap W_k) \leq M - k,$$

as  $\dim(W_k) = M - k$ . For contradiction, we suppose that

$$\dim(\text{span}(A) \cap W_k) = M - k. \quad (2)$$

We have further that

$$W_k \subset \text{span}(A). \quad (3)$$

Pick  $\varphi \in \{\varphi_n\}_{n=1}^N$  so that  $\varphi \notin A$ . Then

$$\dim(\text{span}(A \setminus \psi_k) \cap W_k) \leq \dim(\text{span}(A \setminus \psi_k \cup \varphi) \cap W_k) \leq (M - k) - 1.$$

The last inequality is a result of the first case above.

On the other hand as  $\psi_k \perp W_k$  and  $\psi_k \in A$ , we receive from (2) and (3)

$$\dim(\text{span}(A \setminus \psi_k) \cap W_k) = \dim(\text{span}(A) \cap W_k) = M - k.$$

This contradiction proves (1).

**Corollary 3.** *The finite set of ONB, which are built by the algorithm of random choice of lemma 1, is full spark frame with the probability 1.*

Proof. Let's apply consistently the lemma 1.

**Lemma 2.** *For an integer  $M \geq 2$  let's pick integers  $M - 1 \geq I_1 \geq I_2 \geq \dots \geq I_1 \geq 1$ . There is a real invertible  $M \times M$ -matrix with 0 - 1 instances such that the  $k$ -row has exactly  $I_k$  ones.*

Proof.

We apply induction by  $M$ . The claim is obvious for  $M = 2$ . Let's suppose that the assertion is valid for  $M$ . Let's look at the set of  $M + 1$  numbers such that

$$M = I_1 = \dots = I_s > I_{s+1} \geq \dots \geq I_{M+1} \geq 1$$

for some  $s \leq M + 1$ . By induction assumption for the set of numbers

$$I_1 - 1 = \dots = I_s - 1 \geq I_{s+1} \geq \dots \geq I_M \geq 1$$

there is the invertible  $M \times M$ -matrix  $A = [a_{ij}]_{i,j=1}^M$  with  $I_{k-1} - 1 = M - 1$  ones in  $k$ -row for  $k = 1, \dots, s$  and  $I_k$  ones in  $k$ -row for  $k = s + 1, \dots, M$ . Let's define  $(M + 1) \times (M + 1)$ -matrix  $B = [b_{ij}]_{i,j=1}^{M+1}$  defining

$$b_{ij} = \begin{cases} a_{ij}, & 1 \leq i, j \leq M, \\ 1, & 1 \leq i \leq s, j = M + 1, \\ 1, & i = M + 1, 1 \leq j \leq M + 1, \\ 0, & \text{for other indexes.} \end{cases}$$

The matrix  $B$  has  $I_k$  ones in  $k$ -row for  $k = 1, \dots, M + 1$ . The matrix  $A = [a_{ij}]_{i,j=1}^M = [b_{ij}]_{i,j=1}^M$  is invertible, so the matrix  $B$  by row reduces can be reduced to the step form  $\widetilde{B} = [\widetilde{b}_{ij}]_{i,j=1}^{M+1}$ , where  $[\widetilde{b}_{ij}]_{i,j=1}^M = I_{M \times M}$ , and the row  $(M + 1)$  is not changed. If we suppose that  $\widetilde{B}$  is not invertible, then the row  $(M + 1)$  by row reduces can be reduced to the zero row and hence

$$\sum_{i=1}^{I_{M+1}} \widetilde{b}_{M+1,i} = 0. \tag{5}$$

Let's define for each  $l \in \{1, \dots, I_{M+1}\}$  the matrix  $\widetilde{B}_l$ . It is obtained from the matrix  $\widetilde{B}$  changing  $\widetilde{b}_{M+1,M+1} = 0$  to  $\widetilde{b}_{M+1,l} = 1$ .

If  $\widetilde{B}$  is not invertible, then by row reduces the last row is reduced to the zero row, and we have

$$\sum_{i=1, i \neq l}^{I_{M+1}} \widetilde{b}_{M+1,i} = -1. \tag{6}$$

The equality (6) is valid for any  $l \in \{1, \dots, I_{M+1}\}$ , that's contradict to (5). Hence at least one of the matrixes  $\widetilde{B}$  or  $\widetilde{B}_l$  for some  $l \in \{1, \dots, I_{M+1}\}$  has to be invertible.

**Theorem 6.** *There exists RNP-set in  $\mathbb{R}^M$  consisting from  $2M - 1$  subspaces, dimension of each subspace  $< M - 1$ . Proof.*

Let  $\{\varphi_n\}_{n=1}^{2M-1}$  be the set of vectors in  $\mathbb{R}^M$  with complement property and with additional requirement of orthogonality and normalization ( $\|\cdot\| = 1$ ) to the sets  $\{\varphi_n\}_{n=1}^M$  and  $\{\varphi_n\}_{n=M+1}^{2M-1}$ . The corollary 2 ensures the existence of such set. Let  $I_k \subseteq \{1, \dots, M\}$  for  $k = 1, \dots, M$ , and  $J_k \subseteq \{M + 1, \dots, 2M - 1\}$  for  $k = M + 1, \dots, 2M - 1$ , let  $P_{I_k}$  and  $P_{J_k}$  be projections on span( $\{\varphi_n\}_{n \in I_k}$ ) and span( $\{\varphi_n\}_{n \in J_k}$ ) respectively. The next construction ensures phaseless recovery for  $x \in \mathbb{R}^M$  by  $\|P_{I_k}x\|$  and  $\|P_{J_k}x\|$  for  $k = 1, \dots, 2M - 1$ .

Let  $A = [a_{kz}]_{k,z=1}^M$  be  $M \times M$ -matrix, its rows are agreed with  $I_k$ , i. e.  $a_{kz} = 1$ , for  $z \in I_k$ , and  $a_{kz} = 0$  for other  $z$ .

Similarly we define the matrix  $B = [b_{kz}]_{k,z=1}^{M-1}$  as  $(M - 1) \times (M - 1)$ -matrix with  $b_{kz} = 1$  for  $z + M \in J_k$ , and  $b_{kz} = 0$  for other  $z$ .

Let's look at the subspaces span( $\{\varphi_n\}_{n \in I_k}$ ) for  $k = 1, \dots, M$ . For  $x \in \mathbb{R}^M$  we have

$$\|P_{I_k}x\|^2 = \sum_{n \in I_k} |\langle x, \varphi_n \rangle|^2,$$

whence

$$\begin{bmatrix} \|P_{I_1}x\|^2 \\ \vdots \\ \|P_{I_M}x\|^2 \end{bmatrix} = A \begin{bmatrix} |\langle x, \varphi_1 \rangle|^2 \\ \vdots \\ |\langle x, \varphi_M \rangle|^2 \end{bmatrix}.$$

This equation may be solved upon  $\{|\langle x, \varphi_n \rangle|\}_{n=1}^M$ , if the matrix  $A$  is invertible. Similar equation may be written with the matrix  $B$ . Hence if the matrixes  $A$  and  $B$  are invertible, we obtain the complete set of "measurements"  $\{|\langle x, \varphi_n \rangle|\}_{n=1}^{2M-1}$ . The set  $\{\varphi_n\}_{n=1}^{2M-1}$  has complement property and according to theorem 2, phaseless recovery is possible using subspaces span( $\{\varphi_n\}_{n \in I_k}$ ) and span( $\{\varphi_n\}_{n \in J_k}$ ) for  $k = 1, \dots, 2M - 1$ . To complete the proof we choose  $\{I_k\}_{k=1}^M$  and  $\{J_k\}_{k=M+1}^{2M-1}$  to provide the invertibility of the matrixes  $A$  and  $B$ .

Let's note that the quantity of ones in each row coincides with the dimension of the appropriate subspace. Such selection is possible according to lemma 2 for any subsets  $I_k, J_k$ , with  $1 \leq |I_k| \leq M - 1$  and  $1 \leq |J_k| \leq M - 2$ .

The answer to the next question is unknown [4]:

**Question.** *Is it possible phaseless recovery by norms of projections in  $\mathbb{R}^M$  with the set of subspaces  $\{W_n\}_{n=1}^N$  for  $N < 2M - 1$ ?*

### Acknowledgements

The first author was supported by RFBR grant N 7-01-00138.

### References

[1] Novikov SYa, Likhobabenko MA. Frames in finite-dimensional spaces. Samara: Samarsky Universitet, 2013; 52 p.  
 [2] Christensen O. An Introduction to Frames and Riesz bases. Boston: Birkh auaser, 2003; 403 p.  
 [3] Kashin BS, Kulikova TYu. A note about Frames. Matem. zametki 2002; 72(6): 941–945.  
 [4] Bahmanpour S, Cahill J, Casazza PG, Jasper J, Woodland LM. Phase retrieval and norm retrieval. URL : <https://arxiv.org/abs/1409.8266>.  
 [5] Bandeira AS, Cahill J, Mixon G, Nelson AA. Saving phase: Injectivity and stability. URL: <https://arxiv.org/abs/1302.4618v1>.  
 [6] Alexeev B, Cahill J, Mixon DG. Full Spark Frames. URL: <https://arxiv.org/abs/1110.3548>.  
 [7] Botelho-Andrade S, Casazza PG, Nguyen HV, Tremain JC. Phase retrieval verses phaseless reconstruction. URL: <https://arxiv.org/abs/1507.05815>.  
 [8] Novikov SYa, Fedina ME. Restoring the signal be modules of measurement. Vestnik Samara University 2016; 3-4: 63–74.  
 [9] Bodmann B, Casazza PG, Paulsen V, Speegle D. Spanning and independence properties of frame partitions. Proc. Am. Math. Soc. 2012; 140(7): 2193–2207.  
 [10] Cahill J, Casazza PG, Peterson J, Woodland LM. Phase retrieval by projections. URL: <https://arxiv.org/abs/1305.6226.8858>.  
 [11] Horn R, Johnson Ch. Matrix analysis. Moskva: Mir, 1989; 655 p.  
 [12] Cahill J, Chen X. A note on scalable frames. URL: <https://arxiv.org/abs/1301.7292>.  
 [13] Bodmann B, Haas J. Frame potentials and the geometry of frames. Journal of Fourier Analysis and Applications 2015; 21(6): 1344–1383.  
 [14] Vale R, Waldron S. Tight frames and their symmetries. Constructive Approximation 2004; 21(1): 83–112.

# Cellular automata-based model of group motion of agents with memory and related continuous model

Alexander V. Kuznetsov<sup>1</sup>

<sup>1</sup>Voronezh State University, 1 Universitetskaya pl., Voronezh, 394018, Russia

## Abstract

The paper describes the construction of the motion and interaction model for agents with memory. Agents move on the landscape consisting of squares with different passability. We briefly characterize the cellular automata-based model with one common to all agents layer corresponding to the landscape and many agent-specific layers corresponding to an agent's memory. Also, we develop methods for the random landscape generation and the simulation of a communication system. Finally, we study a connection between the discrete agent motion model and the continuous concentration law for the system of agents.

*Keywords:* cellular automaton; motion model; conflict model; agent system; random landscape generation; landscape metrics; concentration law

## 1. Introduction and definitions

Previously the author studied cellular automaton-based models of motion [1] and communication [2]. The initial idea of the proposed model described in the article [3]. In this paper I continue the previous work, propose cellular automaton that takes into account the history of the movement of agents. Also, I obtain few quantitative characteristics of the model and found the continuous equation and corresponding problem for the partial differential equation which describes a dependence of the agents' number in the direction of agents motion on time and position. Note, that an automaton of the type mentioned above can be viewed as a 0th order reflexive automaton [4].

Give definitions according to the work [5].

**Definition 1.** Let us call landscape  $\mathcal{L}_l(n \times m)$  rectangle from  $n \times m = N$  cells  $\omega_{ij}$ ,  $(i, j) \in I \subset \mathbb{Z}^2$  with equal size belonging to  $l$  different classes and that to  $i$ -th class it belongs  $N_i$  cells, i.e.  $\sum_{i=1}^l N_i = N$ .

Note that for landscapes generated for testing of path-finding algorithms, landscape cells will be divided into classes according to the maximum possible cell-crossing speed.

**Definition 2.** Configuration entropy of the landscape  $\mathcal{L} = \mathcal{L}_l(n \times m)$  is defined as

$$S(\mathcal{L}_l(n \times m)) = - \sum_{i=1}^l \frac{N_i}{N} \ln \frac{N_i}{N}$$

and characterizes landscape heterogeneity in whole.

**Definition 3.** Total Edge is the total number of abutting edges of cells, belonging to different classes, in  $\mathcal{L}$ . We will further denote the Total Edge of the landscape  $\mathcal{L}$  as  $TE(\mathcal{L})$ .

**Definition 4.** Total Edge Density (TED) of the landscape  $\mathcal{L}$  is the ratio  $TE(\mathcal{L})$  to the total cell quantity  $N$  in the  $\mathcal{L}$

$$TED(\mathcal{L}) = TE(\mathcal{L})/N.$$

**Definition 5.** Denote Euclidean distance between  $x = (x_1, \dots, x_n)$ ,  $y = (y_1, \dots, y_n)$ ,  $x, y \in \mathbb{R}^n$  as

$$\|x - y\| = \left( \sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}.$$

## 2. Description of the automaton

Let the  $Ag = \{ag_1, \dots, ag_k\}$  is the system of agents, which move across the landscape  $\mathcal{L}_l(n \times m)$ , and initial and final cells of the landscape are specified for each agent. The idea of the article is that in the model in addition to the total for all agents "layer" corresponding to objective reality, each agent would have been "layer" corresponding to the information about the reality, which is known to this agent.

The behavior of the agents of the system is modeled by a cellular automaton in which the set of cells is  $World = \{(i, j, id) | i, j \in \mathbb{Z}, id = \overline{0, k}\} \subset \mathbb{Z}^3$ . In the set  $World$  we will allocate  $k + 1$  cell planes: a layer of the objective reality  $OWorld = \{(i, j, 0) | i, j \in \mathbb{Z}\}$ , and the layers of subjective reality of agent with identifier  $ag = \overline{1, k}$   $SWorld_{ag} = \{(i, j, ag) | i, j \in \mathbb{Z}\}$ . In this way,

$$World = OWorld \cup \left( \bigcup_{ag=1}^k SWorld_{ag} \right).$$

We assume that the rectangle  $K \subset World$ ,  $K = \{(i, j, id) | i = \overline{0, L_K}, j = \overline{0, L_K}, id = \overline{0, k}\}$  is selected and all cells are in the resting state outside of it.



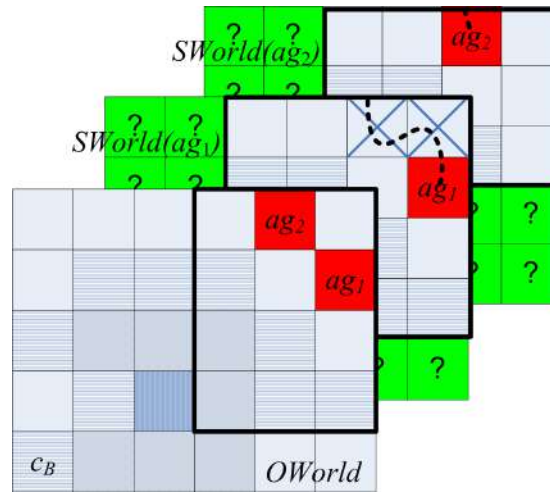


Fig. 1. The sample of the cellular automaton.

### 2.1. Objective reality

Let the objective reality layer  $OWorld$  consists of cells  $(i, j) \in \mathbb{Z}^2$  with different impassability  $u_{ij}$ . The value of  $u_{ij}$  is the number of discrete time units which is required to pass the square  $\omega_{ij}$  with coordinates  $(i, j)$ . If  $\omega_{ij}$  is completely impassable then put  $u_{ij} = -1$ . Also cells can include the information about an agent in a cell, the agent's destination square etc.

### 2.2. Subjective reality

The subjective reality layer consists of cells  $(i, j)$  so that each its cell  $(i, j)$  corresponds the cell  $(i, j)$  of the objective reality layer. Cells of the subjective reality layer  $ag$  contain the information about the current position of the agent  $ag$ , about the history of the  $ag$  motion and about the impassability of known to the agent  $ag$  cells.

### 2.3. The automaton's functioning

Briefly describe the cellular automaton (CA) functioning. Let us denote

$$\mathcal{D} = \{(-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0), (1, 1)\}.$$

**Definition 6.** Let us call agent's cellular route the sequence

$$M = \{(i_1, j_1), (i_2, j_2), \dots, (i_s, j_s) | (i_k, j_k) \in \mathbb{Z}^2, k = \overline{1, s}, (i_{k+1} - i_k, j_{k+1} - j_k) \in \mathcal{D}, k = \overline{1, s-1}\},$$

such as the agent in the square  $\omega_{i_1, j_1}$  will be sequentially move into squares  $\omega_{i_2, j_2}, \dots, \omega_{i_s, j_s}$ . Denote the set of all cellular routes starting in the cell with coordinates  $c_A \in \mathbb{Z}^2$  and ending in the cell with coordinates  $c_B \in \mathbb{Z}^2$  as  $\mathcal{M}(c_A; c_B)$ .

Let us define a function

$$\theta(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

Introduce the notation:

$$\psi_3(u, T_{max}) = \begin{cases} u, & u \geq 0, \\ T_{max}, & u < 0. \end{cases}$$

If cell impassability does not change over time, then it is not necessary to consider the routes containing impassable cells. However, if the impassable cell can become passable, these routes should be taken into account. To do this, define the functional

$$\tilde{T}_h(M) = \sum_{(i,j) \in M} \|d_{ij}\| \psi_3(u_{ij}, T_{max}).$$

We call weight of the route  $M$  for the agent  $ag$  the following:

$$\Lambda(M; \alpha, \beta, \gamma, T_{max}) = \alpha \tilde{T}_h(M) + \beta \sum_{(i,j) \in M} \theta(f_{ij}) + \gamma \sum_{(i,j) \in M} vis_{ij}(ag),$$

where  $u_{ij}$  is the impassability of the square  $\omega_{ij}$ ,  $vis_{ij}$  is the number of visits of the square  $\omega_{ij}$  (it is contained in the subjective reality layer  $SWorld(ag)$ ),  $\alpha, \beta, \gamma$  are parameters.

The agent in the square  $\omega_{ij}$  each discrete time tick tries to find locally optimal (in a neighborhood  $V_o(i, j)$  with radius  $o$ ) route  $M_o$  such that  $\Lambda(M_o; \alpha, \beta, \gamma, T_{max}) \rightarrow \min$  and go through this route. Therefore, the agent's route at whole constructs from locally optimal subroutes.

Agent  $ag$  can apply previously described approach for route searching in undiscovered by this agent areas, i.e. in consisting of cells  $\omega_{ij}$  with  $vis_{ij}(ag) = 0$ . More standard approaches to optimum route search, for example, Dijkstra's algorithm, can be used in areas composed of cells already visited. However, the use of standard methods of searching for an optimal route is limited to a rate of landscape change over time. It is possible that information about the visited cells become outdated (parameter  $time_{ij}(ag)$  is used for determining the actuality of information), or even impassibility of the cells would change directly during the process of passing the route selected as the globally optimal.

Thus, depending on the speed of the landscape changes, it is necessary to find a compromise between the approach "Reacting", which evaluates the current situation immediately near of the agent and the approach "Planning", in which searched globally optimal trajectory. For example, it is pointless to set the radius  $o$  of the neighborhood  $V_o(i, j)$ , in which agent searches locally optimal route more than the number of ticks during which the landscape has remained unchanged.

The example of the described cellular automaton is depicted on the fig. 1. Increasing of the impassability at the mentioned figure is indicated with a darker tone, crosses "x" in the layer  $S World$  mark already visited cells, marks "?" correspond to cells whose status is unknown.

### 3. Function of obstacles

Turn to the continuous formulation of obstacle avoidance problem to construct transfer function for our CA. The agent moves in the domain  $\Omega$  with changing over time obstacles from the point  $A$  to the point  $B$  with route  $r(t)$ ,  $t \in [0, T]$  in the shortest time  $T$ . This problem has the form

$$\|\dot{r}(t)\| = v(t, r(t)), \tag{1}$$

$$r(0) = A, \quad r(T) = B, \tag{2}$$

$$T \rightarrow \min. \tag{3}$$

We will call further the function  $v : [0, T] \times \Omega \rightarrow \mathbb{R}$  as "function of obstacles". Divide the segment  $[0, T]$  onto the  $k$  subsegments with length  $\tau > 0$ , domain  $\Omega$  onto squares  $\omega_{ij}$  with numbers  $(i, j) \in \mathbb{Z}^2$  and the length of a side  $h$ . Approximate at each moment of time  $k\tau \in [0, T]$  on the square  $\omega_{ij}$  function  $v(k\tau, \cdot)$  with the constant function  $v_{ij}^k(h, \tau)$ .

Go to the discrete time for the model simplifying. Let

$$t_h = \frac{h}{\max_{(i,j) \in I_h, k \in T_\tau} v_{ij}^k(h, \tau)}.$$

If relations

$$u_{ij}(k) = \frac{h}{t_h v_{ij}^k(h, \tau)} = \frac{\max_{(i,j) \in I_h, k \in T_\tau} v_{ij}^k(h, \tau)}{v_{ij}^k(h, \tau)}$$

hold, define that square  $\omega_{ij}$  on the state in moment  $k\tau \in [0, T]$  is crossable in non-diagonal direction in  $u_{ij}(k)$  ticks.

Moreover, it is possible to go to the integer values of the  $u_{ij}(k)$  by discarding the fractional part and taking  $\tilde{u}_{ij}(k) = [u_{ij}(k)]$ . We associate with the agent in the square  $\omega_{i,j}$  value  $errc_{ij}$  of the cumulative discrete time error. Also we associate with the square  $\omega_{ij}$  error value  $err_{ij} = \{u_{ij}(k)\}$ . When an agent starts to cross the next square  $\omega_{i',j'}$ , the value  $errc_{i',j'}$  increments on the  $err_{i',j'}$ , sets  $errc_{ij} = 0$  and if  $errc_{i',j'} > 1$  then agent passes one tick independently from the value of the function of obstacles in the square  $\omega_{i',j'}$  and sets  $errc_{i',j'} = errc_{i',j'} - 1$ .

Let

$$T : \mathcal{M}(c_A; c_B) \rightarrow \mathbb{R}$$

the functional of time which is required to going through cellular route.

If values of the  $u_{ij}$  do not change in a time of the movement from the point  $A$  to the point  $B$ , then the problem (1)–(3) can be represented as discrete problem

$$T(M) = t_h \sum_{(i,j) \in M} \|d_{ij}\| u_{ij} \rightarrow \min.$$

It is clear that possible to minimize functional

$$T_h(M) = \sum_{(i,j) \in M} \|d_{ij}\| u_{ij} \rightarrow \min$$

instead functional  $T$ .

It is possibly (but not very easy) to prove that the CA mentioned earlier finds the approximation of the solution of the problem (1)–(3) in some subdomain of the  $\Omega$ . The sequence of such approximations  $r_h$  converges to the optimal solution  $r$ , and the following estimate holds:

$$|r(l(t)) - r_h(l_h(t))| \leq (h\sqrt{2} + \tau)(e^{\|\nabla_{(t,x,y)} v\|_{C([0,T] \times \Omega)} K} - 1) + h\sqrt{2}, \tag{4}$$

where  $l, l_h$  are parameterizations of routes,  $h$  is the length of the square  $\omega_{ij}$  side,  $\tau$  is the length of the time tick,  $K > 0$  is the constant depending on a class of routes considered.

**Definition 7.** Define the obstacle which exists in the moment  $t \in [0, T]$  as simply connected set  $Obst \subset \Omega$  such that any  $r_{obst} \in Obst$  is the point of a local minimum of the function of obstacles  $v(t, \cdot)$  and exists  $r_0 \in \Omega$  such that  $v(t, r_0) > v(t, r_{obst})$ .

#### 4. The model of a communication system and conflict

**Definition 8.** Let us define communication graph as follows

$$\Gamma(t) = (Ag, Comm, \varphi(t), M(t)),$$

where  $Comm$  is the set of channels,  $\varphi(t) : Ag \times Comm \rightarrow \{0, 1\}$  is the incidence function,  $M(t) : Comm \rightarrow \mathbb{R}^n$  is the markup function in the moment of time  $t \in [0, N_T]$ . The  $M$  gives the features vector of the channel  $comm \in Comm$ . This features can be channel bandwidth, radio frequency, etc.

Let's introduce a communication graph connected with the motion model. This means that should be given the function  $pag : [0, N_T] \times Ag \rightarrow \mathbb{Z}^2$  which maps agent's coordinates to an agent in each moment  $t \in [0, T]$ . Also, it means that  $M(t)$  and  $\varphi(t)$  depend on properties of cells in which incident agents are currently placed and on properties of cells between of them. Therefore, connections between agents in the  $\Gamma$  can break and establish depending on the agents' speed and landscape type.

Suppose that each agent  $ag \in Ag$  has an own signal exchange timetable. It is possibly also to define specific signals like "enemy detecting", "grouping", etc. We can study various traffic models depending on the agents' timetables, motion speed, and the landscape type.

Finally, we can define "requirements graph"  $\Pi$  and state that communication graph  $\Gamma(t)$  should be similar with  $\Pi$  in some metric each moment of time. Such graph  $\Pi$  can be viewed as a fuzzy set of communication graphs, as an abstract container or as a generator of the stream of communication graphs.

Also, we developed the conflict model combined with the motion and communication model similar to described in the work [6] and its computer simulation "Bokohod." Agents emerge different kinds of tactics and exchange signals without any external control.

#### 5. Computational experiment

Previously the author had developed the algorithm of the landscape generation with the given configuration entropy. This algorithm constructs the vector of numbers of cells in each class  $V = (N_1, \dots, N_l)$  by the given entropy  $S$  as follows:

Step 1. Solve the equation

$$S = -\frac{\beta(1-\beta^l) - (1-\beta)l\beta^l}{(1-\beta)(1-\beta^l)} \ln \beta + \ln \frac{1-\beta^l}{1-\beta},$$

Step 2. Use the found solution  $0 \leq \beta \leq 1$  and equation

$$N_1 = N \frac{1-\beta}{1-\beta^l}$$

to find  $N_1$ ,

Step 3. Compose the vector  $V_0 = (N_1, \beta N_1, \dots, \beta^{l-1} N_1)$ ,

Step 4. Round the components of the  $V_0$  up to integers and obtain the vector  $V_1$  in this way. It is necessary to make rounding such that the sum of all components of  $V_1$  would be equal to  $N$ .

We generate landscape such as the discrete function of the obstacles  $u : \mathbb{Z}^2 \rightarrow \mathbb{Z}$  would have local maxima strictly in  $N_{obst} = V_1 = \beta^{l-1} N_1$  cells. The author thinks that this method gives more natural-like landscapes as it makes "generally passable" area with some hardly passable subareas. As it known from [5] we can make very different landscapes with the same configuration entropy. By this reason, we will use the special, CA-based way of the filling landscape with cells of different classes. This method guarantees slow, near linear increasing of the TED at the increasing of the entropy.

Examples of obtained landscapes are shown on the fig. 2, (a,b). Sample dependencies of the configuration entropy, TED, and  $N_{obst}$  are shown on the fig. 3.

We set  $l = 9, n = m = 48, o = 6$ , choose  $N_{obst} \in \{5\} \cup \{10i | i = \overline{1, 25}\} \cup \{255\}$ . Generate landscape for the each  $N_{obst}$  value and perform the following experiment<sup>1</sup>. Let an agent moves from the cell  $\omega_{11}$  to the cell  $\omega_{nm}$  100 times according to the previously described algorithm. Next, we compute the time which is required for the experiment completion  $T_{bok}^i$  and the time of the moving from the  $\omega_{11}$  to the  $\omega_{nm}$  by linear straight route  $T_{iup}^i$ . Then we calculate the mean value and standard deviation of the win of time for all of this series:

$$\overline{win} = \frac{1}{50} \sum_{i=1}^{50} \frac{T_{iup}^i}{T_{bok}^i}.$$

<sup>1</sup>The raw data and the data processing program for all experiments are stored at [https://www.researchgate.net/publication/316747096\\_The\\_experimental\\_data\\_for\\_group\\_motion\\_of\\_agents\\_with\\_memory\\_with\\_the\\_program\\_in\\_Wolfram\\_Language](https://www.researchgate.net/publication/316747096_The_experimental_data_for_group_motion_of_agents_with_memory_with_the_program_in_Wolfram_Language).

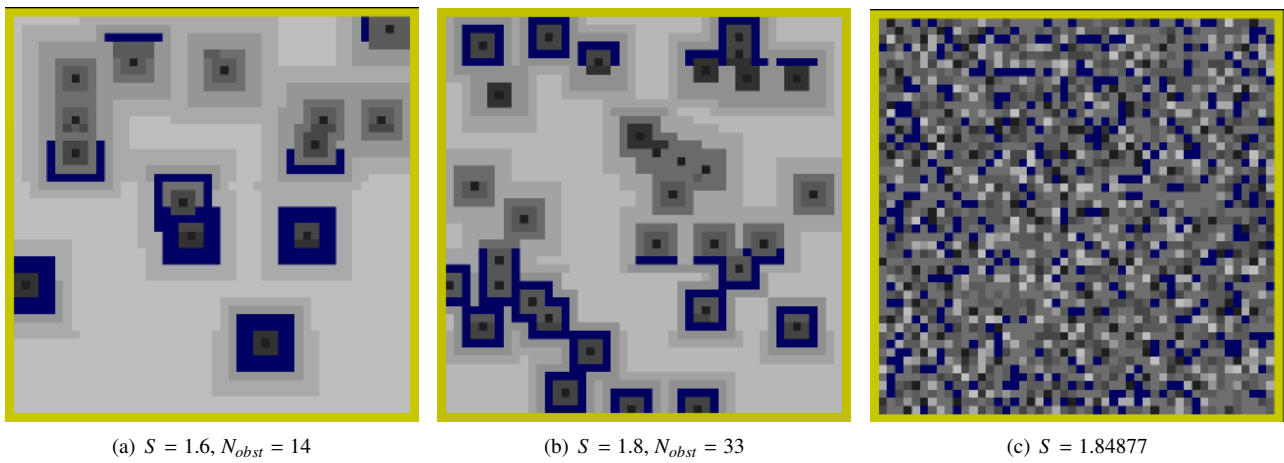
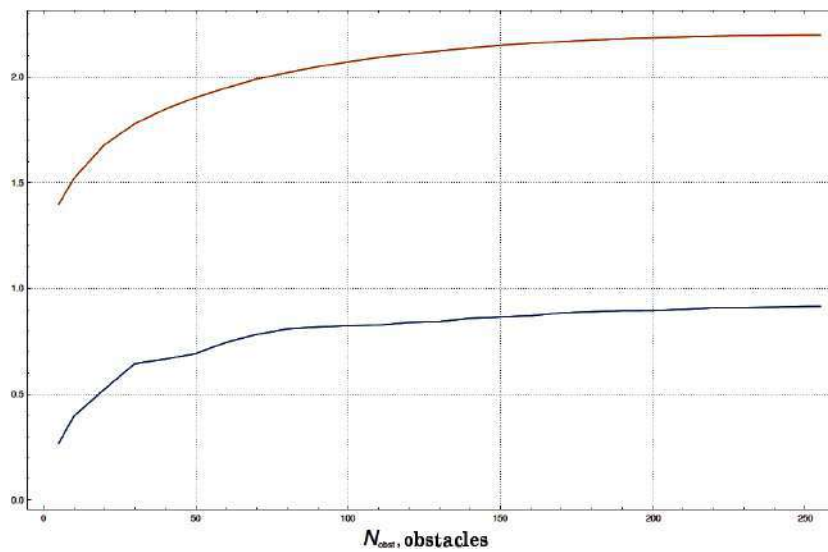


Fig. 2. Examples of Landscapes. A darker cell is more impassable.


 Fig. 3. Sample dependencies of the configuration entropy, TED and  $N_{obst}$ .

$$\sigma = \left( \sum_{i=1}^{50} \left( \overline{win} - \frac{T_{tup}^i}{T_{bok}^i} \right)^2 \right)^{1/2}.$$

It was found that the mean value of the win in transit time  $\overline{win}$  and the configuration entropy of the landscape  $S$  are correlated with a correlation coefficient 0.959556 (see fig. 4). The  $\overline{win}$  and the TED of the landscape are correlated with a correlation coefficient 0.964763. The orange line in the figure corresponds to the curve  $y = (S(N_{obst}) + 1) \ln 9$ ; the brown line corresponds to the curve

$$y = 0.922178 + 0.539383TED(N_{obst}),$$

where  $S(N_{obst})$  and  $TED(N_{obst})$  are the entropy and the total edge density's mean value of the landscape with  $N_{obst}$  obstacles. Vertical bars correspond to the standard deviation of the win of time.

Let's study the dependence of the average number of agents on time moment and position on a landscape. We generate 150 random landscapes with the given entropy by the algorithm mentioned above. The group formed from  $u_0 = 48$  agents moves from the one side of the squared landscape to the opposite one. Compute dependence of the average (through all landscapes generated) number of agents  $u$  at the  $x$ th line of a landscape on the discrete time  $t$ ,  $x = \overline{1, x_{max}}$ ,  $x_{max} = 48$ . Thus, we find that this dependence is determined, mainly, not by the particular kind of landscape, but by the landscape configuration entropy  $S$ . We found the dependence in the form

$$u(x, t) = \frac{u_0}{2} (\text{erf}(\xi_1(S; x, t)) - \text{erf}(\xi_2(S; x, t))), \quad (5)$$

where

$$\begin{aligned} \xi_1(S; x, t) &= \frac{a(S; x)}{\sqrt{t}} + b_1(S; x), \\ \xi_2(S; x, t) &= \frac{a(S; x)}{\sqrt{t}} + \text{sgn}(48 - x)b_2(S; x), \end{aligned}$$

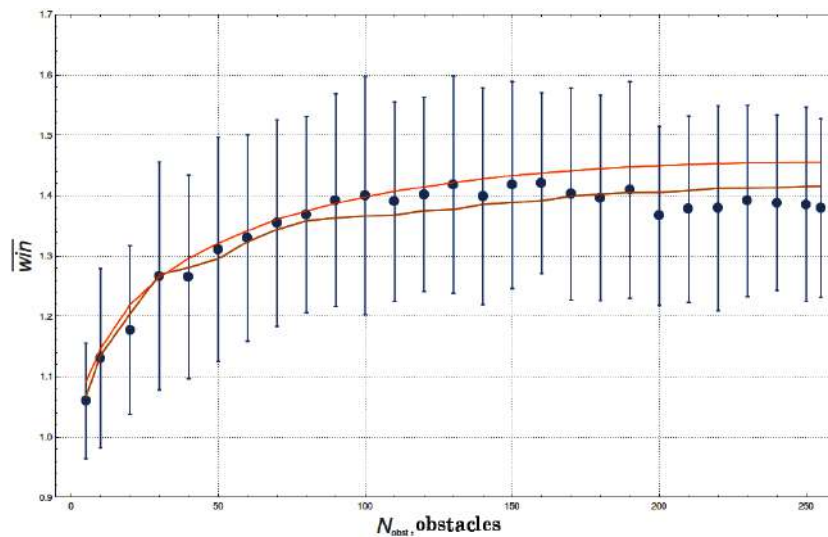


Fig. 4. Win of the time.

$$\text{sgn}(x) = \begin{cases} 1, & x > 0, \\ -1, & x \leq 0. \end{cases}$$

This form of  $\xi_1$  and  $\xi_2$  parameters was assumed by the analogy with the problem of heat propagation along a rod with a thermal diffusivity  $a$  heated on its segment  $[l_1, l_2]$  to the temperature  $u_0$ . This problem has (see, for example, [7]) solution

$$u(x, t) = \frac{u_0}{2} \left( \text{erf} \left( \frac{x - l_1}{2a \sqrt{t}} \right) - \text{erf} \left( \frac{x - l_2}{2a \sqrt{t}} \right) \right).$$

Let us assume that

$$\begin{aligned} a(S; x) &= a_1(S)x + a_2(S), \\ b_1(S; x) &= b_{11}(S) \sqrt{x} + b_{12}(S) + \frac{b_{13}(S)}{x^{3/2}}, \\ b_2(S; x) &= b_{11}(S) \sqrt{x} + b_{22}(S). \end{aligned}$$

These functions allow to approximate  $u = u(x, t)$  with the coefficient of determination  $r^2 > 0.97$ , the mean absolute error  $MAE < 0.1251$ , and the median absolute error  $MedAE < 0.04366$  with every value of the entropy  $S$ . Experimental and approximated values of  $u$  are shown on the fig. 5. For example, when  $N_{obst} = 20$  ( $S = 1.67909$ )

$$\begin{aligned} u(x, t) = 24 \left( \text{erf} \left( \frac{-2.04693x - 1.03949}{\sqrt{t}} + \frac{4.66438}{x^{3/2}} + 1.08597 \sqrt{x} + 0.926477 \right) - \right. \\ \left. - \text{erf} \left( \frac{-2.04693x - 1.03949}{\sqrt{t}} + 1.08597 \sqrt{x} + \text{sgn}(48 - x)0.845118 \right) \right). \end{aligned}$$

Next, we try to find a problem for a partial differential equation which can have a solution in the form (5). Naturally, we can assume that this equation is

$$\frac{\partial u}{\partial t} = C_1 \frac{\partial^2 u}{\partial x^2} + C_2 \frac{\partial u}{\partial x},$$

the initial condition is

$$\begin{aligned} u(x, 0) &= u_0 \delta(x - 1), \\ \delta(x) &= \begin{cases} 1, & x = 0, \\ 0, & x \neq 0, \end{cases} \end{aligned}$$

and the asymptotic condition is

$$\begin{aligned} \lim_{t \rightarrow \infty} u(x, t) &= u_0 \theta(x - x_{\max}), \\ \theta(x) &= \begin{cases} 1, & x \geq 0. \\ 0, & x < 0, \end{cases} \end{aligned}$$

Let us denote

$$\begin{aligned} U_1(x, t) &= e^{-\left( \frac{a(S;x)}{\sqrt{t}} + b_{22}(S) + b_{11}(S) \sqrt{x} \right)^2}, \\ U_2(x, t) &= e^{-\left( \frac{a(S;x)}{\sqrt{t}} + b_{11}(S) \sqrt{x} + b_{12}(S) + \frac{b_{13}(S)}{x^{3/2}} \right)^2}. \end{aligned}$$

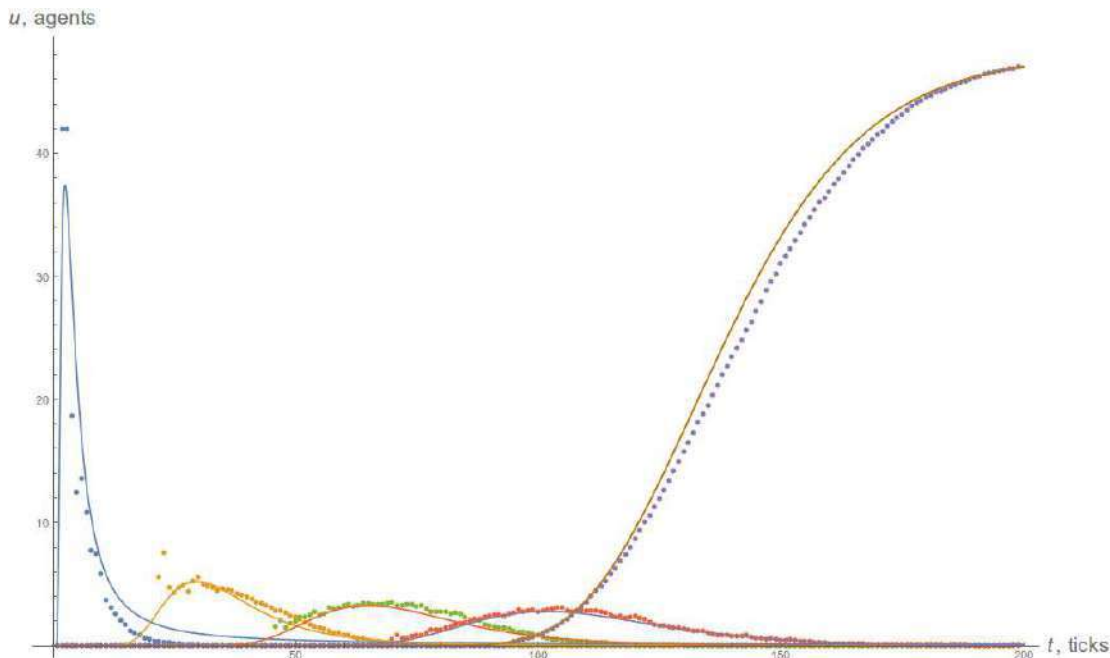


Fig. 5. The dependence of the average number of agents  $u$  on the discrete time  $t$ ,  $N_{obs} = 20$ ,  $x = 2$ ,  $x = 12$ ,  $x = 24$ ,  $x = 36$ , and  $x = 48$  from the left to right.

Compute from the (5) that when  $x \in (1, x_{\max})$

$$\frac{\partial u}{\partial t}(x, t) = \frac{u_0(U_1(x, t) - U_2(x, t))(a_1(S)x + a_2(S))}{2\sqrt{\pi t^3}}, \quad (6)$$

$$\frac{\partial u}{\partial x}(x, t) = -\frac{u_0\left(x^2(U_1(x, t) - U_2(x, t))\left(2a_1(S)\sqrt{x} + b_{11}(S)\sqrt{t}\right) + 3b_{13}(S)\sqrt{t}U_2(x, t)\right)}{2\sqrt{\pi t x^5}}, \quad (7)$$

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}(x, t) = & \frac{u_0}{4\sqrt{\pi t^{3/2} x^{13/2}}}\left(b_{13}(S)\sqrt{t}U_2(x, t)x^3\left(4\left(4a_1(S)^2x^2 + 6a_1(S)a_2(S)x + 7a_1(S)b_{11}(S)\sqrt{tx^3} + 3a_2(S)b_{11}(S)\sqrt{tx}\right) + \right. \\ & \left. + 24a_1(S)b_{12}(S)\sqrt{tx} + t\left(10b_{11}(S)^2x + 12b_{11}(S)b_{12}(S)\sqrt{x} + 15\right)\right) + 8a_1(S)^2\sqrt{tx^{13}}(b_{22}(S)U_1(x, t) - b_{12}(S)U_2(x, t)) - \\ & - 6b_{13}(S)^2tU_2(x, t)x\left(-a_1(S)x^{3/2} + 3a_2(S)\sqrt{x} + b_{11}(S)\sqrt{tx} + 3b_{12}(S)\sqrt{tx}\right) + 8a_1(S)b_{11}(S)tx^6(b_{22}(S)U_1(x, t) - b_{12}(S)U_2(x, t)) + \\ & + 10a_1(S)b_{11}(S)^2tU_1(x, t)x^{13/2} - 10a_1(S)b_{11}(S)^2tU_2(x, t)x^{13/2} + 2b_{22}(S)b_{11}(S)^2t^{3/2}U_1(x, t)x^{11/2} - 2b_{11}(S)^2b_{12}(S)t^{3/2}U_2(x, t)x^{11/2} - \\ & - 18b_{13}(S)^3t^{3/2}U_2(x, t) + (U_1(x, t) - U_2(x, t))\left(8a_1(S)^3x^{15/2} + 8a_1(S)^2a_2(S)x^{13/2} + 8a_1(S)b_{11}(S)\sqrt{tx^6}(2a_1(S)x + a_2(S)) + \right. \\ & \left. + 2a_2(S)b_{11}(S)^2tx^{11/2} + 2b_{11}(S)^3t^{3/2}x^6 + b_{11}(S)t^{3/2}x^5\right). \quad (8) \end{aligned}$$

Solve (7), (8) with respect to  $U_1, U_2$ . Substitute into (6) values found and obtain

$$C_1(x, t) = \frac{P_1(\sqrt{x}, \sqrt{t})}{Q_1(\sqrt{x}, \sqrt{t})}, \quad C_2(x, t) = \frac{P_2(\sqrt{x}, \sqrt{t})}{Q_2(\sqrt{x}, \sqrt{t})},$$

where

$$\begin{aligned} Q_1(\sqrt{x}, \sqrt{t}) = & -3b_{11}(S)tx^2\left(4a_1(S)^2x^{13/2}(b_{12}(S) - b_{22}(S)) - 2b_{13}(S)x^3\left(2a_1(S)x(a_1(S)x + a_2(S)) - 2a_1(S)\sqrt{tx}(b_{22}(S) - 2b_{12}(S)) + t\right) + \right. \\ & \left. + b_{13}(S)^2\left(a_1(S)\sqrt{tx^5} + 3a_2(S)\sqrt{tx^3} + 3b_{12}(S)tx^{3/2}\right) + 3b_{13}(S)^3t\right) + a_1(S)\left(8a_1(S)^2\sqrt{tx^9}(b_{22}(S) - b_{12}(S)) + \right. \\ & \left. + b_{13}(S)\left(4a_1(S)\left(a_1(S)\sqrt{tx^{15}} + 3a_2(S)\sqrt{tx^{13}}\right) - 12a_1(S)tx^{13/2}(b_{22}(S) - 2b_{12}(S)) + 15t^{3/2}x^{11/2}\right) - \right. \\ & - 6b_{13}(S)^2tx^4\left(-a_1(S)x + 3a_2(S) + 3b_{12}(S)\sqrt{t}\right) - 18b_{13}(S)^3t^{3/2}x^{5/2}\right) + 3b_{11}(S)^2t^{3/2}x^4\left(b_{13}(S)\left(3a_1(S)x^{5/2} + a_2(S)x^{3/2} - \right. \right. \\ & \left. \left. - (b_{22}(S) - 2b_{12}(S))\sqrt{tx^3}\right) + 2a_1(S)x^4(b_{22}(S) - b_{12}(S)) - b_{13}(S)^2\sqrt{t}\right) + b_{11}(S)^3t^2x^6\left(x^{3/2}(b_{22}(S) - b_{12}(S)) + 2b_{13}(S)\right), \end{aligned}$$

$$\begin{aligned}
 Q_2(\sqrt{x}, \sqrt{t}) = & 2\sqrt{t}\left(-3b_{11}(S)tx^2\left(4a_1(S)^2x^{13/2}(b_{12}(S) - b_{22}(S)) - 2b_{13}(S)x^3\left(2a_1(S)x(a_1(S)x + a_2(S)) - \right.\right.\right. \\
 & \left.\left.\left.- 2a_1(S)\sqrt{tx}(b_{22}(S) - 2b_{12}(S)) + t\right) + b_{13}(S)^2\left(a_1(S)\sqrt{tx^5} + 3a_2(S)\sqrt{tx^3} + 3b_{12}(S)tx^{3/2}\right) + 3b_{13}(S)^3t\right) + \\
 & \left. + a_1(S)\left(8a_1(S)^2\sqrt{tx^9}(b_{22}(S) - b_{12}(S)) + b_{13}(S)\left(4a_1(S)\left(a_1(S)\sqrt{tx^{15}} + 3a_2(S)\sqrt{tx^{13}}\right) - \right.\right.\right. \\
 & \left.\left.\left.- 12a_1(S)tx^{13/2}(b_{22}(S) - 2b_{12}(S)) + 15t^{3/2}x^{11/2}\right) - 6b_{13}(S)^2tx^4\left(-a_1(S)x + 3a_2(S) + 3b_{12}(S)\sqrt{t}\right) - 18b_{13}(S)^3t^{3/2}x^{5/2}\right) + \\
 & \left. + 3b_{11}(S)^2t^{3/2}x^4\left(b_{13}(S)\left(3a_1(S)x^{5/2} + a_2(S)x^{3/2} - (b_{22}(S) - 2b_{12}(S))\sqrt{tx^3}\right) + 2a_1(S)x^4(b_{22}(S) - b_{12}(S)) - b_{13}(S)^2\sqrt{t}\right) + \right. \\
 & \left. + b_{11}(S)^3t^2x^6\left(x^{3/2}(b_{22}(S) - b_{12}(S)) + 2b_{13}(S)\right)\right),
 \end{aligned}$$

$$P_1(\sqrt{x}, \sqrt{t}) = -3b_{13}(S)\sqrt{tx^{13}}(a_1(S)x + a_2(S)),$$

$$\begin{aligned}
 P_2(\sqrt{x}, \sqrt{t}) = & x^{5/2}(a_1(S)x + a_2(S))\left(-b_{13}(S)\left(4\left(4a_1(S)^2x^5 + 6a_1(S)a_2(S)x^4 + 7a_1(S)b_{11}(S)\sqrt{tx^9} + 3a_2(S)b_{11}(S)\sqrt{tx^7}\right) + \right.\right. \\
 & \left. + 24a_1(S)b_{12}(S)\sqrt{tx^4} + tx^3\left(10b_{11}(S)^2x + 12b_{11}(S)b_{12}(S)\sqrt{x} + 15\right)\right) + 6b_{13}(S)^2\left(-a_1(S)\sqrt{tx^5} + 3a_2(S)\sqrt{tx^3} + b_{11}(S)tx^2 + 3b_{12}(S)tx^{3/2}\right) - \\
 & \left. - 2x^{11/2}(b_{22}(S) - b_{12}(S))\left(2a_1(S)\sqrt{x} + b_{11}(S)\sqrt{t}\right)^2 + 18b_{13}(S)^3t\right).
 \end{aligned}$$

The form of coefficients  $a_1(S)$ ,  $a_2(S)$ ,  $b_{11}(S)$ ,  $b_{12}(S)$ ,  $b_{13}(S)$ ,  $b_{22}(S)$  depends on the landscape generation way. If we generate landscape as described before then we can approximate these parameters as follows

$$\begin{aligned}
 a(S) &= \alpha(S)a(20), \\
 b_1(S; x) &= \beta_1(S)b_1(20; x), \\
 b_2(S; x) &= \beta_2(S)b_2(20; x),
 \end{aligned}$$

where

$$\begin{aligned}
 \alpha(S) &= 0.000344002 \frac{N_{obst}(S)}{10} + 1.00018, \quad MAE < 0.001, \\
 \beta_1(S) = \beta_2(S) &= 1.0582 - 0.0581965 \sqrt{\frac{N_{obst}(S)}{10}} - 1, \quad MAE < 0.005.
 \end{aligned}$$

Results of such approximation are shown into the fig. 6,  $x = x_{max} = 48$ .

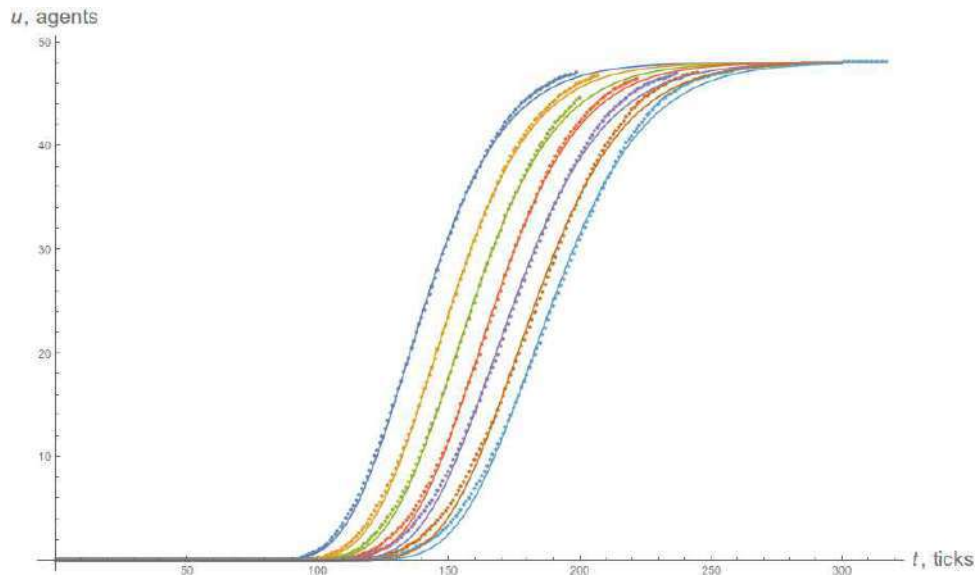


Fig. 6. The dependence  $u$  on the entropy  $S$ , from the left to the right  $N_{obst} = 20, N_{obst} = 40, N_{obst} = 60, N_{obst} = 80, N_{obst} = 100, N_{obst} = 120, N_{obst} = 140$ .

Note that we obtain different results by different landscape generation procedures. The general form of the function  $u$  will be still described by the (5), but exact values of parameters can be entirely different. For example, it is possible to place squares of different classes on the uniformly random way (fig. 2c). The comparison of functions  $u$  for a uniformly random landscape (right) and for landscape generated in the previously described way (left) are depicted on the fig. 7.

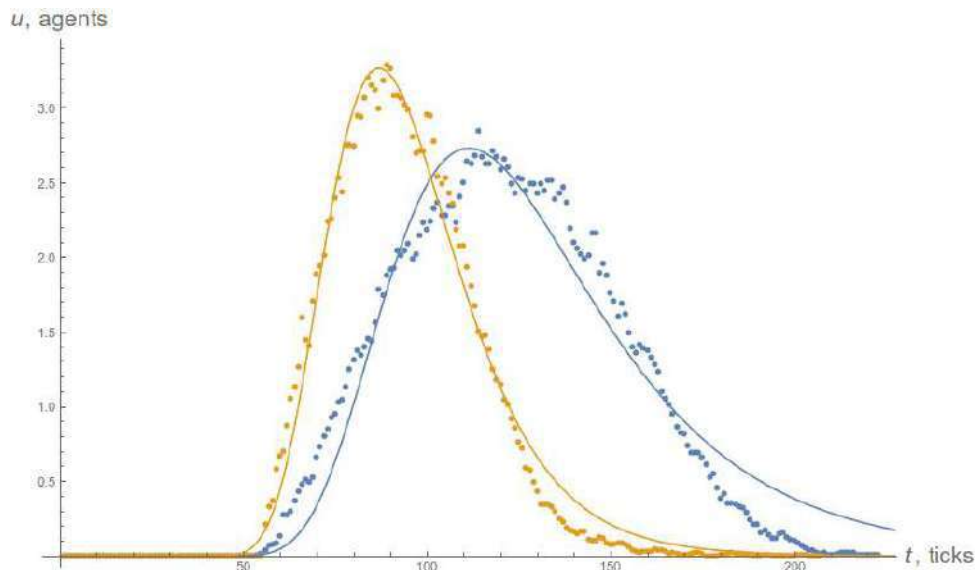


Fig. 7. Functions  $u$  for differently generated landscapes,  $x = 26$ ,  $S = 1.84877$ .

## 6. Conclusion

We obtained dependence of the win of movement by the proposed algorithm on the landscape's configuration entropy for some types of landscapes. The immediately following result may be a comparison of a model of the conflict based on the proposed cellular automaton with the result of solution of the corresponding Osipov-Lanchester equations. Also, we compare models of the "diffusion" of agents into a given sub-area based on the cellular automaton with the solution of the corresponding reaction-diffusion type equation. Finally, we can simulate the sharing of the subjective reality layers between agents. In this case, one agent will use the information about the area, received from other agents and will transmit such information to other agents itself. The algorithm described in the article can be applied to the mobile robot equipped with a transport base, navigation equipment (compass, GPS receiver, etc.), a sensor allowing to determine the impassibility of the terrain and the deciding unit, including memory.

## References

- [1] Kuznetsov, A. A model of the joint motion of agents with a three-level hierarchy based on a cellular automaton. *Computational mathematics and mathematical physics* 2017; 57(2): 340–349.
- [2] Kuznetsov A. Adaptive hierarchical system of the communication, control and decision support. Voronezh : NPF SAKVOEE LLC, 2016; 1: 269–277.
- [3] Malinetskii GG, Stepantsov ME. Application of cellular automata for modeling the motion of a group of people. *Computational Mathematics and Mathematical Physics* 2004; 44(11): 1992–1996.
- [4] Schumann A. Payo cellular automata and reflexive games. *J. Cellular Automata* 2014; 9(4): 287–313.
- [5] Cushman SA. Calculating the configurational entropy of a landscape mosaic. *Landscape Ecology* 2016; 31(3): 481–489. URL: <http://dx.doi.org/10.1007/s10980-015-0305-2>.
- [6] Ilachinski A. *ArtificialWar: Multiagent-Based Simulation of Combat*. Singapore : World Scientific Publishing Company, 2004; 747 p.
- [7] Tikhonov AN, Samarskii AA. *Equations of Mathematical Physics*. New York: Dover Publications, 1990; 800 p.



# Development of software system for analysis and optimization of taxi services efficiency by statistical modeling methods

Pavel Azanov<sup>1</sup>, Andrey Danilov<sup>2</sup>, Nikita Andriyanov<sup>3</sup>

<sup>1</sup>Tango Telecom, 122 Krasnaya street, 426057, Izhevsk, Republic Udmurtia, Russia

<sup>2</sup>Taxi Ulyanovsk, 1/3 Narimanova prospect, 432071, Ulyanovsk, Russia

<sup>3</sup>Ulyanovsk State Technical University, 32 Severniy Venets stret, 432027, Ulyanovsk, Russia

---

## Abstract

The text considers using of statistical models for taxi service data analysis and forecasting. Special attention is paid to the model parameters identification and short-term forecasting. We suggest to use the mathematical models of images to account the alternating character, associated with the dependence of the taxi orders number on various parameters. In addition the possibility of improving the effectiveness of evaluation by use of mixed random fields models is shown.

*Keywords:* random processes; mixed models; time series forecasting; taxi service; data analysis; image processing

---

## 1. Introduction

The following algorithm of the taxi service was quite common recently. Firstly, a dispatcher received the call, then the dispatcher communicated with a driver. During the communication the driver could accept the order or reject it. Usually all connections were provided by the radio devices. However, promising opportunities for use Internet in the taxi order service have appeared [1] due to the rapid Internet development. Now it is not difficult to order a taxi directly on the portal on the Internet or by using special applications for smartphones. In such cases, a very important source of receipt of orders from customers, that we will call customers from the phone, is not taken into account.

At the same time, it should be noted that such processing also provides a sufficient collection of statistics, the analysis of which may allow in the future to improve the quality of the taxi service. Increasing the volume of the telephone calls database warrants the possibility of analyzing the talk time, determining the most popular places in the city, etc. You will get a fairly complete statistical description of the taxi service operation adding to this statistics for orders, including the time of their execution, the waiting time of the car, the distribution by hours and other parameters.

Thus, there is an urgent task of analyzing an information collected in order to increase the efficiency of the service. So, for example, you can anticipate the number of dispatchers and drivers in advance by making precise forecasts of the calls number and tracking the orders percentage. In this case, both time series [2] and various models of random processes (RP) can be used to work with accumulated information [3,4].

## 2. Service architecture and statistics collection

Consider the taxi service project based on the contact center. At the same time, telephony is sent to operators through the Internet, and it requires only having a computer with a headset. To organize a dispatch taxi, you need a powerful software and hardware system. Its application allows several thousand taxi cars to work in real time.

Obviously, the use of this technology allows you to effectively manage resources, increase the speed of processing orders, always have exact customer numbers, reduce the time for applications.

For the contact center organization we need the presence of a multi-channel phone number, which will allow receiving many calls simultaneously. That's why we should use IP-telephony technologies. One of the most common telephony servers (PBX) is the Asterisk server [5], which allows to use SIP-telephony [6]. Such a telephone PBX should be set up to make calls distribution to taxi service operators. To process incoming calls we use a special program that represents the operator the form of a taxi order based on the Internet browser. To store information about calls, a database server is used, for example, MySQL or MsSQL server. Tariffs are set up using a separate module called Tarifficator. This module is programmed for its use in the web.

Thus, it is advisable to use virtualization methods to separate different servers, including a telephony server, a data base of telephony server, and a web server. In addition, an application server is needed. It provides information transfer from the contact center to the drivers. The special program for taxi service implements such transfer. And we suggest to use one more database server to store order information.

Fig. 1 presents full architecture of the considered taxi service.

The application for the Taxi program can have a version running just under java or common modern devices running by Android and iOS.

When a particular driver receives an order, the database is updated. The updates include information about the car, time of order picking, etc. These data can be used to inform the client about the assigned car.

The statistics is collected using database servers, but to present information in a convenient form it is necessary to use the Tari\_cator. Tari\_cator program allows you to display statistics either in a text document or in an excel format document. Fig. 2 presents the revised information on the distribution of orders, preserving the properties of the real sequence. We will make models fit according to this data.

It should be noted that the process in Fig. 2 has a heterogeneous structure, as well as some recurrent features. It is therefore necessary to select the most adequate model to more accurately describe all the peculiar distribution characteristics.

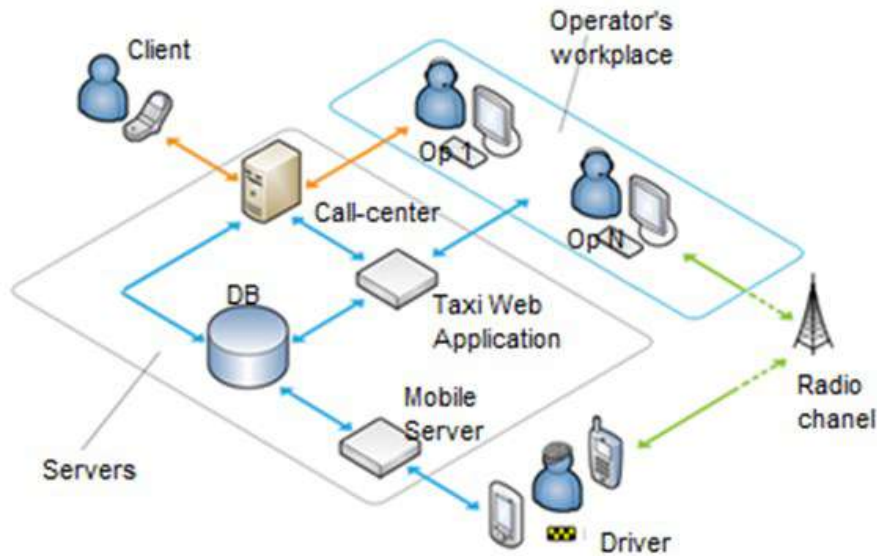


Fig. 1. Block diagram of the taxi service.

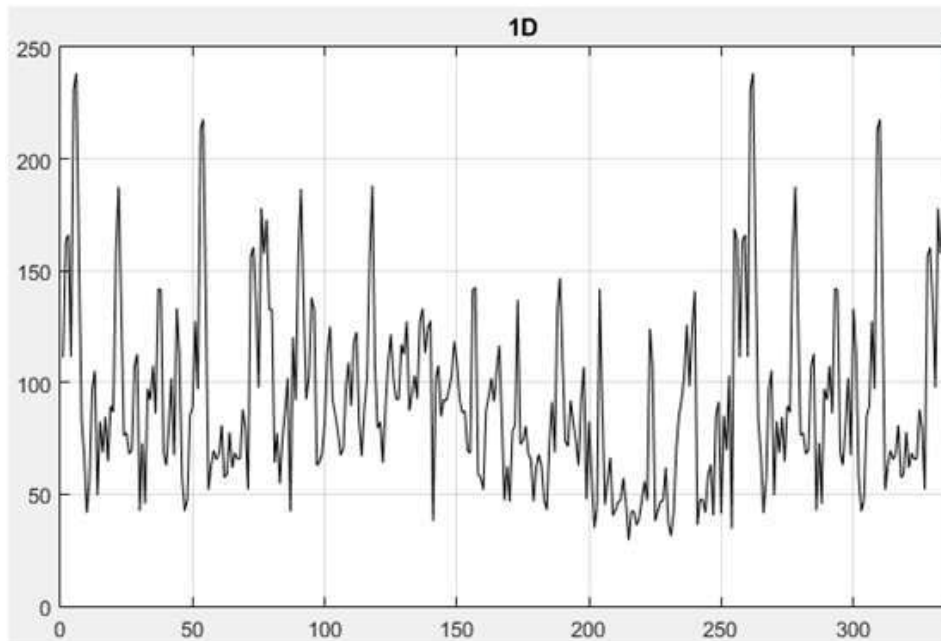


Fig. 2. Distribution of orders daily with the conversion (along the X axis is the number of orders, along the Y axis is the certain day).

### 3. Mathematical models for the presentation of taxi service statistics

Let's consider some variants of the description of the collected statistics on service. Let the data be collected from the beginning of the year (from January) and until the end of the year (to December) with some simplification, which will be used in the presentation approach in the form of an image.

#### 3.1. One-dimensional Autoregressive process

Let's imagine a sequence of data available on orders  $fOg$  using an expression for the Autoregressive (AR) of the first order

$$O_i = \rho O_{i-1} + \xi_i, i = 1, \dots, N \tag{1}$$

where  $\rho$  is a coefficient of correlation throughout the sequence and can easily be evaluated on the basis of existing data;  $\xi_i$  is

accidental admixture with zero mathematical expectation and variance  $\sigma_{\xi_i}^2 = \sigma_O^2(1-\rho^2)$ .

Besides the variance for orders is also estimated on the basis of the sample.

AR processes of higher orders can be used for a more accurate description. In this case, it is need to use the Yule-Walker equations [7] to determine the correlation parameters.

#### 3.2. One-dimensional doubly stochastic model of Random Process

Descriptions of the heterogeneity and periodicity of real data can be achieved using mixed models of Random Fields (RF). One of the variants to realize mixed models is the doubly stochastic model [8,9], whose correlation parameters also represent the implementation of the RF:

$$O_i = \rho_i O_{i-1} + \xi_i, i = 1, \dots, N \tag{2}$$

where  $\xi_i$  is the random additive value with zero mathematical expectation and variance  $\sigma_{\xi_i}^2 = \sigma_o^2(1-\rho_i^2)$ ;  $\rho_i$  is a sequence of correlation parameters

$$\rho_i = \tilde{\rho}_i + m_\rho, \tilde{\rho}_i = r\tilde{\rho}_{i-1} + \sqrt{\sigma_\rho^2(1-r^2)}\zeta_i, i = 1, \dots, N \quad (3)$$

where  $r$  is the constant correlation coefficient;  $m_\rho$  is the average value of the basic correlation coefficient;  $\sigma_\rho^2$  is the dispersion of the process describing change in the correlation parameters;  $\{\zeta_i\}$  is a field of Gaussian random variables with zero mathematical expectation and variance of unit.

An increase in the order of the process can also be used for the model (2) and its parameters (3), respectively. However, Fig. 1 shows the process which looks fairly "prickly". This fact allows the use of first-order models.

It is important that the estimation of all parameters of the model can be performed by mathematical statistics using the available sample, but also satisfactory results can be obtained with a slight increase in complexity, for example, in estimating all the parameters of the model in a sliding window [10] or using a nonlinear Kalman filter [11]. In addition, such algorithms can be adapted to different dimensionalities of the models.

### 3.3. Presentation in the form of a Random Field

The observed quasi-periodicity of the process shown in Fig. 2, allows us to conclude that it is possible to use models of random fields to represent information of this kind. Consider, for example, the doubly stochastic models of images that allow describing heterogeneous signals [12]. As an example, we will use the following model:

$$\begin{aligned} O_{i,j} = & 2\rho_{xi,j}O_{i-1,j} + 2\rho_{yi,j}O_{i,j-1} - 4\rho_{xi,j}\rho_{yi,j}O_{i-1,j-1} - \rho_{xi,j}^2O_{i-2,j} - \rho_{yi,j}^2O_{i,j-2} + \\ & + 2\rho_{xi,j}^2\rho_{yi,j}O_{i-2,j-1} + 2\rho_{xi,j}\rho_{yi,j}^2O_{i-1,j-2} - \rho_{xi,j}^2\rho_{yi,j}^2O_{i-2,j-2} + b_{i,j}\xi_{i,j}, i = 1, \dots, N_1, j = 1, \dots, N_2, \end{aligned} \quad (4)$$

where  $O_{i,j}$  is modeled RF with a normal distribution having  $M\{O_{i,j}\} = 0, M\{O_{i,j}^2\} = \sigma_o^2$ ;  $\{\xi_{i,j}\}$  is RF of independent standard Gaussian variables with  $M\{\xi_{i,j}\} = 0, M\{\xi_{i,j}^2\} = \sigma_\xi^2 = 1$ ;  $\rho_{xi,j}$  and  $\rho_{yi,j}$  are correlation coefficients of the model with multiple roots of characteristic equations of frequency rate (2,2) [13];  $b_{i,j}$  is a scale coefficient of simulated RF.

Random variables  $\rho_{xi,j}$ ;  $j$  and  $\rho_{yi,j}$  have the Gaussian probability distribution function and can be described by AR equations of the first order or higher orders.

It is easy to see that the model (4) is a transformation of the usual two-dimensional autoregressive model of the first order. This model of RF can also be used to describe a two-dimensional array of data and has the form:

$$\begin{aligned} O_{i,j} = & 2\rho_x O_{i-1,j} + 2\rho_y O_{i,j-1} - 4\rho_x\rho_y O_{i-1,j-1} - \rho_x^2 O_{i-2,j} - \rho_y^2 O_{i,j-2} + \\ & + 2\rho_x^2\rho_y O_{i-2,j-1} + 2\rho_x\rho_y^2 O_{i-1,j-2} - \rho_x^2\rho_y^2 O_{i-2,j-2} + b_{i,j}\xi_{i,j}, i = 1, \dots, N_1, j = 1, \dots, N_2. \end{aligned} \quad (5)$$

Note that the model (4), unlike the model with constant parameters (5), imitates heterogeneous in the structure of the RF, so it can fairly well reflect sharp surges on the number of orders on weekends and holidays. In order to estimate the parameters of such an image, we can use a vector (row-by-row) nonlinear Kalman filter. It requires to combine the elements of the image string into a vector  $\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{iN})$ . Then the model for a single frame of the image can be written as following equation:

$$\vec{x}_i = \text{diag}(\vec{\rho}_{xi})\vec{x}_{i-1} + \vartheta(\rho_{xi}, \rho_{yi})\vec{\xi}_i, \vec{\rho}_{xi} = r_{1x}\vec{\rho}_{x(i-1)} + \vartheta_{\rho_x}\vec{\xi}_{xi}, \vec{\rho}_{yi} = r_{1y}\vec{\rho}_{y(i-1)} + \vartheta_{\rho_y}\vec{\xi}_{yi},$$

where  $\text{diag}(\vec{\rho}_{xi})$  is the diagonal matrix with elements  $\vec{\rho}_{xi}$  on the main diagonal;  $\vartheta$  is down triangle matrix determined by the decomposition of covariance matrix:  $V_x = \vartheta\vartheta^T$ .

The evaluation process is described by the Kalman nonlinear filter:

$$\begin{aligned} \hat{\vec{x}}_{pi} = & \hat{\vec{x}}_{epi} + P_i \frac{\partial \Phi^T}{\partial \vec{x}_{pi}} V_n^{-1} (\vec{x}_i - \hat{\vec{x}}_{epi}), \\ \vec{x}_{pi} = & \begin{pmatrix} \vec{x}_i \\ \vec{\rho}_{xi} \\ \vec{\rho}_{yi} \end{pmatrix} = \Phi(\vec{\rho}_{x(i-1)}, \vec{x}_{i-1}) + \vartheta(\vec{\rho}_{x(i-1)}, \vec{\rho}_{y(i-1)})\vec{\xi}_i, \end{aligned}$$

where

$$\vec{x}_{epi} = \Phi(\vec{x}_{p(i-1)}), \Phi_p(\vec{x}_{p(i-1)}) = \begin{pmatrix} \Phi(\rho, x) \\ r_{1x}\vec{\rho}_{x(i-1)} \\ r_{1y}\vec{\rho}_{y(i-1)} \end{pmatrix}, \vec{\xi}_i = \begin{pmatrix} \xi_i \\ \xi_{xi} \\ \xi_{yi} \end{pmatrix}.$$

The use of this algorithm is possible if characteristics of information RF is exactly known, i.e. when we know the correlation coefficients  $r_{1x}, r_{2x}, r_{1y}, r_{2y}$ , as well as average values by row and column correlation, variance of correlation parameters and variance of information signal. Otherwise, a preliminary assessment of these parameters is required. Pseudogradient assessment procedures, as well as expressions for covariation function for doubly stochastic models can be used for this purpose. Produced at the output sequence of parameters can then be further parsed and replaced with any model. Also you can use and evaluation in the sliding window.

Fig. 3 shows the transformation of the original process to the image.

Thus, we see that the resulting image, on the one hand, is not strongly correlated, and on the other hand, there are several regions with higher brightness values on the image, which indicates the properties of the heterogeneity. We propose 6 variants of the models to describe the available data. Let's compare them in detail.

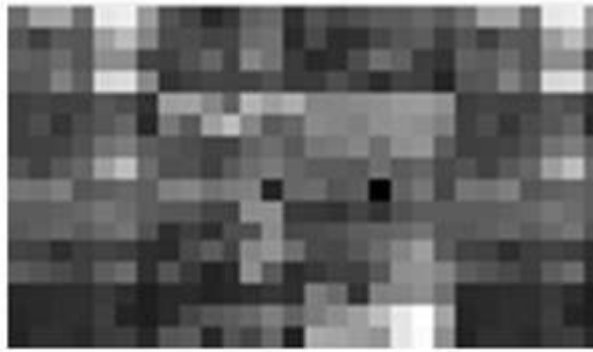


Fig. 3. Representation of orders statistics as an image.

**4. Comparative analysis of efficiency of prediction based on different models**

We will perform the necessary parameter estimation for models (1), (2), (4) and (5). So we produce forecasting the past 21 values of a sequence on the basis of models which was considered. It should be noted that the image data will be structured by seasons and weeks, as presented in Table 1.

Table 1. Data structure when converting it to image.

Month	January							February							March						
Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 1	Data							Data							Data						
Week 2																					
Week 3																					
Week 4																					
Month	April							May							June						
Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 1	Data							Data							Data						
Week 2																					
Week 3																					
Week 4																					
Month	July							August							September						
Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 1	Data							Data							Data						
Week 2																					
Week 3																					
Week 4																					
Month	October							November							December						
Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week 1	Data							Data							Data						
Week 2																					
Week 3																					
Week 4																					

The latter values will form a rectangular area in the lower right corner of the image, which is also useful for predicting and comparing the results of prediction based on various models. Denote the forecasting methods as follows:

- 1) A1 is the prediction based on one-dimensional AR model;
- 2) A2 is the prediction based on one-dimensional doubly stochastic model;
- 3) A2\* is the prediction based on one-dimensional mixed model with the evaluation parameters through the Kalman filter;
- 4) A3 is the prediction based on two-dimensional AR model;
- 5) A4 is the prediction based on two-dimensional doubly stochastic model;
- 6) A4\* is the prediction based on mixed model with evaluation parameters through the Kalman filter in two-dimension mode.

Fig. 4 presents the results of statistical modeling.

Relative variance of the prediction error of the last twenty one value, respectively, are as following:

- 1) It equals 10.88 for one-dimensional (1D) AR model;
- 2) It equals 0.254 for one-dimensional (1D) doubly stochastic model;
- 3) It equals 0.067 for one-dimensional (1D) doubly stochastic model with Kalman filter evaluation;
- 4) It equals 0.870 for two-dimensional (2D) AR model;
- 5) It equals 0.174 for two-dimensional (2D) doubly stochastic model;
- 6) It equals 0.049 for two-dimensional (2D) doubly stochastic model with Kalman filter evaluation.

Thus, analysis of the different models predicting results allows to say that using AR model leads to unsatisfactory results when predicting of complex data. Improving the effectiveness of predicting by the statistical models can be get using models of images. But such assessment will also not effective enough. So doubly stochastic models provide the best indicators because such models take into account the heterogeneity inherent in real data. Moving to the multivariate case leads to better forecast

because of the characteristics of the analyzed data set. In addition, the highest accuracy of prediction algorithms which were considered is provided by doubly stochastic models of the images. For such models estimation of parameters is performed using the Kalman filter.

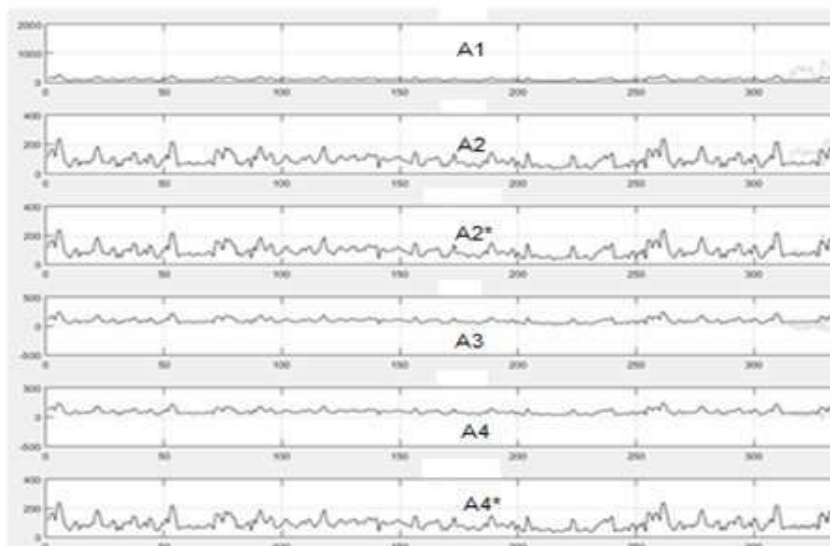


Fig. 4. Predicting the past values of the taxi service orders and real data (on X axis we have converted number of orders, on Y axis we have the certain day of the year).

## 5. Software package for statistical analysis of data on taxi service

As mentioned earlier the following algorithm of the taxi service is widely known. A dispatcher receives a call from the client and then communicates with a driver on the radio and transmits the order details to the driver. However, there is an alternative variant to this scheme of work at present time. The rapid growth of Internet traffic and the possibilities of IP-telephony (SIP), the availability of smartphones running under iOS or Android in each family allow you to abandon the use of radios when organizing a taxi order service.

At the same time, the task of optimizing the work of the taxi service is quite relevant, because this type of service is still in demand even during a finance crisis. Moreover, the opportunity to improve the efficiency of the service and save money by switching to automated mode is a very promising task.

Thus, the solution of this problem implies research at the junction of information and telecommunications systems. Indeed, it is necessary to realize not only communication networks that allow to exchange the information between operators, taxi drivers and customers, but also have software implementation of algorithms for handling calls and orders. At the same time, an important study is the statistical analysis of orders data.

We have solved the number of tasks during implementation of the software. First of all, the structure of the database has been developed. All the connections were thought out in the database, all the necessary information was collected. Secondly, it was suggested to use mixed or doubly stochastic autoregressive models to solve the orders forecasting problem. Third, we suggested a number of procedures based on the forecast data. We described how to calculate call traffic, determine the required number of operators. Fourthly, a Web-based interface has been developed that allows you to quickly change the settings on the telephony server.

The organization of the taxi service is performed on the basis of integration with the contact center (Telephony Server). Furthermore, the service includes:

- the database server;
- the Web server;
- the application server running the special Taxi program.

Fig. 5 shows the work of the Contact Center in more detail when the operator processes the order form. After such processing the database is updated and the order for taxi drivers is distributed.

Using the programming languages PHP and JavaScript, we developed the web-based interface for analyzing order data. As we mentioned earlier the interface can be conditionally called Tarifficator and allows you to obtain various statistical characteristics, as well as implement database modifications that are aimed at changing prices. In addition, you can view statistics on orders in real time using Tarifficator.

Another application, implemented by PHP and JavaScript, is the calculator of complex routes. The program allows you to calculate the cost of an order in the case when the driver passes several points in sequence. For example, cabbie drives first from point A to point B, and then from point B to point C.

The module for data analysis has been improved for convenience of the operating with different statistics in the languages PHP and JavaScript. This module (Fig. 6) allows to draw various statistical graphs using the library flot.js, and it also allows to make changes in the database related to setting prices. In addition, it is very important that in the Tarifficator module all necessary statistics is collected in real time mode.



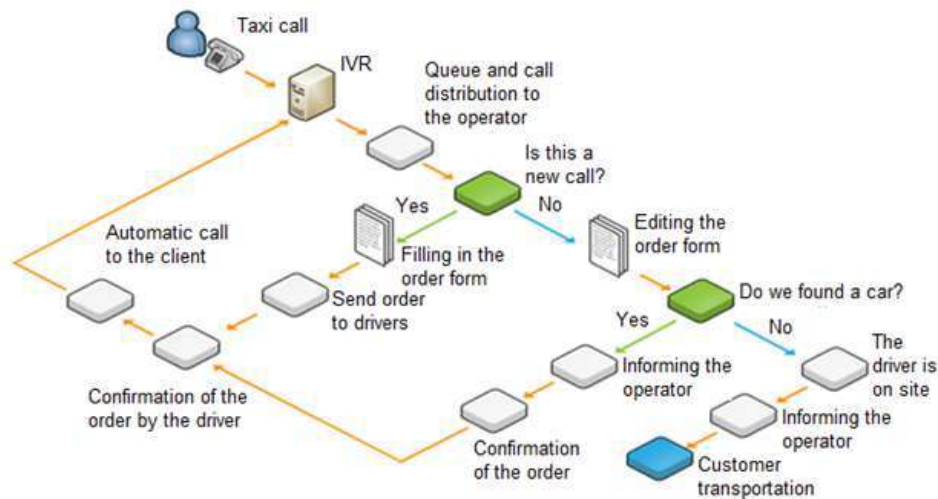


Fig. 5. Call processing algorithm.

You also can use module that allows the fitting of real data for operating the statistics module. So you can use statistical models of random sequences. Parameter identification may be implemented for the distribution of orders daily, calculated by the common AR model. Using these data the module will give forecast for the following days. Doubly stochastic models allow to consider the non-stationary in the distribution of data (bursts on weekends). For the such models you can use parameter identification algorithms based on a combination of algorithms of pseudogradient search and nonlinear Kalman filter [14].

The developed program complex allows you to accurately forecasting based on the doubly stochastic models of the images. Thus, improving the efficiency of taxi services is possible through the right choice of the necessary number of drivers in different time intervals. Similarly, it is possible to calculate, for example, the required number of call-center staff for different time periods.

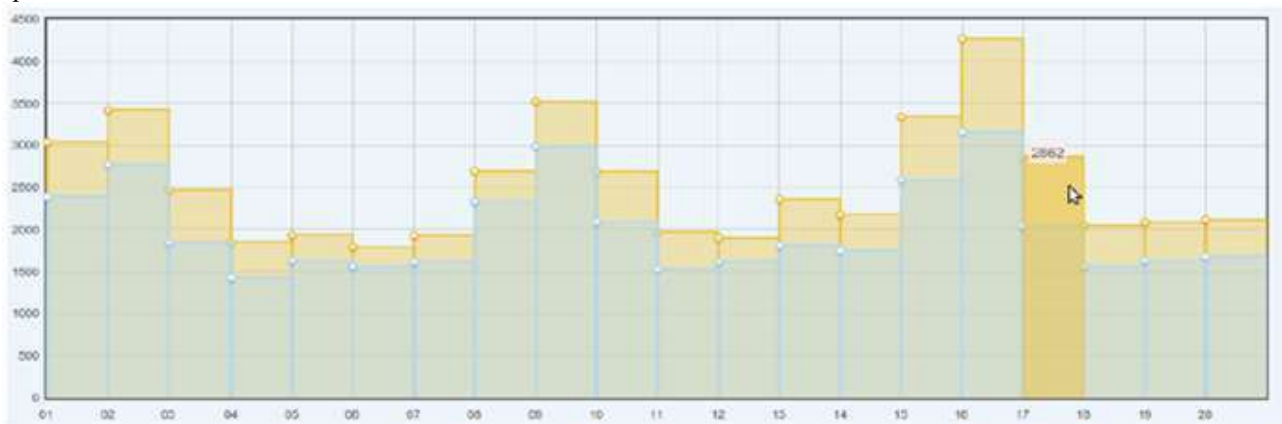


Fig. 6. Example of presenting statistics in the Tarifficator module.

Thus, you can get rid of radio communication and go to a software complex that handles data through the Internet. At the same time, telephony in the contact center means not the operators attached to the handset, but the people who process the data directly on the computer. In our project the dispatching taxi is organized with the help of the powerful software and hardware complex. Furthermore, thousands of cars and more can work simultaneously. So it is possible to completely abandon the use of a radio for a taxi service. Obviously, the use of such technology allows you to effectively manage resources, increase the speed of processing orders, always have exact customer numbers, reduce the time for applications running. And in order that the work of the taxi dispatcher was possible and necessary condition is the presence of a standard computer and a headset with a microphone.

## 6. Conclusion

The problem of analysis and optimization of the taxi order service efficiency is considered. It is suggested to use the doubly stochastic models of images to account for the heterogeneity of the data. A comparative analysis of forecasting based on 6 different models is carried out. In this case, the gain in comparison with autoregressive ones can reach several orders, and by applying the Kalman vector nonlinear filter it is possible to increase the forecast efficiency by another 4-5 times. A powerful software and hardware complex was developed. It will be used in the work of taxi order services and provide a solution to the task of real-time forecasting.

## Acknowledgements

This work was supported by the OOO "EIS-PFO" (Ulyanosk, Russia). We express our special gratitude for provided information used in the research.

## References

- [1] Andriyanov NA, Danilov AN. Taxi service with forecasting statistics based on complex mathematical models *Advances of modern science* 2016; 2(10): 114–116. (in Russian)
- [2] Yarushkina NG, Afanasyeva TV, Perfilieva IG. Time series mining. Students book. Ulyanovsk: UIGTU, 2010; 320 p. (in Russian)
- [3] Prokis J. Digital communications. Translated from eng. Edited by Klovskiy DD. Moscow: Radio and communications, 2000; 800 p.
- [4] Borovkov AA. Probability Theory. Springer Science and Business Media; 536 p.
- [5] Meggelen J, Madsen L, Smith J. Asterisk: future of the telephony. 2-nd edition, translated from eng. SPb: Symbol-Plus, 2009; 656 p.
- [6] Goldstein BS, Zarubin AA, Samorezov VV. Session Initiation Protocol (SIP): Reference book. Series: Telecommunication protocols of Russia, 2005; 456 p. (in Russian)
- [7] Andriyanov NA, Dement'ev VE. The application of the system of equations of the Yule-Walker to simulate isotropic random fields. *Modern trends of technical sciences. IV International Scientific Conference materials. Kazan, Russia, 2015: 2–6.* (in Russian)
- [8] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Doubly stochastic models of images. *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)* 2015; 25(1): 105–110. DOI: 10.1134/S1054661815010204.
- [9] Andriyanov NA. Doubly stochastic models based on the correlation interval changes. *Mathematical methods and models: theory, application and role in education* 2014; 3: 6–8. (in Russian)
- [10] Andriyanov NA. Method of fitting images based on random field model with changing parameters. *Advances of modern science* 2016; 5(9): 98–100. (in Russian)
- [11] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Application of mixed models for solving the problem on restoring and estimating image parameters. *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)* 2016; 26(1): 240–247. DOI: 10.1134/S1054661816010284.
- [12] Dement'ev VE, Andriyanov NA. The using of doubly stochastic models of random processes and fields to describe complex heterogeneous signals. *Actual problems of physical and functional electronics. Materials of 19-th all-Russian youth scientific schoolseminar. Ulyanovsk: UIGTU, 2016; 98–99.* (in Russian)
- [13] Vasiliev KK, Krashenninikov VR. Statistical image analysis. Ulyanovsk: UIGTU, 2014; 214 p. (in Russian)
- [14] Vasiliev KK, Dement'ev VE, Andriyanov NA. Parameter estimation of doubly stochastic random fields. *Radio* 2014; 7: 103–106. (in Russian)

# Detuning and dipole-dipole interaction effects on the entanglement of two qubits interacting with quantum fields of resonators

Eugene K Bashkirov<sup>1</sup>

<sup>1</sup> Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We investigate the entanglement dynamics between two dipole-coupled qubits interacting with vacuum or thermal fields of lossless resonators. Double Jaynes-Cummings model and two-atom Jaynes-Cummings model are considered taking into account detuning and direct dipole-dipole interaction. Using the dressed-states technique we derive the exact solutions for models under consideration. The computer modeling of the time dependence of qubit-qubit negativity is carried out for different strength of the dipole-dipole interaction and detuning. Results show that dipole-dipole interaction and detuning may be used for entanglement operating and controlling.

*Keywords:* Entanglement, Superconducting qubits, Detuning, Dipole-dipole interaction, Vacuum field, Thermal field

---

## 1. Introduction

Quantum computers are devices that store information on quantum variables such as spins, photons, and atoms, and that process that information by making those variables interact in a way that preserves quantum coherence. To perform a quantum computation, one must be able to prepare qubits in a desired initial state, coherently manipulate superpositions of a qubits two states, couple qubits together, measure their state, and keep them relatively free from interactions that induce noise and decoherence [1]. Qubits have been physically implemented in a variety of systems, including cavity quantum electrodynamics, superconducting qubits, atoms and ions in traps, quantum dots, spins and hybrid systems [2]. The connection between qubits can be arranged through their interaction with quantum fields of resonators. Basic protocols of quantum physics calculations are based on the use of entangled states [1]. Therefore, great efforts have been made to investigate entanglement characterization, entanglement control, and entanglement production in different systems. It is well known that the Jaynes-Cummings model (JCM) [3] is the simplest possible physical model that describes the interaction of a natural or artificial two-level atom (qubit) with a single-mode cavity [2], and has been used to understand a wide variety of phenomena in quantum optics and condensed matter systems, such as superconducting circuits, spins, quantum dots, atoms or ions in a cavity [2]. In order to explore a wider range of phenomena caused by the interaction of the qubits with the quantum fields in resonators the numerous generalizations of the JCM have been investigated in recent years (see references in [4]-[8]). Yöncü et al. [9] have proposed the so-called double JCM (DJCM), consisting of two two-level atoms and two resonator modes, provided that each atom interacts only with one field of the resonator, and investigated the pairwise entanglement dynamics of this model. Recently, the DJCM have been extensively investigated [10]-[17].

The direct dipole-dipole interaction between the qubits is the natural mechanism of entanglement producing and controlling. It's very important that the effective dipole-dipole interaction for superconducting Josephson qubits may be much greater than the coupling between the qubit and cavity field [18, 19]. The numerous references to the theoretical papers devoted to investigation of entanglement in two-qubit systems taking into account the dipole-dipole interaction are cited in our works [20]-[24]. In this paper, we considered two two-atom Jaynes-Cummings models taking into account the direct dipole-dipole interaction between qubits. We concerned our attention on two-atom double JCM and two-atom JCM with common cavity field. We investigated the entanglement between qubits, and



discussed the dependence of the entanglement on the parameters of the considered systems, such as the intensity of dipole-dipole interaction and the detuning between the atomic transition frequency and the cavity field frequencies.

## 2. Double Jaynes-Cummings model

In this section we consider two identical superconducting qubits labeled A and B, and two cavity modes of coplanar resonators labeled a and b. Qubit A not-resonantly interacts with a single-mode cavity field a, and qubit B not-resonantly interacts with a single-mode cavity field b. Due to the randomness of the qubits positions in the cavity, it is very difficult to control the couplings between different atom-cavity systems to be the same. Therefore the coupling constants between the atoms and cavities are assumed to be unequal. For superconducting qubits interacting with microwave coplanar resonators or LC superconducting circuits the intensity of effective dipole-dipole interaction can be compared with the atom-cavity coupling constant. In this case the dipole-dipole interaction should be included in the model Hamiltonian. Therefore the Hamiltonian for the system under rotating wave approximation can be written as

$$H = (\hbar\omega_0/2) \sigma_A^z + (\hbar\omega_0/2) \sigma_B^z + \hbar\omega_a a^\dagger a + \hbar\omega_b b^\dagger b + \hbar\gamma_a (\sigma_A^+ a + a^\dagger \sigma_A^-) + \gamma_b (\sigma_B^+ b + b^\dagger \sigma_B^-) + \hbar J (\sigma_A^+ \sigma_B^- + \sigma_A^- \sigma_B^+), \quad (1)$$

where  $(1/2)\sigma_i^z$  is the inversion operator for the  $i$ th qubit ( $i = A, B$ ),  $\sigma_i^+ = |+\rangle_{ii}\langle -|$ , and  $\sigma_i^- = |-\rangle_{ii}\langle +|$  are the transition operators between the excited  $|+\rangle_i$  and the ground  $|-\rangle_i$  states in the  $i$ th qubit,  $a^\dagger$  and  $a$  are the creation and the annihilation operators of photons of the cavity mode a,  $b^\dagger$  and  $b$  are the creation and the annihilation operators of photons of the cavity mode b,  $\gamma_a$  is the coupling constant between qubit A and the cavity field a and  $\gamma_b$  is the coupling constant between qubit A and the cavity field a,  $\delta_a = \omega_a - \omega_0$  and  $\delta_b = \omega_b - \omega_0$  are the detunings for mode a and b and  $J$  is the coupling constant of the dipole interaction between the qubits A and B. Here  $\omega_0$  is the qubit frequency and  $\omega_a$  and  $\omega_b$  are the frequencies of the cavities modes.

Firstly we take two qubits initially in the Bell-like pure state of the following form

$$|\Psi(0)\rangle_A = \cos\theta|+, -\rangle + \sin\theta|-, +\rangle, \quad (2)$$

where  $0 \leq \theta \leq \pi$  and the cavity fields initially are in vacuum states  $|0, 0\rangle = |0\rangle \otimes |0\rangle$ . We take into account that optimal temperature at which the superconducting qubits are used for quantum computing is mK. For such temperature the influence of thermal photons of the microwave cavity field on the dynamics of qubits can be neglected.

Then the full initial state is

$$|\Psi(0)\rangle = (\cos\theta|+, -\rangle + \sin\theta|-, +\rangle) \otimes |0, 0\rangle. \quad (3)$$

The evolution of the system is confined in the subspace  $|-, -, 0, 1\rangle, |-, -, 1, 0\rangle, |-, +, 0, 0\rangle, |+, -, 0, 0\rangle$ . To obtain the time-dependent wave function of considered model one can use the so-called dressed states or eigenvectors of the Hamiltonian (1). We have obtained these for general case when parameters of the Hamiltonian (1) take the arbitrary values. But the general expressions for eigenvectors are too cumbersome to display here. Therefore, we present below the eigenvectors and eigenvalues of the Hamiltonian (1) for special case when  $\delta_a = -\delta_b = \delta$  and  $\gamma_a = \gamma_b = \gamma$ .

In this case the eigenvectors of the Hamiltonian (1) in a frame rotating with the qubit frequency  $\omega_0$  can be written as

$$|\Phi_i\rangle = \xi_i (X_{i1}|-, -, 0, 1\rangle + X_{i2}|-, -, 1, 0\rangle + X_{i3}|-, +, 0, 0\rangle + X_{i4}|+, -, 0, 0\rangle) \quad (i = 1, 2, 3, 4),$$

where

$$\xi_i = 1 / \sqrt{|X_{i1}|^2 + |X_{i2}|^2 + |X_{i3}|^2 + |X_{i4}|^2}$$

and

$$X_{11} = \frac{2\alpha}{\alpha^2 + \Delta^2 - B + \sqrt{2}\Delta\sqrt{A-B}}, \quad X_{12} = \frac{\sqrt{2}}{\Delta\sqrt{2} - \sqrt{A-B}}, \quad X_{13} = \frac{-\alpha^2 - \Delta^2 + B + \sqrt{2}\Delta\sqrt{A-B}}{\alpha(-2\Delta + \sqrt{2}\sqrt{A-B})}, \quad X_{14} = 1;$$

$$X_{21} = \frac{2\alpha}{\alpha^2 + \Delta^2 - B - \sqrt{2}\Delta\sqrt{A-B}}, \quad X_{22} = \frac{\sqrt{2}}{\Delta\sqrt{2} + \sqrt{A-B}}, \quad X_{23} = \frac{\alpha^2 + \Delta^2 - B + \sqrt{2}\Delta\sqrt{A-B}}{\alpha(2\Delta + \sqrt{2}\sqrt{A-B})}, \quad X_{24} = 1,$$

$$\begin{aligned}
 X_{31} &= \frac{2\alpha}{\alpha^2 + \Delta^2 + B + \sqrt{2}\Delta\sqrt{A+B}}, & X_{32} &= \frac{\sqrt{2}}{\Delta\sqrt{2} - \sqrt{A+B}}, & X_{33} &= \frac{\alpha^2 + \Delta^2 + B - \sqrt{2}\Delta\sqrt{A+B}}{2\alpha\Delta - \sqrt{2}\alpha\sqrt{A+B}}, & X_{34} &= 1, \\
 X_{41} &= \frac{2\alpha}{\alpha^2 + \Delta^2 + B - \sqrt{2}\Delta\sqrt{A+B}}, & X_{42} &= \frac{\sqrt{2}}{\Delta\sqrt{2} + \sqrt{A+B}}, & X_{43} &= \frac{\alpha^2 + \Delta^2 + B + \sqrt{2}\Delta\sqrt{A+B}}{\alpha(2\Delta + \sqrt{2}\sqrt{A+B})}, & X_{44} &= 1,
 \end{aligned}$$

where  $\Delta = \delta/\gamma$ ,  $\alpha = J/\gamma$  and  $A = 2 + \alpha^2 + \Delta^2$ ,  $B = \sqrt{\alpha^4 + 4\Delta^2 + \Delta^4 - 2\alpha^2(-2 + \Delta^2)}$ .

The corresponding eigenvalues are

$$E_1 = -\hbar\gamma\sqrt{A-B}/\sqrt{2}, \quad E_2 = \gamma\hbar\sqrt{A-B}/\sqrt{2}, \quad E_3 = -\hbar\gamma\sqrt{A+B}/\sqrt{2}, \quad E_4 = \hbar\gamma\sqrt{A+B}/\sqrt{2}.$$

For entanglement modeling we can obtain the time dependent wave function

$$|\Psi(t)\rangle = e^{-iHt/\hbar}|\Psi(0)\rangle. \quad (4)$$

Using the eigenvalues and eigenvectors of Hamiltonian (1) and the initial state (3) we can derive from (4)

$$|\Psi(t)\rangle = C_1^{(1)}(t)|-, -, 0, 1\rangle + C_2^{(1)}(t)|-, -, 1, 0\rangle + C_3^{(1)}(t)|-, +, 0, 0\rangle + C_4^{(1)}(t)|+, -, 0, 0\rangle, \quad (5)$$

where

$$\begin{aligned}
 C_1^{(1)} &= \cos\theta Z_{11} + \sin\theta Z_{12}, & C_2^{(1)} &= \cos\theta Z_{21} + \sin\theta Z_{22}, \\
 C_3^{(1)} &= \cos\theta Z_{31} + \sin\theta Z_{32}, & C_4^{(1)} &= \cos\theta Z_{41} + \sin\theta Z_{42}
 \end{aligned}$$

and

$$\begin{aligned}
 Z_{11} &= e^{-iE_1t/\hbar} \xi_1 Y_{41} X_{11} + e^{-iE_2t/\hbar} \xi_2 Y_{42} X_{21} + e^{-iE_3t/\hbar} \xi_3 Y_{4n} X_{31} + e^{-iE_4t/\hbar} \xi_4 Y_{44} X_{41}, \\
 Z_{12} &= e^{-iE_1t/\hbar} \xi_1 Y_{31} X_{11} + e^{-iE_2t/\hbar} \xi_2 Y_{3n} X_{21} + e^{-iE_3t/\hbar} \xi_3 Y_{33} X_{31} + e^{-iE_4t/\hbar} \xi_4 Y_{34} X_{41}, \\
 Z_{21} &= e^{-iE_1t/\hbar} \xi_1 Y_{41} X_{12} + e^{-iE_2t/\hbar} \xi_2 Y_{42} X_{22} + e^{-iE_3t/\hbar} \xi_3 Y_{43} X_{32} + e^{-iE_4t/\hbar} \xi_4 Y_{44} X_{42}, \\
 Z_{22} &= e^{-iE_1t/\hbar} \xi_1 Y_{31} X_{12} + e^{-iE_2t/\hbar} \xi_2 Y_{32} X_{22} + e^{-iE_3t/\hbar} \xi_3 Y_{33} X_{32} + e^{-iE_4t/\hbar} \xi_4 Y_{34} X_{42}, \\
 Z_{31} &= e^{-iE_1t/\hbar} \xi_1 Y_{41} X_{13} + e^{-iE_2t/\hbar} \xi_2 Y_{42} X_{23} + e^{-iE_3t/\hbar} \xi_3 Y_{43} X_{33} + e^{-iE_4t/\hbar} \xi_4 Y_{44} X_{43}, \\
 Z_{32} &= e^{-iE_1t/\hbar} \xi_1 Y_{31} X_{13} + e^{-iE_2t/\hbar} \xi_2 Y_{32} X_{23} + e^{-iE_3t/\hbar} \xi_3 Y_{33} X_{33} + e^{-iE_4t/\hbar} \xi_4 Y_{34} X_{43}, \\
 Z_{41} &= e^{-iE_1t/\hbar} \xi_1 Y_{41} X_{14} + e^{-iE_2t/\hbar} \xi_2 Y_{42} X_{24} + e^{-iE_3t/\hbar} \xi_3 Y_{43} X_{34} + e^{-iE_4t/\hbar} \xi_4 Y_{44} X_{44}, \\
 Z_{42} &= e^{-iE_1t/\hbar} \xi_1 Y_{31} X_{14} + e^{-iE_2t/\hbar} \xi_2 Y_{32} X_{24} + e^{-iE_3t/\hbar} \xi_3 Y_{33} X_{34} + e^{-iE_4t/\hbar} \xi_4 Y_{34} X_{44},
 \end{aligned}$$

where  $Y_{ij} = \xi_j X_{ji}^*$ .

We also can consider an another type of Bell-like pure initial state of two qubits

$$|\Psi(0)\rangle_A = \cos\theta|+, +\rangle + \sin\theta|-, -\rangle.$$

For this initial atomic state and vacuum cavities fields the full initial state of the system is

$$|\Psi(0)\rangle = (\cos\theta|+, +\rangle + \sin\theta|-, -\rangle) \otimes |0, 0\rangle. \quad (6)$$

For initial state (6) the time-dependent wave function can be written in the form

$$\begin{aligned}
 |\Psi(t)\rangle &= C_1^{(2)}(t)|+, +, 0, 0\rangle + C_2^{(2)}(t)|+, -, 0, 1\rangle + C_3^{(2)}(t)|-, +, 1, 0\rangle + C_4^{(2)}(t)|+, -, 1, 0\rangle + \\
 &+ C_5^{(2)}(t)|-, +, 0, 1\rangle + C_6^{(2)}(t)|-, -, 2, 0\rangle + C_7^{(2)}(t)|-, -, 0, 2\rangle + C_8^{(2)}(t)|-, -, 1, 1\rangle + C_9^{(2)}(t)|-, -, 0, 0\rangle. \quad (7)
 \end{aligned}$$

The coefficients  $C_i(t)$  may be obtained by using the way which is described in previous case. But these are too cumbersome. Therefore, we will use below the numerical results for coefficients under consideration.

For two-qubit system described by the reduced density operator  $\rho_A(t)$ , a measure of entanglement or negativity can be defined in terms of the negative eigenvalues  $\mu_i^-$  of partial transpose of the reduced atomic density matrix  $\rho_A^{T_1}$  [26, 27]

$$\varepsilon = -2 \sum \mu_i^- \quad (8)$$

Using the wave functions (5) or (7) one can obtain the density operator for the whole system as

$$\rho(t) = |\Psi(t)\rangle\langle\Psi(t)| \quad (9)$$

Taking a partial trace over the field variable one can obtain from (9) the reduced atomic density operator in the two-qubit basis  $|+, +\rangle$ ,  $|+, -\rangle$ ,  $|-, +\rangle$ ,  $|-, -\rangle$  for initial state (3) in the form

$$\rho_A(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & V(t) & H(t) & 0 \\ 0 & H(t)^* & W(t) & 0 \\ 0 & 0 & 0 & R(t) \end{pmatrix} \quad (10)$$

The matrix elements of (10) are

$$V(t) = |C_4^{(1)}(t)|^2, \quad W(t) = |C_3^{(1)}(t)|^2, \quad R(t) = |C_1^{(1)}(t)|^2 + |C_2^{(1)}(t)|^2, \quad H(t) = C_4^{(1)}(t)C_3^{(1)}(t)^*.$$

The partial transpose of the reduced atomic density matrix (10) is

$$\rho_A^{T_1}(t) = \begin{pmatrix} 0 & 0 & 0 & H(t)^* \\ 0 & V(t) & 0 & 0 \\ 0 & 0 & W(t) & 0 \\ H(t) & 0 & 0 & R(t) \end{pmatrix} \quad (11)$$

From equation (11), we obtain four eigenvalues. Three of them are always positive. The eigenvalue  $\mu_4^- = 1/2(R - \sqrt{4H^2 + R^2})$  is always negative. As a result, the negativity can be written as

$$\varepsilon(t) = \sqrt{R(t)^2 + 4|H(t)|^2} - R(t) \quad (12)$$

The partial transpose of the reduced atomic density matrix  $\rho_A^{T_1}$  for initial state (6) has the form

$$\rho_A^{T_1}(t) = \begin{pmatrix} U_1(t) & 0 & 0 & \tilde{H}_1(t)^* \\ 0 & V_1(t) & H_1(t)^* & 0 \\ 0 & H_1(t) & W_1(t) & 0 \\ \tilde{H}_1(t) & 0 & 0 & R_1(t) \end{pmatrix} \quad (13)$$

where one can obtain with using (7)

$$U_1(t) = |C_1^{(2)}(t)|^2, \quad H_1(t) = C_1^{(2)}(t)C_9^{(2)}(t)^*, \quad \tilde{H}_1(t) = C_2^{(2)}(t)C_5^{(2)}(t)^* + C_4^{(2)}(t)C_3^{(2)}(t)^*,$$

$$V_1(t) = |C_2^{(2)}(t)|^2 + |C_4^{(2)}(t)|^2, \quad W_1(t) = |C_3^{(2)}(t)|^2 + |C_5^{(2)}(t)|^2, \quad R_1(t) = |C_6^{(2)}(t)|^2 + |C_7^{(2)}(t)|^2 + |C_8^{(2)}(t)|^2 + |C_9^{(2)}(t)|^2.$$

Two eigenvalues of matrix (13) may be negative. Then, the negativity can be written as a superposition of two terms

$$\varepsilon(t) = \sqrt{(U_1(t) - R_1(t))^2 + 4|\tilde{H}_1(t)|^2} - U_1(t) - R_1(t) + \sqrt{(V_1(t) - W_1(t))^2 + 4|H_1(t)|^2} - V_1(t) - W_1(t) \quad (14)$$

The first term is taken into account if and only if  $|\tilde{H}_1|^2 > U_1 R_1$  and the second term is taken into account if and only if  $|H_1|^2 > V_1 W_1$ .

### 3. Two-atom Jaynes-Cummings model

In this section we consider two-atom JCM with common thermal cavity field. We have two identical qubits A and B (spins, quantum dots etc.) non-resonantly interacting with common one-mode quantum electromagnetic field of resonator. As in a previous case we assume that the direct dipole-dipole interaction between qubits takes place. But in contrast with previous case we investigated the entanglement induced by a thermal field. In a frame rotating with the field frequency, the Hamiltonian for the system under rotating wave approximation can be written as

$$H = \hbar\delta\sigma_A^z + \hbar\delta\sigma_B^z + \hbar\gamma \sum_{i=A}^B (\sigma_i^+ a + a^+ \sigma_i^-) + \hbar J(\sigma_A^+ \sigma_B^- + \sigma_A^- \sigma_B^+). \quad (15)$$

Here  $\sigma_A^z$  and  $\sigma_B^z$  are the inversion operators for qubit A and B respectively,  $\delta = \omega - \omega_0$  is detuning, where  $\omega$  is the cavity field frequency and  $\omega_0$  is the atom transition frequency. The other notations are similar to these used in Section 2. Let us note that concurrence dynamics for system with Hamiltonian (15) without dipole-dipole interaction has been earlier investigated by Zhang [25].

We consider two type of initial atomic states: separable state  $|+, -\rangle$  (or  $|-, +\rangle$ ) and entangled state (2). The initial cavity mode state are assumed to be the thermal one-mode state  $\rho_F(0) = \sum_n p_n |n\rangle\langle n|$ , where the weight functions are  $p_n = \bar{n}^n / (1 + \bar{n})^{n+1}$ . Here  $\bar{n}$  is the mean photon number in a cavity mode,  $\bar{n} = (\exp[\hbar\omega_i/k_B T] - 1)^{-1}$ ,  $k_B$  is the Boltzmann constant and  $T$  is the equilibrium cavity temperature.

Before considering the interaction between two qubits and thermal field, it is straightforward to first study two qubits simultaneously interacting with Fock state. Suppose that the excitation number of the atom-field system is  $n$  ( $n \geq 0$ ). The evolution of the system is confined in the subspace

$$|-, -, n+2\rangle, \quad |+, -, n+1\rangle, \quad |-, +, n+1\rangle, \quad |+, +, n\rangle.$$

On this basis, the eigenfunctions of the Hamiltonian (15) can be written as

$$|\Phi_{in}\rangle = \xi_{in}(X_{i1n}|-, -, n+2\rangle + X_{i2n}|+, -, n+1\rangle + X_{i3n}|-, +, n+1\rangle + X_{i4n}|+, +, n\rangle) \quad (i = 1, 2, 3, 4),$$

where

$$\begin{aligned} X_{11n} &= 0, & X_{12n} &= -1, & X_{13n} &= 1, & X_{14n} &= 0, \\ X_{i1n} &= -\frac{2\sqrt{1+n}\sqrt{2+n}}{4+2n+2\Delta+E_{in}-2\Delta E_{in}-E_{in}^2}, & X_{i2n} &= -\frac{\sqrt{1+n}(2\Delta+E_{in})}{4+2n+2\Delta+E_{in}-2\Delta E_{in}-E_{in}^2}, \\ X_{i3n} &= -\frac{\sqrt{1+n}(2\Delta+E_{in})}{4+2n+2\Delta+E_{in}-2\Delta E_{in}-E_{in}^2}, & X_{i4n} &= 1 \quad (i = 2, 3, 4). \end{aligned}$$

The corresponding eigenvalues are

$$\begin{aligned} E_{1n} &= -\hbar\gamma \alpha, & E_{2n} &= (1/3) \hbar\gamma (\alpha + A_n/B_n + B_n), \\ E_{3n} &= (1/6) \hbar\gamma \operatorname{Re} [2\alpha - (1+i\sqrt{3})A_n/B_n + i(i+\sqrt{3})B_n], \\ E_{4n} &= (1/6) \hbar\gamma \operatorname{Re} [2\alpha + i(i+\sqrt{3})A_n/B_n - (1+i\sqrt{3})B_n]. \end{aligned}$$

Here

$$\begin{aligned} A_n &= 18 + 12n + \alpha^2 + 12\Delta^2, \\ B_n &= \left( \alpha^3 - 54\Delta + 9\alpha(3+2n-4\Delta^2) + \frac{1}{2} \sqrt{-4(18+12n+\alpha^2+12\Delta^2)^3 + 4(\alpha^3-54\Delta+9\alpha(3+2n-4\Delta^2))^2} \right)^{1/3}. \end{aligned}$$

To derive the full dynamics of our model one can consider also the basis states  $|-, -, 1\rangle, |+, -, 0\rangle, |-, +, 0\rangle$ .

Assume that the system is initially prepared in the state  $|+, -, n\rangle$  ( $n \geq 0$ ), then at time  $t$ , the whole system will evolve to

$$|\Psi(t)\rangle = Z_{12,n}|-, -, n+2\rangle + Z_{22,n}|+, -, n+1\rangle + Z_{32,n}|-, +, n+1\rangle + Z_{42,n}|+, +, n\rangle. \quad (16)$$

Here

$$\begin{aligned}
 Z_{12,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{21n} X_{11n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{22n} X_{21n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{23n} X_{31n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{24n} X_{41n}, \\
 Z_{22,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{21n} X_{12n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{22n} X_{22n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{23n} X_{32n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{24n} X_{42n}, \\
 Z_{32,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{21n} X_{13n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{22n} X_{23n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{23n} X_{33n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{24n} X_{43n}, \\
 Z_{42,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{21n} X_{14n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{22n} X_{24n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{23n} X_{34n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{24n} X_{44n},
 \end{aligned}$$

where  $Y_{ijn} = \xi_{jn} X_{jin}^*$ .

If the initial state of our system is  $|+, -, 0\rangle$ , the time dependent wave function takes the form

$$|\Psi(t)\rangle = Z_{12}|-, -, 1\rangle + Z_{22}|+, -, 0\rangle + Z_{32}|-, +, 0\rangle, \quad (17)$$

where

$$\begin{aligned}
 Z_{12} &= -2ie^{-(\alpha-2\Delta)t/2} \sin(\Omega t/2)/\Omega, \quad Z_{22} = e^{-i(\alpha-2\Delta)t/2} \left( e^{i(3\alpha-2\Delta)t/2} + \Omega \cos(\Omega t/2) - 2t \sin(\Omega t/2) \right) / (2\Omega), \\
 Z_{32} &= e^{-i(\alpha-2\Delta)t/2} \left( -e^{i(3\alpha-2\Delta)t/2} + \Omega \cos(\Omega t/2) - 2t \sin(\Omega t/2) \right) / (2\Omega)
 \end{aligned}$$

and  $\Omega = \sqrt{8 + (\alpha + 2\Delta)^2}$ .

For initial state  $|-, +, n+1\rangle$  ( $n \geq 0$ ) the time-dependent wave function is

$$|\Psi(t)\rangle = Z_{13,n}|-, -, n+2\rangle + Z_{23,n}|+, -, n+1\rangle + Z_{33,n}|-, +, n+1\rangle + Z_{43,n}|+, +, n\rangle. \quad (18)$$

Here

$$\begin{aligned}
 Z_{13,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{31n} X_{11n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{32n} X_{21n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{33n} X_{31n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{34n} X_{41n}, \\
 Z_{23,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{31n} X_{12n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{32n} X_{22n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{33n} X_{32n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{34n} X_{42n}, \\
 Z_{32,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{31n} X_{13n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{32n} X_{23n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{33n} X_{33n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{34n} X_{43n}, \\
 Z_{42,n} &= e^{-iE_{1n}t/\hbar} \xi_{1n} Y_{31n} X_{14n} + e^{-iE_{2n}t/\hbar} \xi_{2n} Y_{32n} X_{24n} + e^{-iE_{3n}t/\hbar} \xi_{3n} Y_{33n} X_{34n} + e^{-iE_{4n}t/\hbar} \xi_{4n} Y_{34n} X_{44n}.
 \end{aligned}$$

If the initial state is  $|-, +, 0\rangle$ , the time dependent wave function takes the form

$$|\Psi(t)\rangle = Z_{13}|-, -, 1\rangle + Z_{23}|+, -, 0\rangle + Z_{33}|-, +, 0\rangle, \quad (19)$$

where  $Z_{13} = Z_{12}$ ,  $Z_{23} = Z_{22}$ ,  $Z_{33} = Z_{32}$ .

Now we go back to the theme of this Section. Using the equations (16)-(19) one can obtain the density operator for the whole system. Taking a partial trace over the field variables one can obtain the reduced atomic density operator and partial transpose of the reduced atomic density matrix  $\rho_A^{T_1}$ . For initial atomic state  $|+, -\rangle$  the partial transpose of the reduced atomic density operator has the form

$$\rho_A^{T_1}(t) = \begin{pmatrix} U_2(t) & 0 & 0 & H_2(t)^* \\ 0 & V_2(t) & 0 & 0 \\ 0 & 0 & W_2(t) & 0 \\ H_2(t) & 0 & 0 & R_2(t) \end{pmatrix}. \quad (20)$$

where

$$\begin{aligned}
 U_2(t) &= \sum_{n=0}^{\infty} p_n |Z_{42,n}(t)|^2, & V_2(t) &= \sum_{n=1}^{\infty} p_n |Z_{22,n-1}(t)|^2 + p_0 |Z_{22}(t)|^2, \\
 W_2(t) &= \sum_{n=1}^{\infty} p_n |Z_{32,n-1}(t)|^2 + p_0 |Z_{32}(t)|^2, & R_2(t) &= \sum_{n=1}^{\infty} p_n |Z_{12,n-1}(t)|^2 + p_0 |Z_{12}(t)|^2, \\
 H_2(t) &= \sum_{n=1}^{\infty} p_n Z_{22,n-1}(t) Z_{32,n-1}(t)^* + p_0 Z_{22}(t) Z_{32}(t)^*.
 \end{aligned} \tag{21}$$

Only one of the eigenvalues of matrix (20) may be negative. Therefore the negativity can be written in the form

$$\epsilon(t) = \sqrt{(|R_2(t)| - |U_2(t)|)^2 + 4|H_2(t)|^2} - |R_2(t)| - |U_2(t)|.$$

For initial atomic state  $|-, +\rangle$  the partial transpose of the reduced atomic density operator has the form (20). Its matrix elements may be obtained from (21) by replacing the coefficients  $Z_{i2n}$  with  $Z_{i3n}$ , where  $i = 1, 2, 3, 4$ . For entangled initial atomic state (2) the partial transpose of the reduced matrix also has the form (20). The elements of this matrix may be obtained by combining the elements of two partial transpose matrix for initial states  $|+, -\rangle$  and  $|-, +\rangle$ .

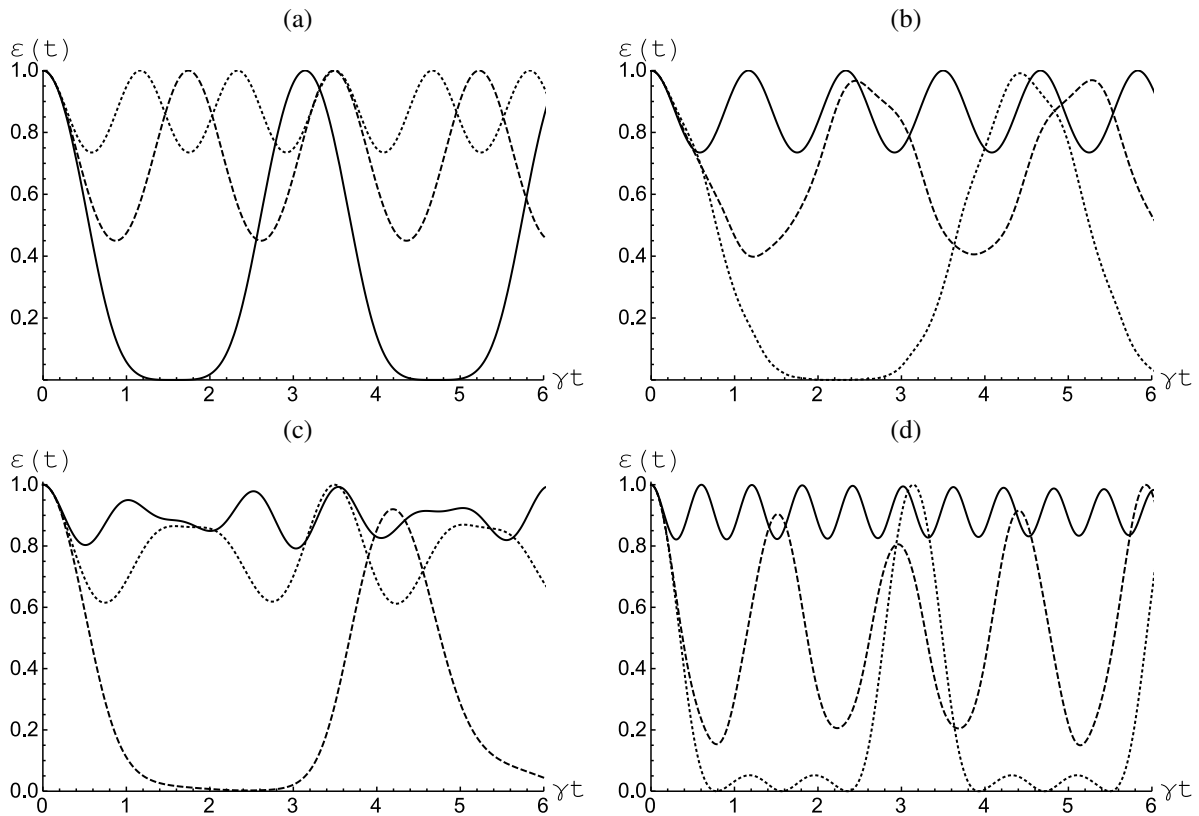


Figure 1: The negativity as a function of  $\gamma t$  for double JCM and initial state (3) with  $\theta = \pi/4$ . Parameters  $\delta_a = \delta_b = 0$ ,  $\gamma_b = \gamma_a$  (a),  $\delta_a = -\delta_b = 5$ ,  $\gamma_b = \gamma_a$  (b),  $\delta_a = \delta_b = 5$ ,  $\gamma_b = \gamma_a$  (c) and  $\delta_a = \delta_b = 0$ ,  $\gamma_a = 2\gamma_b$  (d). The strength of dipole interaction  $\alpha = 0$  (dotted),  $\alpha = 3$  (dashed) and  $\alpha = 5$  (solid).

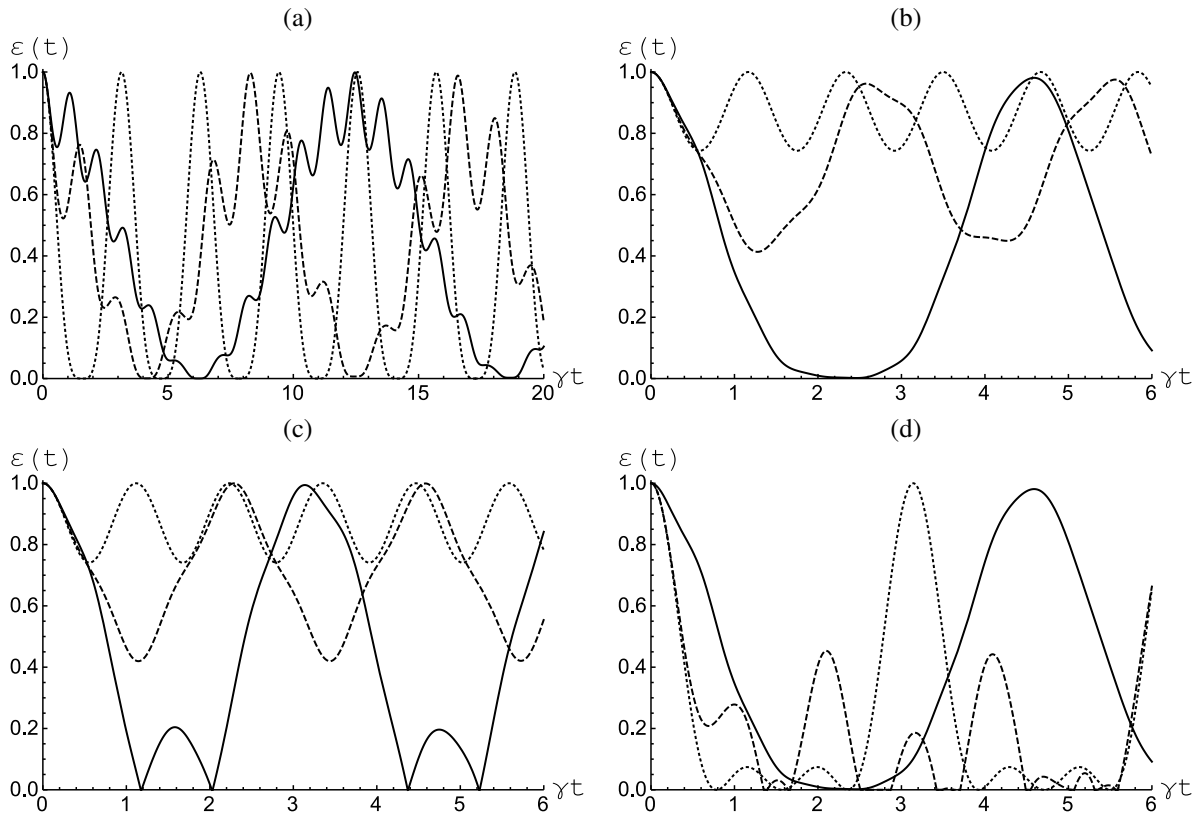


Figure 2: The negativity as a function of  $\gamma t$  for double JCM and initial state (6) with  $\theta = \pi/4$ . Parameters  $\delta_a = \delta_b = 0$ ,  $\gamma_b = \gamma_a$  (a),  $\delta_a = -\delta_b = 5$ ,  $\gamma_b = \gamma_a$  (b),  $\delta_a = \delta_b = 5$ ,  $\gamma_b = \gamma_a$  (c) and  $\delta_a = \delta_b = 0$ ,  $\gamma_a = 2\gamma_b$  (d). The strength of dipole interaction  $\alpha = 0$  (dotted),  $\alpha = 3$  (dashed) and  $\alpha = 5$  (solid).

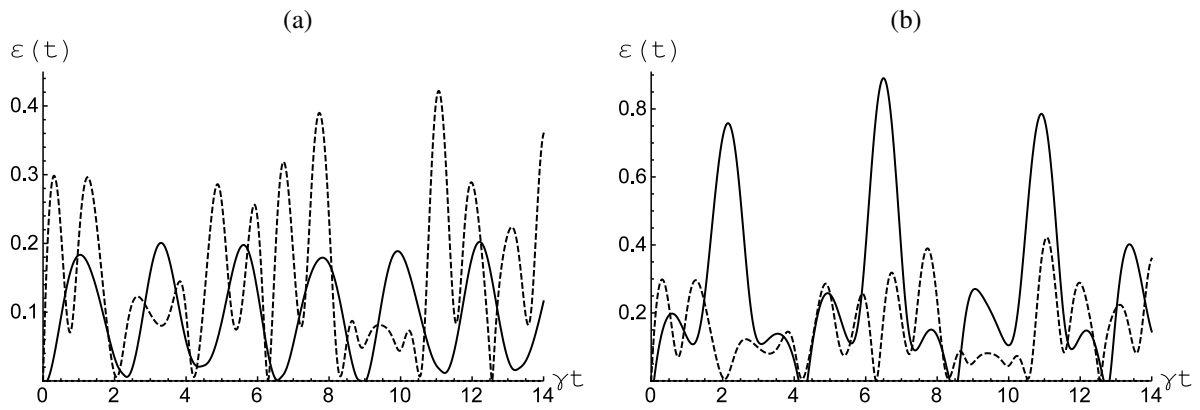


Figure 3: The negativity as a function of  $\gamma t$  for two-atom JCM with common field and initial atomic state  $|+, -\rangle$ . The strength of dipole interaction  $\delta = 0$  (a) and  $\delta = 0.5$  (b). The detuning  $\delta = 0$  (solid) and  $\delta = 1$  (dashed). Mean photon number  $\bar{n} = 0.1$ .

#### 4. Modeling of qubits entanglement dynamics

The results of calculations of entanglement parameter (12) for double JCM and initial state (3) are shown in Figs. 1(a)-(d). Results of calculations of entanglement parameter (14) for initial state (6) are displayed in Figs. 2(a-

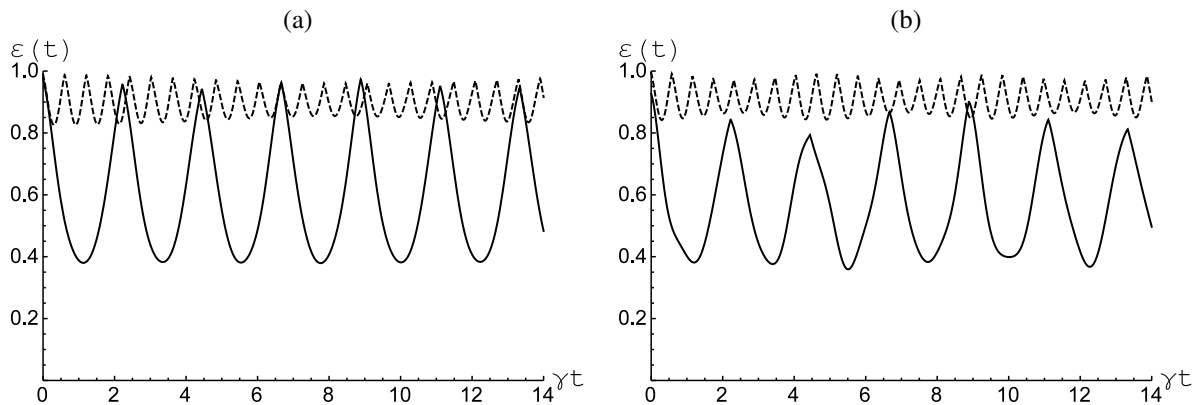


Figure 4: The negativity as a function of  $\gamma t$  for two-atom JCM with common field and entangled initial atomic state (2) with  $\theta = \pi/4$ . The strength of dipole interaction  $\alpha = 0$  (a) and  $\alpha = 0.5$  (b). The detuning  $\delta = 0$  (solid) and  $\delta = 5$  (dashed). Mean photon number  $\bar{n} = 0.1$ .

d). Fig. 1(a) shows that for exact resonance the negativity evolves periodically between 0 and 1, but the period is affected by the coupling constant. For resonance interaction the inclusion of the dipole-dipole interaction leads to a stabilization of entanglement behavior. Figs. 1(b)-1(d) show the effect of dipole-dipole interaction on negativity for non-resonant interaction and different couplings. When qubits A and B interact with a single-mode cavity fields via non-zero detuning the presence of dipole-dipole interaction with intermediate strength leads to increasing of the amplitudes of the negativity oscillations. But for large values of dipole-dipole interaction strength one can see the stabilization of entanglement oscillations as in the case of exact resonance. Figs. 2(a)-(d) show the time dependence of negativity for initial state (6) and different strength of dipole-dipole interaction. Fig. 2(a) gives the entanglement behavior for exact resonance. This behavior is different from that obtained for initial state (3) in resonance regime. The dipole-dipole interaction does not lead to stabilization of the entanglement, but has only an effect on the periods and amplitudes of the oscillations of entanglement. However, for non-resonant interaction between dipole-coupled qubits and fields the reverse behavior of atom-atom entanglement is true. For large values of the dipole-dipole interaction strength we have to deal with the stabilization of entanglement. Figs. 3 and 4 show the influence of detuning and dipole-dipole strength on atom-atom entanglement for two atoms interacting with common thermal field of resonator. Fig. 3(a) shows the entanglement time behavior for different couplings and separable atomic state  $|+, -\rangle$  ignored the dipole-dipole interaction. One can easily find that as the detuning increases, higher entanglement is obtainable. Zhang [25] earlier discovered such behavior and noted that when the atom-field detuning is large enough, the atoms tend to exchange energy with each other instead of with the field, and the field, which acts as a medium, is virtually excited during the atom-atom coupling process. Fig. 3(b) shows the negativity behavior for dipole-coupled qubits. For this case the reverse behavior of the entanglement is true. It seems like the negativity for qubits decreases as the detuning increases. We can also consider the negativity behavior for entangled initial state (2). In this case the inclusion of the detuning leads to a stabilization of entanglement behavior both to the model with dipole-dipole interaction and to the model without such interaction.

## 5. Conclusion

In this paper, we investigated the entanglement between two qubits interacting with fields of resonators in the framework of two types of JCM: double JCM with different coupling constants and detunings and two-qubit JCM with common cavity field taking into account the direct dipole-dipole interaction. For double JCM we discussed the influence of dipole-dipole interaction on qubit-qubit entanglement for resonance and non-resonance interactions. The results showed that these parameters have great impact on the amplitude and the period of the atom-atom entanglement evolution. In addition, the presence of sufficiently large dipole-dipole interaction leads to stabilization of entanglement for all Bell-type initial qubit states and different couplings and detuning. For two-qubit JCM with common field we



investigated the entanglement dynamics taking into account the dipole-dipole interaction for separable and entangled initial qubits states and thermal cavity field. For dipole coupled qubits prepared in a separable state the entanglement decreases as the detuning increases. For dipole uncoupled qubits the reverse behavior of the entanglement is true. For entangled initial states the inclusion of the detuning leads to a stabilization of entanglement behavior.

## References

- [1] Nielsen MA, Chuang IL. Quantum Computation and Quantum Information Cambridge: Cambridge University Press, 2000; 700 p.
- [2] Buluta I, Ashhab S, Nori F. Natural and artificial atoms for quantum computation. Rep. Prog. Phys. 2001; 74: 104401.
- [3] Scully MO, Zubairy MS. Quantum optics. Cambridge: Cambridge University Press, 1997; 630 p.
- [4] Bashkirov EK. Entanglement induced by the two-mode thermal noise. Laser Physics Letters 2006; 3(3): 145–150.
- [5] Bashkirov EK. Dynamics of the Two-Atom Jaynes-Cummings Model with Nondegenerate Two-Photon Transitions. Laser Physics 2006; 16: 1218–1226.
- [6] Bashkirov EK, Stupatskaya MP. The entanglement of two dipole-dipole coupled atoms induced by nondegenerate two-mode thermal noise. Laser physics 2009; 19: 525–530.
- [7] Bashkirov EK. Entanglement in the degenerate two-photon Tavis-Cummings model. Physica Scripta 2010; 82: 015401.
- [8] Bashkirov EK, Mastuygin MS. The dynamics of entanglement in two-atom Tavis-Cummings model with non-degenerate two-photon transitions for fourqubits initial atom-field entangled states. Optics Communications 2014; 313: 170–174.
- [9] Yonac MY, Yu T, Eberly JH. Sudden death of entanglement of two Jaynes Cummings atoms. J. Phys. B: At. Mol. Opt. Phys. 2006; 39: S621–S625.
- [10] Hu Y-H, Fang M-F, Cai J-W, Zeng K, Jiang C-L. Effect of the Stark shift on entanglement in a double two-photon JC model. J. Mod. Opt. 2008; 55(21): 3551–3562.
- [11] Hu Y-H, Fang M-F, Cai J-W, Zeng K, Jiang C-L. Sudden Death and Long-Lived Entanglement Between Two Atoms in a Double JC Model System. Int. J. Theor. Phys. 2008; 47: 2554–2565.
- [12] Du M, Fang M-F, Liu X. Sudden birth of entanglement between two atoms in a double JC model. Chin. Opt. Lett. 2009; 7(5): 443–445.
- [13] Xie Q, Fang M-F. Entanglement Dynamics of the Double Intensity-Dependent Coupling Jaynes-Cummings Models. Int. J. Theor. Phys. 2012; 51: 778–786.
- [14] Liao Q, Nie W, Zhou N, Liu Y, Ahmad MA. The Entanglement Dynamics of Two Atoms in a Double Two-Photon Jaynes-Cummings Model. Chin. J. Phys. 2013; 51(2): 404–411.
- [15] Vieira AR, de Oliveira Junior JGG, Peixoto de Faria JG, Nemes MC. Geometry in the Entanglement Dynamics of the Double Jaynes-Cummings Model. Braz. J. Phys. 2014; 44: 19–29.
- [16] Baghshahi HR, Tavassoly MZ, Faghghi MJ. Entanglement Criteria of Two Two-Level Atoms Interacting with Two Coupled Modes. Int. J. Theor. Phys. 2015; 54(8): 2839–2854.
- [17] Zhu W-T, Ren Q-B, Duan L-W, Chen Q-H. Entanglement Dynamics of Two Qubits Coupled Independently to Cavities in the Ultrastrong Coupling Regime: Analytical Results. Chin. Phys. Lett. 2016; 33(5): 050302(1–4).
- [18] Izmailkov A et al. Evidence for Entangled States of Two Coupled Flux Qubits. Phys. Rev. Lett. 2004; 93: 037003.
- [19] Majer JB et al. Spectroscopy on two coupled flux qubits. Phys. Rev. Lett. 2005; 94: 090501.
- [20] Bashkirov EK, Mastuygin MS. Entanglement of two superconducting qubits interacting with two-mode thermal field. Computer Optics 2013; 37(3): 278–285.
- [21] Bashkirov EK, Mastuygin M.S. The influence of the dipole-dipole interaction and atomic coherence on the entanglement of two atoms with degenerate two-photon transitions. Optics and Spectroscopy 2014; 116(4): 630–634.
- [22] Bashkirov EK, Mastuygin MS. The influence of atomic coherence and dipole-dipole interaction on entanglement of two qubits with nondegenerate twophoton transitions. Pramana - Journal of Physics 2015; 84(1): 127–135.
- [23] Bashkirov EK, Mastuygin MS. Entanglement Between Qubits Interacting with Thermal Field. EPJ Web of Conferences 2015; 103: 03002.
- [24] Bashkirov EK, Mastuygin MS. Entanglement between two qubits induced by thermal field. Journal of Physics: Conference Series 2016; 735: 012025.
- [25] Zhang B. Entanglement between two qubits interacting with a slightly detuned thermal field. Opt. Comm. 2010; 283: 4676–4679.
- [26] Peres A. Separability Criterion for Density Matrices. Phys. Rev. Lett. 1996; 77: 1413–1415.
- [27] Horodecki R, Horodecki M, Horodecki P. Separability of Mixed States: Necessary and Sufficient Condition. Phys. Lett. 1996; A223: 333–339.

# Reduction of flexible joint manipulator mathematical model

O.V. Vidilina<sup>1</sup>, N.V. Voropaeva<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoye shosse, 443086, Samara, Russia

## Abstract

The singularly perturbed differential systems which describe the dynamics of the manipulator with flexible joints are investigated under the condition of weak dissipation. The method of integral manifolds is used to construct the reduced model of robot. Integral manifolds may be constructed as an asymptotic power series. The simplified model is used to construct the control law for the robot with two flexible joints.

*Keywords:* mathematical model; integral manifolds; reduction; asymptotic methods

## 1. Introduction

The dynamic and control problems for robotic systems are connected with difficulties caused by high dimensions of models and availability of several time scales. Thereby the reduction problem ( the problem of the construction the lower order corrected models) is topical.

We investigate the model of  $n$ -links robot-manipulator with flexible joints where dissipation is small. The dynamics of such manipulators is described by quasi-oscillating singularly perturbed differential systems, which contain small parameter at the leading derivative. The conditions ensuring the possibility of using classical asymptotic methods are described in the well-known Tikhonov's theorem. The main of them is the asymptotic stability of the so-called boundary layer system. For investigated class of systems this condition of Tikhonov's theorem is not satisfied.

One of the approaches, which allows to reduce the complex multirate dynamic systems, is based on the theory of integral manifolds [1–14]. The conditions of the existence of an attractive slow integral manifold are investigated. This makes it possible to use the slow subsystem, which describes the motion on the manifold, as the simplified model of the manipulator. Similar questions for other classes of quasi-oscillating systems are studied in [10-13].

We consider the dynamic model of  $n$ -links robot with flexible joints Fig. 1.

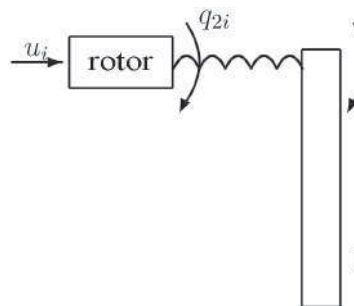


Fig. 1. The link of the manipulator.

The dynamics of manipulator is described by the system [15 – 18]

$$\begin{aligned} D(q_1)\ddot{q}_1 + c(q_1, \dot{q}_1) + K(q_1 - q_2) + B(\dot{q}_1 - \dot{q}_2) &= 0, \\ J\ddot{q}_2 - K(q_1 - q_2) - B(\dot{q}_1 - \dot{q}_2) &= u, \end{aligned} \quad (1)$$

where the coordinates of vectors  $q_1 \in R^n$  and  $q_2 \in R^n$  are the angles which characterize links and rotors positions respectively,  $D(q_1)$  is inertia matrix due to the links,  $J$  is diagonal inertia matrix of drive rotors, vector  $c(q_1, \dot{q}_1)$  is determined by coriolis, centrifugal and gravitation components. The flexibility of joint is represented by torsion spring with sufficiently large elastic coefficient. Let  $K = k \text{diag}(\tilde{K}_1, \dots, \tilde{K}_n)$  be the matrix of elastic coefficients,  $B = \text{diag}(B_1, \dots, B_n)$  be the matrix of damping ratios,  $u$  be the unit torque.

Let  $\mu = 1/k$  be a small positive parameter. Note that the more hard restriction on the matrix  $B$  was imposed in [15 – 18]. It was supposed that  $B_j = \tilde{B}_j/\mu$ , or  $B_j = \tilde{B}_j/\sqrt{\mu}$ . In fact, it was assumed that there is a sufficiently high dissipation. Such assumption guarantees the fulfillment of Tichonov's theorem condition about the asymptotic stability of the boundary layer system. Let us suppose that  $B_j = O(1)$ . Then the main condition of this theorem is not fulfilled.

**2. Reduction of the model**

Putting  $q = q_1$ ,  $z = k(q_1 - q_2)$  gives us the following system

$$\begin{aligned} \ddot{q} &= a_1(q, \dot{q}) + A_1(q)z + \mu A_3(q)\dot{z}, \\ \mu \ddot{z} &= a_2(q, \dot{q}) + A_2(q)z + \mu A_4(q)\dot{z} + M_2u, \end{aligned} \tag{2}$$

where

$$\begin{aligned} a_1(q, \dot{q}) &= a_2(q, \dot{q}) = -D^{-1}(q)c(q, \dot{q}), \quad A_1(q) = -D^{-1}(q)\widetilde{K}, \\ A_2(q) &= -(D^{-1}(q) + J^{-1})\widetilde{K}, \quad A_3(q) = -D^{-1}(q)B, \\ A_4(q) &= -(D^{-1}(q) + J^{-1})B, \quad M_2 = -J^{-1}. \end{aligned}$$

Using the coordinates  $x_1 = q$ ,  $x_2 = \dot{q}$ ,  $y_1 = z$ ,  $y_2 = \dot{z}$  we can rewrite system (2) to the form

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= a_1(x) + A_1(x_1)y_1 + \mu A_3(x_1)y_2, \\ \mu \dot{y}_1 &= \mu y_2, \\ \mu \dot{y}_2 &= a_2(x) + A_2(x_1)y_1 + \mu A_4(x_1)y_2 + M_2u(t, x, \mu). \end{aligned} \tag{3}$$

We obtained [14] the conditions for the existence of attractive slow integral manifold of system (3)

$$y = h(t, x, \mu), \tag{4}$$

where  $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ ,  $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ ,  $h(t, x, \mu) = \begin{pmatrix} h_1(t, x, \mu) \\ h_2(t, x, \mu) \end{pmatrix}$ .

Let function  $u(t, x, \mu)$  is represented in the form  $u(t, x, \mu) = u_0(t, x) + \mu u_1(t, x) + \dots$

Integral manifold (4) may be constructed as an asymptotic power series of the small parameter  $\mu$

$$y_i = h_i^{(0)}(t, x) + \mu h_i^{(1)}(t, x) + \mu^2 h_i^{(2)}(t, x) + \dots, \quad i = 1, 2 \tag{5}$$

with any degree of accuracy. Substituting (5) to the equations

$$\begin{aligned} \frac{\partial h_1}{\partial t} + \frac{\partial h_1}{\partial x_1}x_2 + \frac{\partial h_1}{\partial x_2}(a_1(x) + A_1(x_1)h_1 + \mu A_3(x_1)h_2) &= h_2, \\ \mu \left( \frac{\partial h_2}{\partial t} + \frac{\partial h_2}{\partial x_1}x_2 + \frac{\partial h_2}{\partial x_2}(a_1(x) + A_1(x_1)h_1 + \mu A_3(x_1)h_2) \right) &= \\ = a_2(x) + A_2(x_1)h_1 + \mu A_4(x_1)h_2 + M_2u(t, x, \mu), & \\ h_i &= h_i(t, x, \mu) \end{aligned} \tag{6}$$

and equating the coefficients at the same powers of  $\mu$  we can get  $h_i^{(j)} = h_i^{(j)}(t, x)$  for any  $j$ . In particular

$$\begin{aligned} h_1^{(0)} &= -A_2^{-1}(x_1)[a_2(x) + M_2u_0(t, x)], \quad h_2^{(0)} = \frac{\partial h_1^{(0)}}{\partial t} + \frac{\partial h_1^{(0)}}{\partial x_1}x_2 + \frac{\partial h_1^{(0)}}{\partial x_2}[a_1(x) + A_1(x_1)h_1^{(0)}], \\ h_1^{(1)} &= A_2^{-1}(x_1) \left[ \frac{\partial h_2^{(0)}}{\partial t} + \frac{\partial h_2^{(0)}}{\partial x_1}x_2 + \frac{\partial h_2^{(0)}}{\partial x_2}[a_1(x) + A_1(x_1)h_1^{(0)}] - A_4(x_1)h_2^{(0)} - M_2u_1(t, x) \right], \\ h_2^{(1)} &= \frac{\partial h_1^{(1)}}{\partial t} + \frac{\partial h_1^{(1)}}{\partial x_1}x_2 + \frac{\partial h_1^{(1)}}{\partial x_2}[a_1(x) + A_1(x_1)h_1^{(0)}] + \frac{\partial h_1^{(0)}}{\partial x_2}[A_1(x_1)h_1^{(1)} + A_3(x_1)h_2^{(0)}]. \end{aligned}$$

For  $h_i^{(j)}$ ,  $i = 1, 2$  from (6) we have

$$\begin{aligned} h_1^{(j)} &= A_2^{-1}(x_1) \left[ M_2u_j(t, x) - A_4(x_1)h_2^{(j-1)} - \frac{\partial h_2^{(j-1)}}{\partial t} - \frac{\partial h_2^{(j-1)}}{\partial x_1}x_2 - \frac{\partial h_2^{(j-1)}}{\partial x_2}[a_1(x) + A_1(x_1)h_1^{(0)}] - \right. \\ &\quad \left. - \sum_{s=0}^{j-2} \frac{\partial h_2^{(s)}}{\partial x_2}[A_1(x_1)h_1^{(j-s-1)} + A_3(x_1)h_2^{(j-s-2)}] \right], \\ h_2^{(j)} &= \frac{\partial h_1^{(j)}}{\partial t} + \frac{\partial h_1^{(j)}}{\partial x_1}x_2 + \frac{\partial h_1^{(j)}}{\partial x_2}[a_1(x) + A_1(x_1)h_1^{(0)}] + \sum_{s=0}^{j-1} \frac{\partial h_1^{(s)}}{\partial x_2}[A_1(x_1)h_1^{(j-s)} + A_3(x_1)h_2^{(j-1-s)}]. \end{aligned} \tag{7}$$

The system, which describes the motion on the slow integral manifold, is

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= a_1(x) + A_1(x_1)h_1(t, x, \mu) + \mu A_3(x_1)h_2(t, x, \mu). \end{aligned} \quad (8)$$

The dimension of this system is half of the dimension of the initial system. The slow subsystem has not fast variables, but nonetheless reliably describes the behavior of full system near the slow integral manifold. This allows to use it as a simplified model of flexible joints robot. The proposed approach to construct the reduced model is used in [11 – 13] to solve the problems of control and estimation for robot with one flexible joint.

### 3. Example

Let us consider the control problem for the robot with two flexible joints. The dynamics of manipulator is described by system (1), where [18]

$$\begin{aligned} D(q_1) &= \begin{pmatrix} \theta_1 + \theta_2 + 2\theta_3 \cos \varphi_2 & \theta_2 + \theta_3 \cos \varphi_2 \\ \theta_2 + \theta_3 \cos \varphi_2 & \theta_2 \end{pmatrix}, \quad q_1 = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \quad q_2 = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}, \\ \theta_1 &= m_1 l_{c_1}^2 + m_2 l_1^2 + I_1, \quad \theta_2 = m_2 l_{c_2}^2 + I_2, \quad \theta_3 = m_2 l_1 l_{c_2}, \quad \theta_4 = m_1 l_{c_1}, \quad \theta_5 = m_2 l_1, \quad \theta_6 = m_2 l_{c_2}, \\ c(q_1, \dot{q}_1) &= \theta_3 \sin \varphi_2 \begin{pmatrix} -2\dot{\varphi}_1 \dot{\varphi}_2 - \dot{\varphi}_2^2 \\ \dot{\varphi}_1^2 \end{pmatrix} + \begin{pmatrix} (\theta_4 + \theta_5)g \cos \varphi_1 + \theta_6 g \cos(\varphi_1 + \varphi_2) \\ \theta_6 g \cos(\varphi_1 + \varphi_2) \end{pmatrix}, \\ J &= \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}, \quad K = \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}, \quad B = \begin{pmatrix} b & 0 \\ 0 & b \end{pmatrix}. \end{aligned}$$

Using the coordinates  $x_1 = q$ ,  $x_2 = \dot{q}$ ,  $y_1 = z$ ,  $y_2 = \dot{z}$  we rewrite system (2) to the form (3), where

$$\begin{aligned} x &= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad x_1 = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \quad x_2 = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \end{pmatrix} = \begin{pmatrix} \dot{\varphi}_1 \\ \dot{\varphi}_2 \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \\ a_1(x) = a_2(x) &= \frac{(-\theta_3 \sin x_2^{(1)}(2x_1^{(2)} + x_2^{(2)})x_2^{(2)} + \theta_4 g \cos x_1^{(1)} + \theta_6 g \cos(x_1^{(1)} + x_2^{(1)}))}{\Delta} \begin{pmatrix} -\theta_2 \\ (\theta_2 + \theta_3 \cos x_2^{(1)}) \end{pmatrix} + \\ &+ \frac{(\theta_3 \sin x_2^{(1)}(x_1^{(2)})^2 + \theta_6 g \cos(x_1^{(1)} + x_2^{(1)}))}{\Delta} \begin{pmatrix} (\theta_2 + \theta_3 \cos x_2^{(1)}) \\ -(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)}) \end{pmatrix}, \quad \Delta = \theta_1 \theta_2 - \theta_3^2 \cos^2 x_2^{(1)}, \end{aligned}$$

$$A_1(x) = \frac{1}{\Delta} \begin{pmatrix} -\theta_2 & \theta_2 + \theta_3 \cos x_2^{(1)} \\ \theta_2 + \theta_3 \cos x_2^{(1)} & -(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)}) \end{pmatrix},$$

$$A_2(x) = \frac{k}{\Delta J_1 J_2} \begin{pmatrix} -J_2(\theta_2 J_1 + \Delta) & J_1 J_2(\theta_2 + \theta_3 \cos x_2^{(1)}) \\ J_1 J_2(\theta_2 + \theta_3 \cos x_2^{(1)}) & J_1(J_2(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)}) + \Delta) \end{pmatrix},$$

$$A_3(x) = \frac{b}{\Delta} \begin{pmatrix} -\theta_2 & J_1 J_2(\theta_2 + \theta_3 \cos x_2^{(1)}) \\ (\theta_2 + \theta_3 \cos x_2^{(1)}) & -(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)}) \end{pmatrix},$$

$$A_4(x) = \frac{b}{\Delta J_1 J_2} \begin{pmatrix} -J_2(\theta_2 J_1 + \Delta) & J_1 J_2(\theta_2 + \theta_3 \cos x_2^{(1)}) \\ J_1 J_2(\theta_2 + \theta_3 \cos(x_2^{(1)})) & J_1(J_2(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)}) + \Delta) \end{pmatrix}, \quad M_2 = \begin{pmatrix} -\frac{1}{J_1} & 0 \\ 0 & -\frac{1}{J_2} \end{pmatrix}.$$

The slow integral manifold (4) takes the form (5), where coefficients  $h_{j,k}^{(i)}$  are obtained from (7) by using the computer algebra system Maple. In particular

$$\begin{aligned} h_{1,1}^{(0)} &= -\frac{1}{kS}(u_1^{(0)}(\theta_1 \theta_2 + J_2 \theta_1 + J_2 \theta_2 - \theta_3^2 (\cos x_2^{(1)})^2 + 2J_2 \theta_3 \cos x_2^{(1)}) + u_2^{(0)} J_1 (\theta_2 + \theta_3 \cos x_2^{(1)}) - \\ &- J_1 \theta_3^2 (x_1^{(2)})^2 \cos x_2^{(1)} \sin x_2^{(1)} - J_1 \theta_3 \sin x_2^{(1)} (\theta_2 (x_1^{(2)} + x_2^{(2)})^2 + J_2 x_2^{(2)} (2x_1^{(2)} + x_2^{(2)})) + \\ &+ J_1 g (\theta_4 + \theta_5) \cos x_1^{(1)} (\theta_2 + J_2) + \cos(x_1^{(1)} + x_2^{(1)}) J_1 g \theta_6 (J_2 - \theta_3 \cos x_2^{(1)}), \\ h_{1,2}^{(0)} &= -\frac{1}{kS}(u_1^{(0)} J_2 (\theta_2 + \theta_3 \cos x_2^{(1)}) + u_2^{(0)} (\theta_1 \theta_2 + J_1 \theta_2 - \theta_3^2 (\cos x_2^{(1)})^2) - J_2 g (\theta_4 + \theta_5) \cos x_1^{(1)} (\theta_2 + \theta_3 \cos x_2^{(1)}) + \end{aligned}$$

$$\begin{aligned}
 & + J_2 g \theta_6 \cos(x_1^{(1)} + x_2^{(1)})(\theta_1 + J_1 + \theta_3 \cos x_2^{(1)}) + J_2 \theta_3^2 \cos x_2^{(1)} \sin x_2^{(1)} (2x_1^{(2)} x_2^{(2)} + (x_2^{(2)})^2 + 2(x_1^{(2)})^2) + \\
 & + J_2 \theta_3 \sin x_2^{(1)} ((\theta_1 + \theta_2)(x_1^{(2)})^2 + 2\theta_2 x_1^{(2)} x_2^{(2)} + \theta_2 (x_2^{(2)})^2 + J_1 (x_1^{(2)})^2), \\
 S & = (J_2 \theta_1 + \theta_1 \theta_2 + J_1 \theta_2 + J_2 \theta_2 + J_1 J_2 - \theta_3^2 (\cos x_2^{(1)})^2 + 2J_2 \theta_3 \cos x_1^{(1)}).
 \end{aligned}$$

The reduced system (8) takes the form

$$\begin{aligned}
 \dot{x}_1^{(1)} & = x_1^{(2)}, \quad \dot{x}_2^{(1)} = x_2^{(2)}, \\
 \dot{x}_1^{(2)} & = \frac{1}{S} (u_1^{(0)}(\theta_2 + J_2) - u_2^{(0)}(\theta_2 + \theta_3 \cos x_2^{(1)}) - g(\theta_4 + \theta_5) \cos x_1^{(1)}(\theta_2 + J_2) + \theta_3^2 \cos x_2^{(1)} \sin x_2^{(1)} (x_1^{(2)})^2 + \\
 & + \theta_3 \sin x_2^{(1)} (\theta_2 (x_1^{(2)} + x_2^{(2)})^2 + J_2 x_2^{(2)} (x_2^{(2)} + 2x_1^{(2)})) + g\theta_6 \cos(x_1^{(1)} + x_2^{(1)})(\theta_3 \cos x_2^{(1)} - J_2) + O(\mu), \\
 \dot{x}_2^{(2)} & = \frac{1}{S} (u_1^{(0)}(\theta_2 + \theta_3 \cos x_2^{(1)}) - u_2^{(0)}(\theta_1 + \theta_2 + 2\theta_3 \cos x_2^{(1)} + J_1) - \cos x_1^{(1)}(\theta_4 + \theta_5)g(\theta_2 + \theta_3 \cos x_2^{(1)}) + \\
 & + \theta_3 \sin x_2^{(1)} (\theta_2 x_2^{(2)} (x_2^{(2)} + 2x_1^{(2)}) + (x_1^{(2)})^2 (\theta_1 + \theta_2 + J_1)) + \theta_3^2 \cos x_2^{(1)} \sin x_2^{(1)} (x_1^{(2)} + x_2^{(2)})^2 + \\
 & + g\theta_6 \cos(x_1^{(1)} + x_2^{(1)})(\theta_3 \cos x_2^{(1)} + J_1 + \theta_1)) + O(\mu),
 \end{aligned}$$

We omitted here the terms containing  $\mu$  because of their bulkiness.

We construct the control law to move both links of manipulator to the fixed stable positions. We choose the control law based on the reduced system in accordance with the concept of linearizing feedback.

The Fig. 1 and Fig. 2 demonstrate the dynamics of the angles which characterize links positions with the control law formed by reduced system for the following parameters

$$m_1 = 10, m_2 = 5, l_1 = 1, l_2 = 1, l_{c_1} = 0.5, l_{c_2} = 0.5, J_1 = 1, J_2 = 1, \bar{K}_1 = 1, \mu = 0.01, \bar{K}_2 = 1.$$

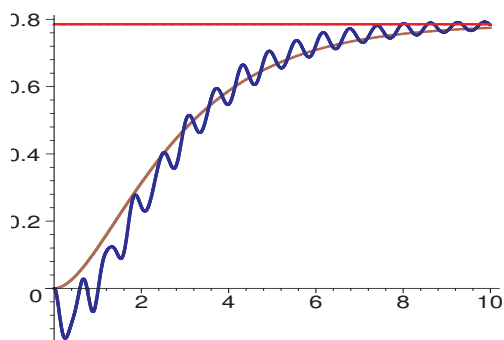


Fig. 1. The first link angle  $\varphi_1(t)$ .

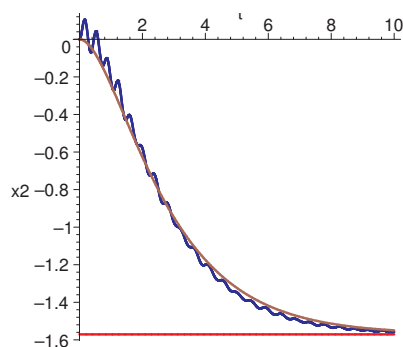


Fig. 2. The second link angle  $\varphi_2(t)$ .

It can be seen that the trajectories of initial system, which is characterized by damped high-frequency oscillations tend to the trajectories of reduced system, and those tend to the required fixed positions.

#### 4. Conclusion

The application of the integral manifolds method allows us to reduce the dimension and simplify the problem of the control law construction.

#### Acknowledgements

The research has been supported by the Russian Foundation for Basic Research and Government of the Samara region (grant 16-41-630524).

#### References

- [1] Sobolev VA. Integral manifolds and decomposition of singularly perturbed systems. Syst. & Control Lett. 1984; 5: 169–279.
- [2] Voropaeva NV, Sobolev VA. Geometric decomposition of singularly perturbed systems. Fizmatlit: Moscow, 2009. [in Russian]
- [3] Shchepakina E, Sobolev V, Mortell MP. Singular Perturbations: Introduction to System Order Reduction Methods with Applications. In: Springer Lecture Notes in Mathematics, Cham: Springer International Publishing, 2014.
- [4] Voropaeva NV, Sobolev VA. Decomposition of a linear-quadratic optimal control problem with fast and slow variables. Automation and Remote Control 2006; 67(8): 1185–1193.
- [5] Voropaeva NV. Decomposition of problems of optimal control and estimation for discrete systems with fast and slow variables. Automation and Remote Control 2008; 69(6): 920–928.

- [6] Sobolev VA. Singular perturbations in linearly quadratic optimal control problems. *Automation and Remote Control* 1991; 52(2): 180-189.
- [7] Vidilina OV, Voropaeva NV. The construction of the observers for dynamic systems with fast and slow variables. *CEUR Workshop Proceedings*, 2016; 1638: 750–758.
- [8] Smetannikova EN, Sobolev VA. Regularization of cheap periodic control problems. *Automation and Remote Control* 2005; 66(6): 903-916.
- [9] Strygin VV, Sobolev VA. Effect of geometric and kinetic parameters and energy dissipation on orientation stability of satellites with double spin. *Cosmic Research* 1976; 14(3): 366–371.
- [10] Strygin VV, Sobolev VA. Separation of motions by the method of integral manifolds. Nauka: Moscow, 1988. [in Russian]
- [11] Osintsev MS, Sobolev VA. Dimensionality Reduction in Optimal Control and Estimation Problems for Systems of Solid Bodies with Low Dissipation. *Automation and Remote Control* 2013; 74(8): 121–137.
- [12] Mortell MP, O'Malley R, Pokrovskii A, Sobolev V. *Singular Perturbation and Hysteresis*. SIAM: Philadelphia, 2005.
- [13] Aksenova NK, Sobolev VA. Control of a one rigid-link manipulator in the case of nonsmooth trajectory. *CEUR Workshop Proceedings* 2016; 1638: 493–497.
- [14] Vidilina OV, Voropaeva NV. Reduction of mathematical model of robot with elastic joints. *Vestnik SamGU. Estestvennonauchnaya seriya*. 2014; 3(114): 16–29.[in Russian]
- [15] Spong MW. Modeling and control of elastic joint robots. *Journal of Dynamic Systems, Measurement and Control* 1987; 109: 310–319.
- [16] Spong MW, Khorasani K, Kokotovic PV. An integral manifold approach to feedback control of flexible joint robots. *IEEE Journal of Robotics and Automation* 1987; 3(4): 291–301.
- [17] Moberg S. On modeling and control of flexible manipulators. Linkoping University: Linkoping, 2007.
- [18] Spong MW. On the robust control of robot manipulators. *IEEE Trans. Automatic Control* 1992; 37(11): 1782–1786.

# Model for constructing an option's portfolio with a certain payoff function

M.E. Fatyanova<sup>1</sup>, M.E. Semenov<sup>1</sup>

<sup>1</sup>Tomsk Polytechnic University, 30, Lenin ave., 634050, Russia, Tomsk

---

## Abstract

The portfolio optimization problem is a basic problem of financial analysis. In the study, an optimization model for constructing an option's portfolio with a certain payoff function has been proposed. The model is formulated as an integer linear programming problem and includes an objective payoff function and a system of constraints. In order to demonstrate the performance of the proposed model, we have constructed the portfolio on the European call and put options of Taiwan Futures Exchange. The optimum solution was obtained using the MATLAB software. Our approach is quite general and has the potential to design option's portfolios on financial markets.

*Keywords:* option strategies; payoff function; portfolio selection problem; combinatorial model; linear programming problem

---

## 1. Introduction

Interest to the options market steadily grows. In general case, brokers are creating financial portfolios based on a combination of standard European call and put options, cash, and the underlying assets itself, which are not associated with an investor's goal. Sometimes this goal is simply to insurance and hedge [1, 2], while, in other cases, the investor will wish to gain access to cash without currently paying tax [3] or the manager will can choice an investment technology [4] as well as speculative purposes [5]. The option's structures appeared in 1990's and became a popular tool of protection against falling prices [5, 6, 7]. Most of the company's hedges were conducted through option's portfolio (three-way collars), which involve selling a call, buying a put, and selling a put [1, 2, 7].

In the study [1] the theoretical model of zero-cost option's strategy in hedging of sales was demonstrated. In the model the options prices were evaluated by banks. There are seven different cases and it is shown that the put option was not exercised in either one of the researched cases. Therefore, it is difficult to talk about the positive effect of hedging.

In the study [2] authors provide an empirical analysis of the zero cost collar option contracts for commodity hedging and its financial impact analysis. Authors assessed option's portfolio as the hedging instrument on a quantitative basis using two scenarios in which assets prices are changed in the certain range.

In the study [8] authors proved that option's combinations are very popular on major option markets. They show the most popularly traded combinations in order of contract volume: straddles, ratio spreads, vertical spreads, and strangles. If European options were available with every single possible strike, any *smooth* payoff function could be created [9]. Authors [9] gives the decomposition formula for the replication of a certain payoff, was shown that any twice differentiable payoff function can be written as a sum of the payoffs from a static position on bonds, calls, and puts. Note that authors did not impose assumption regarding the stochastic price path. Analytical forms and graphs of the typical payoff profiles of option trading strategies can be found in publications [3, 10].

The portfolio optimization problem is a basic problem of financial analysis. In modern portfolio theory, developed by H. Markowitz, investors attempt to construct the portfolios by taking some alternatives into account: a) the portfolios with the lowest variance correspond to their preferred expected returns and vice versa, b) the portfolios with the highest expected returns correspond to their preferred variance. The Markowitz optimization is usually carried out by using historical data. The objective is to optimize the security's weight so that the overall portfolio variance is minimum for a given portfolio return. In this approach, options and structured products do not have a chance to be included into optimal mean-variance portfolios, but they have a place in optimal behavioral portfolios [9]. Theoretical option pricing models generally assume that the underlying asset return follows a normal distribution.

Another possible alternative is to apply some criterions to the payoff function as well as to the initial cost of a portfolio: for example, market risk measure [10, 11], probabilistic [12], fuzzy goal problem [13].

In the paper [11] authors have proposed a method to optimize portfolios without the normal distribution assumption of the portfolio's return. The objective function of the portfolio optimization problem is the expected return which is written as an integral over product of portfolio weights underlying assets and the joint density function of underlying asset returns. Also, the optimization problem includes constraints on short sale as well as the probability of not reaching thresholds.

In the study [5] authors have described a class of stock and option strategies, involving a long or short position in a stock, combined with a long or short position in an option. It was found that only the standard deviation, skewness, and kurtosis of the returns distribution of the underlying stock affected the optimal strategy, i.e. yield maximum returns. In the study [10] also, first four moments (mean, variance, skewness and kurtosis) were used to approximate the empirical distributions of the returns. Then the authors [10] have stated the multi-asset stochastic portfolio optimization model that incorporates European options and the

portfolio has a multi-currency structure. The objective function is to minimize the tail risk of the portfolio's value at the end of the strategy term,  $T$ . In the model, the Conditional Value-at-Risk (CVaR) metric of tail losses over the strategy term was used. The hedge option's portfolio is optimal in the sense of the CVaR metric. However, the proposed model depends on the quality of a scenario tree which additionally must be test on containing arbitrage opportunities.

Authors [10, 11] have found that optimal behavioral portfolios are likely to include the combination of derivative securities: put options, call options, and other structured products. Moreover, it must be noted that portfolios might include put options as well as call options on the same underlying assets.

In the paper [14] there is a proposed model for constructing a multi-period hedged portfolio which includes an European-type options. The objective function of optimization problem is recorded as the difference between the expected value of the portfolio and the expected regret of an investor. The model takes into account, the features of long-term investment: the risk aversion level is added into the objective function as well as the options contract with the different time to maturity were used.

In the paper [12] the two-step problem of optimal investment using the probability as an optimality criterion was studied. Various cases of distribution of returns were investigated. It was found that the structure of the optimal investment portfolio is almost identical despite of one or another distribution.

In the paper [13] a model for the construction of an option's hedged portfolio was proposed under a fuzzy objective function. The model does not explicitly takes transaction costs into account, but the entered membership functions are aimed at minimizing transaction lots. Thus, the authors have implicitly tried to reduce the potential transaction costs.

In the study [4] proposed a model based on using exotic options – lookback call options. Authors denoted that lookback calls have positive payoffs for both maximum and minimum asset's prices, and thus have features similar to a portfolio on call and put options.

In the paper [15] a computer-based system of identifying the informed trader activities in European-style options and their underlying asset was proposed, then the mathematical procedure of informed trader activity monitoring was built.

Previous studies considered the set of option's prices as a price function of an underlying asset:  $(1 + \delta) \times S_t$ , where  $|\delta| \leq 0.1$ . In contrast, in this paper the optimization is performed over the ask- and bid-prices and their combinations. We do not use the historical empirical distribution of returns.

Options are popular in Europe, USA, Russia, India and Taiwan [2, 4, 5, 8, 11, 15]. In the paper [8] authors have collected market data sets and shown that more than 55 percent of the trades of 100 contracts or larger are option's combinations and they account for almost 75 percent of the trading volume attributable to trades of 100 contracts or larger. In the numerical examples section of this article we will use the prices quoted on the Taiwan Futures Exchange (TAIFEX). According to the Report<sup>1</sup> in 2016 options amounted to almost 70 percent of total volume of derivative market in Taiwan.

The purpose of this study is to construct an option's strategy with the piecewise linear payoff function. With this in mind, the purpose for this study can be defined as the following problem statement: How does the personal investor's goal can be realized with an option's portfolio? In order to answer this, two research questions have been put forward.

- **Q1:** How can the method of establishment of the option's strategy be described from a theoretical perspective?
- **Q2:** How to validate a proposed method on option's market?

The remainder of this paper is organized as follows. In Section 2 we present the main definitions and assumptions which allow us to formulate the model as an optimization problem, including the objective function and the constraints for option's strategy. Section 3 then demonstrates the results of numerical experiments, including data description, and the integer solution of the optimization problem. Finally, Section 4 presents conclusions and future research.

## 2. Basic Definitions and Method

Option contracts were originally developed and put into circulation in order to reduce financial risks (hedging, insurance), associated with the underlying assets. In addition to existing standard option strategies [3] actively developing trading models, oriented to the objectives of a particular trader (speculative trading) [5], construction of synthetic positions [10], .

An *option* is a contract that will give an option holder a right to buy (or sell) the underlying asset. An *options premium* is the amount of money that investors pay for a call or put option.

A *call option* is a contract that will give its holder a right, but not the obligation, to purchase at a specified time, in the future, certain identified underlying assets at a previously agreed price. A *put option* is a security that will give its holder a right, but not the obligation, to sell at a specified time, in the future, certain identified underlying assets at a previously agreed price.

The *strike price* (exercise price) is defined as the price at which the holder of an options can buy (in the case of a call option) or sell (in the case of a put option) the underlying security when the option is exercised.

An *American option* may be exercised at the discretion of the option buyer at any time before the option expires. In contrast, a *European option* can only be exercised on the day the contract expires.

A *covered option* involves the purchase of an underlying asset (equity, bond or currency) and the writing a call option on that same asset. Short selling is the sale of a security that is not owned by a trader, or that a trader has borrowed.

<sup>1</sup>[http://www.taifex.com.tw/eng/eng3/eng3\\_3.asp](http://www.taifex.com.tw/eng/eng3/eng3_3.asp)



A *zero-cost option strategy* is an option trading strategy in which one could take a free options position for hedging or speculating in equity, forex and commodity markets.

There are main types of option's portfolio in real-world applications in terms of the time to expiration: American-, European-type options [16], and their combination.

The various option combinations represent strategies designed to exploit expected changes of the options values: the price of the underlying asset, its volatility, the time to expiration, the risk-free interest rate, the cost to enter [3, 8]. There is a large number of possible option combinations. When there are only two possible strike prices and two times to expiration we can design 36 combinations on one call and one put which may be either bought or sold. This number of combinations will increase significantly when any options values (strike prices, times to expiration, underlying assets) will be expanded insignificantly.

In this study we propose the strategy which involves European call and put options on the same underlying asset with the same maturity date  $T$ , but different strikes in a series. Let  $K_c = \{k_c^i \in \mathbb{Z}_{>0}, i \in I\}$  and  $K_p = \{k_p^i \in \mathbb{Z}_{>0}, i \in I\}$  be the call and put strikes,  $K_c, K_p$  are the increasing sequence of positive integers:

$$k_c^i < k_c^{i+1}, k_p^i < k_p^{i+1}, \forall i = 1, 2, \dots, n - 1,$$

$K = \{K_c \cup K_p\}$  is the set of unique strikes,  $\mathbb{Z}_{>0} = \{x \in \mathbb{Z} : x > 0\}$  denotes the set of positive integers. Let the number of call and put options be

$$X_c = \{x_i^c \in \mathbb{Z} : L \leq x_i^c \leq U, L < 0, U > 0, i \in I\},$$

$$X_p = \{x_i^p \in \mathbb{Z} : L \leq x_i^p \leq U, L < 0, U > 0, i \in I\},$$

with  $x_i^c, x_i^p > 0$  for buying,  $x_i^c, x_i^p < 0$  for selling, if  $x_i^c$  or  $x_i^p$  equal to 0 it means that the contract does not include in the portfolio,  $L$  and  $U$  represents the lower and upper bounds of the integer search space, respectively,  $I = \{1, 2, \dots, n\}$  is the set of indices, and  $S_t$  is a price of the underlying asset at calendar time,  $0 \leq t \leq T$ ,  $\hat{S}_T$  is an expected (forecasting) price of the underlying asset at the end of the strategy term,  $T$  (single period). Prices  $S_t, \hat{S}_T \in \mathbb{R}_{>0}$ , where  $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$  denotes the set of positive real numbers. We assume that the initial capital  $W$  is given and that no funds are added to or extracted from the portfolio,  $0 \leq t \leq T$ .

In order to determine the number of call and put options  $X = \{X_c, X_p\}$  for the implementation of the individual investor goal we propose the following assumptions that have impact on the payoff  $V(T, X)$  and an initial cost  $C(t, X)$  of portfolio:

- (i) the strategy should have protection on the downside and upside of strike prices,
- (ii) the strategy should effectively limit the upside earnings and downside risk with a maximal loss,  $\mathcal{L}$ ,
- (iii) the strategy should have the certain initial cost to enter  $C(t, X)$  at time  $t = 0$ .

### 2.1. Objective Function of Payoff

To establish the strategy we propose to use a combination long and short positions in put and call contracts based on the same underlying asset with different strike prices. We consider that one can take a static position (buy-and-hold), and the portfolio can include  $x_i^c, x_i^p$  units of European call and put options,  $i \in I$ . Its value at time  $t$  is given by the formula

$$V(T, X) = \sum_{i=1}^n x_i^c (S_t - k_c^i)^+ + x_i^p (k_p^i - S_t)^+, \tag{1}$$

the first term is the value of the call option payoff and the second is the value of the put option payoff, and

$$X^+ = \max(X, 0).$$

Let the best ask- and bid-prices for buying and selling of call and put options at time  $t = 0$  be

$$A_c = \{a_c^i \in \mathbb{R}_{>0}, i \in I\}, B_c = \{b_c^i \in \mathbb{R}_{>0}, i \in I\},$$

$$A_p = \{a_p^i \in \mathbb{R}_{>0}, i \in I\}, B_p = \{b_p^i \in \mathbb{R}_{>0}, i \in I\},$$

which we will name the *input constants*:

$$b_c^i < a_c^i, b_p^i < a_p^i, i \in I \text{ and}$$

$$a_c^i > a_c^{i+1}, a_p^i < a_p^{i+1}, b_c^i > b_c^{i+1}, b_p^i < b_p^{i+1},$$

$i = 1, 2, \dots, n - 1$ . The initial cost portfolio at time  $t = 0$  can be expressed as

$$C(t, X) = \sum_{i=1}^n x_i^c \cdot g_c(x_i^c) + x_i^p \cdot g_p(x_i^p), \tag{2}$$

where the functions  $g_c(x_i^c)$  and  $g_p(x_i^p)$  are defined as

$$g_c(x_i^c) = \begin{cases} a_c^i \in A_c, & \text{if } x_i^c > 0, \\ b_c^i \in B_c, & \text{if } x_i^c \leq 0, \end{cases} \tag{3}$$

$$g_p(x_i^p) = \begin{cases} a_p^i \in A_p, & \text{if } x_i^p > 0, \\ b_p^i \in B_p, & \text{if } x_i^p \leq 0, \end{cases} \quad (4)$$

$i \in I$ . Thus taking into account Eq. (1) and Eq. (2) the overall profit and loss at time  $T$  will be the final payoff minus the initial cost. So the objective function can be expressed as

$$F(X) = V(T, X) - C(t, X) = \sum_{i=1}^n x_i^c ((\hat{S}_T - k_c^i)^+ - g_c(x_i^c)) + x_i^p ((k_p^i - \hat{S}_T)^+ - g_p(x_i^p)), \quad (5)$$

which is a linear function of the *decision* variable  $X = \{X_c, X_p\}$ . The objective functional  $F(X)$  maps the entire stochastic process (cash flow) to a single real number

$$F(X) : \mathbb{Z}^n \mapsto \mathbb{R}, \text{ where } \mathbb{Z}^n \mapsto \mathbb{Z} \times \dots \times \mathbb{Z}.$$

### 2.2. Selection of Input Parameters

Using the conditional functions  $g_c(\cdot)$  and  $g_p(\cdot)$  in the objective function Eq. (5) leads us to solve a sequence of optimization problems. There are four input parameter values for each call and put options: ( $A_c$  or  $B_c$ ) and ( $A_p$  or  $B_p$ ). In this case, the number of permutations based on the selection between alternative prices (ask or bid) and possible contracts (call or put) equal to  $N = 2^n \times 2^n = 2^{2n}$ . In the numerical examples section of this article (Section 3) we will use the ask- and bid-prices quoted on the Taiwan Futures Exchange (TAIFEX).

Let  $C$  denote the set of all  $2 \times n$ -tuples of elements of given ordered sets of ask- and bid-prices  $A_c, B_c, A_p$  and  $B_p$ . The set  $C$  can be expressed as

$$C = \{(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n) : x_i = a_c^i \text{ or } b_c^i, y_i = a_p^i \text{ or } b_p^i, i \in I\}. \quad (6)$$

Thus  $C = \{c_1, c_2, \dots, c_N\}$  is the set of ordered permutations without replacement of two  $n$ -elements sets  $A_c, B_c$  and two  $n$ -elements sets  $A_p, B_p$ .

The calculation of the portfolio's terminal payoff under each price combination in the vector notation takes the form:

$$\max_X \{X_c^T ((\hat{S}_T - K_c)^+ - G_c(X_c)) + X_p^T ((K_p - \hat{S}_T)^+ - G_p(X_p))\} = \max_X \{F_C(X)\}, \quad (7)$$

where  $C$  denotes the set of ordered permutations of model input constants Eq. (6), and  $G_c(X_c), G_p(X_p)$  are the vector notation of conditional functions defined in Eq. (3) and Eq. (4). The objective function Eq. (7) maximizes the option's payoff over the holding period  $[0, T]$ .

### 2.3. System of Constraints

Each objective function  $F_C(X)$  from the set Eq. (7) is the piecewise linear function. Taking into account the assumptions mentioned in Section 2 we should determine the slope of the objective function Eq. (7) in the unique strike intervals. We separately investigate the intervals  $0 \leq S_T \leq k_1, k_2 \leq S_T \leq k_3, \dots, k_m \leq S_T < +\infty$ , here  $k_1 = \min(K_c, K_p)$  is the smallest strike and  $k_m = \max(K_c, K_p)$  is the largest strike.

The horizontal slope of the function (7) in the first closed interval  $[0, k_1]$  and the left-closed interval  $[k_m, +\infty)$  are specified respectively by:

$$\sum_{i=1}^n x_i^c = 0, \text{ if } S_T \in [0, k_1], \quad (8)$$

$$\sum_{i=1}^n x_i^p = 0, \text{ if } S_T \in [k_m, +\infty). \quad (9)$$

Positive and negative slopes of the function (7) in the interior intervals  $[k_q, k_{q+1}]$  are provided by:

$$\sum_{i:k_c^i \leq k_q} x_i^c - \sum_{j:k_p^j \geq k_{q+1}} x_j^p \text{ is } \begin{cases} \geq 0, & \text{if } k_q \leq k, \\ \leq 0, & \text{if } k_q > k, \end{cases} \quad (10)$$

here  $k \in K$  is an inflection point of the function (7).

The next balance constraint defines the bound of the downside risk with a maximal loss,  $\mathcal{L}$ , over the holding period  $0 \leq t \leq T$ :

$$V(T, X) = -\mathcal{L}. \quad (11)$$

The objective function value (7) at the terminal time  $T$  must be positive:

$$V(T, X) > 0, S_T = \hat{S}_T. \quad (12)$$

TAIEX:	8,067.60(13.91)					HIGH:	8,091.48	LOW:	8,003.93	Total Vol:	535,433	
TXO:	8,067.60(13.91)	UNDERLYING STATUS:	TC	HIGH:	8,091.48	LOW:	8,003.93	Total Vol:	531,132			
CALL						TXO	201605	PUT				
Bid	Ask	Last	Change	IttVol	Time	Strike	Bid	Ask	Last	Change	IttVol	Time
163.000	175.000	168.000	34.000	3,250	13:44:56	7900	5.400	5.500	5.000	-11.500	39,899	13:44:59
121.000	130.000	122.000	27.000	5,612	13:44:57	7950	10.500	11.000	11.000	-15.500	38,993	13:44:59
80.000	83.000	82.000	19.000	28,732	13:44:56	8000	19.500	20.000	20.000	-24.000	52,576	13:44:59
47.000	48.000	48.000	11.500	36,679	13:44:56	8050	35.000	35.500	35.500	-32.500	30,582	13:44:59
23.000	24.000	23.000	4.500	51,480	13:44:59	8100	61.000	63.000	61.000	-40.000	21,272	13:44:58
9.800	10.000	10.000	0.600	31,489	13:44:59	8150	98.000	100.000	98.000	-43.000	4,697	13:44:59
3.600	3.700	3.600	-1.400	25,761	13:44:59	8200	140.000	142.000	141.000	-45.000	4,153	13:44:58
0.800	0.900	0.800	-1.400	14,106	13:44:57	8250	183.000	190.000	184.000	-46.000	227	13:43:59

Fig. 1. Option Snapshot Quotes of the Taiwan Futures Exchange at May 16, 2016.

The model has the liquidity constraints, we assume that an investor can buy at least  $U$  and sell at least  $L$  contracts at each strike price  $k_c^i, k_p^i \in K$ :

$$L \leq x_i^c, x_i^p \leq U, L < 0, U > 0, i \in I. \quad (13)$$

The inflection point  $k$  introduced in Eq. (10), the maximal loss  $\mathcal{L}$ , and the expected (forecasting) price  $\hat{S}_T$  Eq. (12) should be specified by an investor. Thus to address research question **Q1** we formulated the integer linear programming of option's portfolio selection problem (7), subject to the portfolio constraints (8)-(13).

We assume that an investor can use the money received from the sale of some contracts to buy other contracts in the portfolio, then the initial cost of portfolio  $C(t, X)$  Eq. (2) can be either a positive number or zero, or even negative number. In the Section 3 we will represent the series of numerical experiments for these three cases. Thus, the system of constraints (8)-(13) can be (optionally) extended with the constraint on the initial cost of portfolio:

$$C(t, X) \geq 0.$$

Another series of numerical experiments will be conducted to define the sensitivity of the solution on the liquidity constraints Eq. (13).

### 3. Data Collection and Processing

To address research question **Q2** we apply the optimization problem (7)–(13) and construct the option's strategy with the certain payoff function on the derivatives of Taiwan Futures Exchange. All options in the Taiwan's market are European-style. The expiration periods of TXO options have spot month, the next two months, and the next two quarterly calendar months<sup>2</sup>. We will be designing option's portfolio from the daily closing ask- and bid-prices for TXO options<sup>3</sup>. We select TXO options; they are liquid assets and comprise above 60% of trading volume of TAIEX. Here we have taken a single date, May 16, 2016, selected at random, to illustrate the portfolio design in practice.

The strikes and the ask- and bid-prices of options are required inputs to the portfolio optimization model. The available TXO prices are denominated in New Taiwan Dollars (NTD). The TAIEX index closed at 8,067.60 on May 16, 2016 (Fig. 1), and the May options contracts expired 9 days later, on May 25, 2016. The option price equals to  $S_0 = 8,067.60$  NTD at May 16, 2016. Next, we will consider two cases for the expected price,  $\hat{S}_T$ . Suppose that the price will significant move up to 1)  $\hat{S}_T = 8,300.00$  NTD, 2)  $\hat{S}_T = 8,400.00$  NTD at May 25, 2016.

The investor then wants to monetize his position,  $C(t, X) = 100$  NTD at the time of purchase,  $t = 0$ , and to limit the maximum loss by  $\mathcal{L} = -100$  NTD, if the price of underlying asset will come out of the certain range from 8,000 to 8,400 NTD, respectively. The margin requirement for short positions and transaction costs are not accounted. To establish the proposed strategy we use 12 strike prices: sequential  $n = 6$  strike prices are corresponding to call

$$K_c = (\mathbf{8050}, 8150, 8250, 8350, 8400, 8500)$$

and sequential  $n = 6$  strike prices are corresponding to put

$$K_p = (7850, 7950, \mathbf{8050}, 8150, 8250, 8350)$$

at the same expiration date May 25, 2016. The central strike of the option is  $K = 8050$ , is marked with bold. The number of combinations of ask- and bid-price for the 12 options equal to  $2^n \times 2^n = 2^6 \times 2^6 = 4096$ , thus the cardinality of set of feasible

<sup>2</sup><http://www.taifex.com.tw/eng/eng4/Calendar.asp>

<sup>3</sup><http://www.taifex.com.tw/eng/eng2/TX0.asp>

**Table 1. Optimal portfolios with the different initial costs,  $C$ , and the expected price  $\hat{S}_T$ , NTD**

Strike Price	$\hat{S}_T = 8400$						$\hat{S}_T = 8300$					
	$C = 100$		$C = -100$		$C = 0$		$C = 100$		$C = -100$		$C = 0$	
	Call	Put	Call	Put	Call	Put	Call	Put	Call	Put	Call	Put
7850		0		0		0		0		0		0
7950		0		0		0		0		0		0
<b>8050</b>	4	-3	7	-6	3	-3	7	-7	3	-2	5	-5
8150	-8	8	-10	9	1	2	-8	9	-2	4	-4	6
8250	10	-5	4	0	-5	6	4	3	0	0	0	4
8350	-8	0	-3	-3	-5	-5	-3	-5	-1	-2	-5	-5
8400	-5		-2		2		-9		-7		-2	
8500	7		4		4		9		7		6	
$\max_x \{F_C(X)\}$	<b>700</b>		400		600		<b>400</b>		250		300	
Total number of contracts	58		48		36		64		28		42	

**Table 2. Optimal portfolios with the different liquidity constraints,  $|L| = U$ , contracts, and the expected price  $\hat{S}_T$ , NTD**

Strike Price	$\hat{S}_T = 8400$						$\hat{S}_T = 8300$					
	$ L  = 10^a)$		$ L  = 50$		$ L  = 100$		$ L  = 10^a)$		$ L  = 50$		$ L  = 100$	
	Call	Put	Call	Put	Call	Put	Call	Put	Call	Put	Call	Put
7850		0		0		0		0		0		0
7950		0		0		0		0		0		0
<b>8050</b>	4	-3	-24	24	-51	51	7	-7	-24	24	-50	50
8150	-8	8	50	-50	100	-100	-8	9	47	-47	100	-100
8250	10	-5	-11	30	-4	49	4	3	-4	22	-12	50
8350	-8	0	-27	-4	-89	0	-3	-5	-21	1	-40	0
8400	-5		-1		21		-9		-15		-35	
8500	7		13		23		9		17		37	
$\max_x \{F_C(X)\}$	700		1800		4400		400		800		1800	
Total number of contracts	58		234		488		64		222		474	

<sup>a)</sup> The column  $|L| = 10$  corresponds to the column  $C = 100$  of Table 1.

portfolios  $|C| = 4096$ . The cardinality of the set  $|K| = |K_p \cup K_c|$  is 8, therefore, we have 7 pairs of sequential strike prices  $[k_q, k_{q+1}]$ , and that these pairs produce the system of 7 inequalities from Eq. (10), and we should add two equalities on the first closed interval and the left-closed interval from Eq. (8). In our example, we assumed that one can buy or sell at least  $L = -10$ ,  $U = 10$  contracts at each strike price. This assumption does not limit our approach because the total volume (*TitVol*, Fig. 1) is bigger for all strike prices. Then we calculate the price for call and put in accordance with the specific strike (Fig. 1). Next, we maximized the objective function  $F_C(X)$ , proposed in Eq. (7) with the system of constraints (8)–(13). The optimal portfolio was obtained in approximately two minutes on a personal computer, using the MATLAB software.

As a result, the optimum solution in the case  $\hat{S}_T = 8400$ ,  $C = 100$  is

$$X = (\underbrace{4, -8, 10, -8, 5, 7}_{\text{call}}, \underbrace{0, 0, -3, 8, -5, 0}_{\text{put}})$$

the first six elements correspond to call options, the second six – to put options, the total number of contracts are 58 out of which 34 are for buying and 24 are for selling, with objective function value equal to 700 NTD (bold in Table 1). The optimum solution in the case  $\hat{S}_T = 8300$ ,  $C = 100$  is

$$X = (\underbrace{7, -8, 4, -3, -9, 9}_{\text{call}}, \underbrace{0, 0, -7, 9, 3, -5}_{\text{put}})$$

– the total number of contracts are 64 out of which 32 are for buying and 32 are for selling, with objective function value equal to 400 NTD (bold in Table 1). From the Table 1 one can see that the maximum values of the payoff function are achieved at  $C = 100$  NTD, and the minimum values at  $C = -100$  NTD.

At the end of the strategy term, May 25, 2016, the price of underlying index increased to  $S_T = 8,396.20$  NTD. The forecast come true, the amount of loss was limited by the maximal loss  $\mathcal{L} = -100$ .

### 3.1. The Sensitivity of the Solution to the Constraints Variation

The initial cost  $C(t, X)$  Eq. (2) can be either a positive number or zero, or even negative number. We calculated the alternative portfolios with the different initial costs  $C(t, X) = \{-100, 0, 100\}$  with fixed liquidity constraints  $|L| = U = 10$  Eq. (13), and then the values of liquidity constraints were varied  $|L| = U = \{10, 50, 100\}$  with the fixed initial cost  $C(t, X) = 100$ , results are represent in Tables 1, 2. Table 2 shows the strong dependence: with the increase in the number of liquidity constraints, i.e.  $|L| = U$ , the maximum value of the payoff function grows too. Fig. 2 show the payoff functions from the proposed option's portfolio, taking into account, the different values of: a), b) the initial cost  $C(t, X)$ , Eq. (2) and c), d) liquidity constraints  $L, U$ , Eq. (13), respectively. Our strike prices are the  $x$ -axis and the payoff functions are the  $y$ -axis.

The expansion of the boundaries on the liquidity constraints  $|L| = U \in \{10, 50, 100\}$  makes the options portfolio more attractive from the point of view of the terminal payoff amount. On the other hand, there is the difficulty of forming a strategy in view of the buy/sell of a sufficiently large number of underlying assets, a transaction which is difficult to implement for a short time.

## 4. Conclusion and Future Research

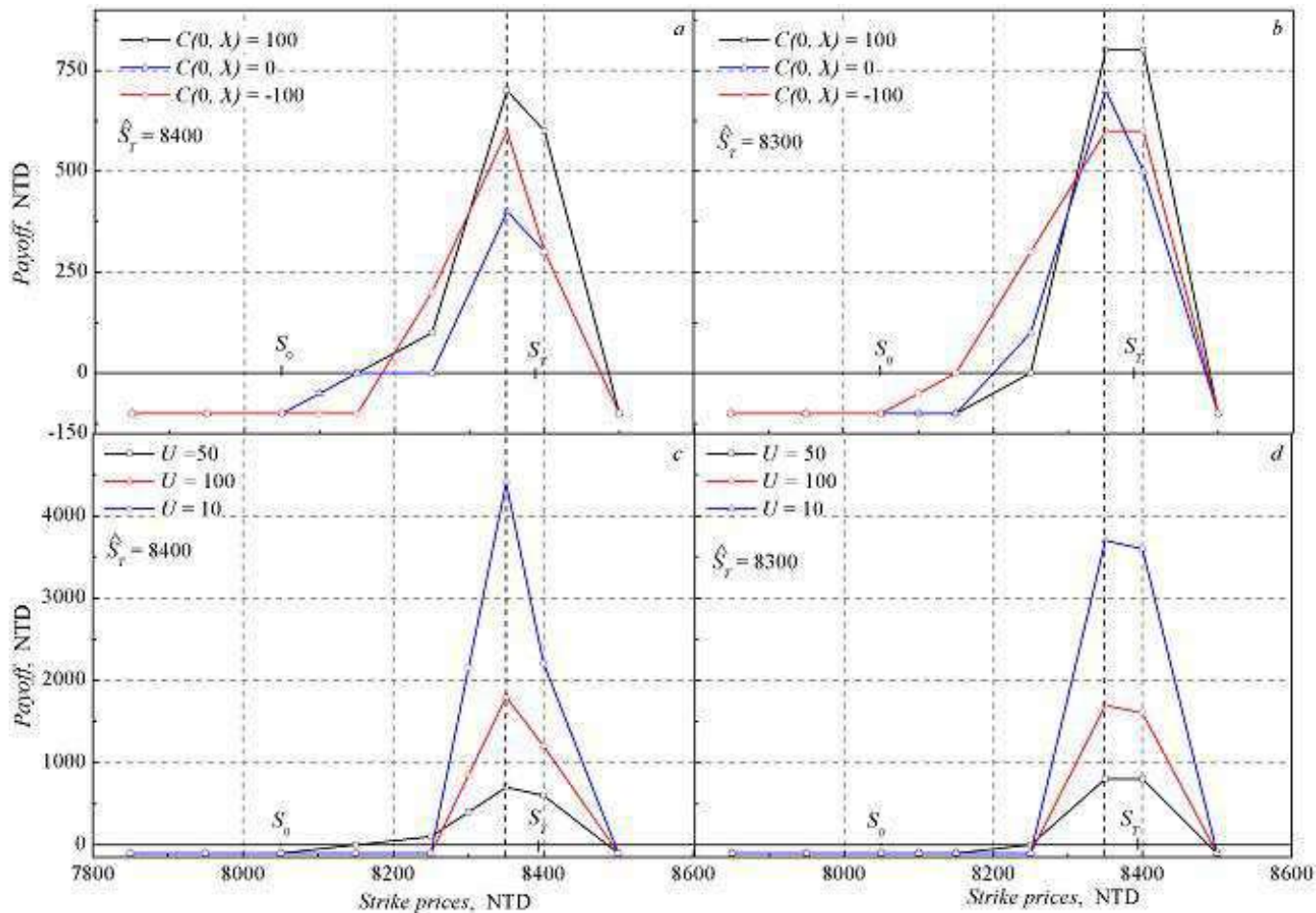
In this study, the description of the method of construction of the option's strategy with the piecewise linear payoff function is carried out. We take into account the next set of assumptions of the proposed option's strategy: strategy should effectively limit the upside earnings and downside risks; strategy should have an initial cost to enter; strategy should have protection on the downside and upside of the underlying asset price.

To address research question **Q1** we have formulated the mathematical model as an integer linear programming problem which includes the system of constraints in the form of equalities and inequalities. The optimum solution was obtained using the MATLAB software.

To address research question **Q2** we have demonstrated the possibility of the proposed model on the European-style TXO options of Taiwan Futures Exchange.

In the study we do not use the historical empirical distribution of returns. Our approach is statical, quite general and has the potential to design option's portfolios on financial markets.

In option's strategies, in addition to the forecast of the price of the underlying asset, various parameters can be taken into account: exercise price, volatility, the time to expiration, the risk-free interest rate, option premium, transaction costs. In this case, even an insignificant change of the parameter's values can lead to a significant change in the number of possible combinations of option's strategies. In our numerical experiments the total number of contracts for different cases varieties from 28 to 64 (Table 1) and from 58 to 488 (Table 2). In papers [14, 17, 18], it is noted that transactional costs in the dynamic management of the portfolio of options are one of the key factors without which it is impossible to talk about the feasibility of using the proposed models. The use of option strategies, including covered options, leads to deformation of the initial distribution of returns – it becomes truncated and asymmetric. The payoff of the option's portfolio is asymmetric and non-linear, therefore, from the point of view of risk management. The use of a portfolio with various options contracts is preferable and effective, but the problem of choosing an optimal portfolio is significantly complicated too.



**Fig. 2. Payoff functions of proposed option's portfolios, underlying  $S_0 = 8,067.60$  NTD. Different initial costs  $C(t, X)$  : 100, 0, -100 NTD: a)  $S_T = 8,400.00$  NTD, b)  $S_T = 8,300.00$  NTD. Different liquidity constraints  $L, U$ , contracts: 10, 50, 100: c)  $S_T = 8,400.00$  NTD, d)  $S_T = 8,300.00$  NTD.**

In this paper, we used the European type options in a series only. Pricing and valuation of the American option, even the single-asset option, is a hard problem in a quantitative finance [16, 19]. One of possible approach in the pricing and considering early exercise of the American option is dynamic programming.

The further research of our study can be continued in the following directions. At first, it is a portfolio optimization under transaction costs (exchange commissions, brokerage fees) and the margin requirement for short positions which are essential in the options market. At second, it is using options with different time to the expiration. At third, it is necessary to extend the system of constraints and add the budget constraint. Such extensions allows us to make the proposed approach more realistic and flexible.

### 5. Acknowledgments

Thanks to the editor and referees for several comments and suggestions that were instrumental in improving the paper. We are grateful to Dr. Sergey V. Kurochkin (National Research University Higher School of Economics, Russia) and Mr. Ashu Prakash (Indian Institute of Technology, Kanpur, India) for valuable comments and suggestions that improved the work and resulted in a better presentation of the material.

### References

- [1] Bartonova M. Hedging of sales by zero-cost collar and its financial impact. *Journal of Competitiveness* 2012; 4: 111–127.
- [2] Kaur A, Rattol AS. Commodity hedging through zero-cost collar and its financial impact. *Journal of Energy and Management* 2016; 1:44–57.
- [3] Hull J. *Fundamentals of Futures and Options Markets*. Financial Times, New Jersey, 2002.
- [4] Ju N, Leland H, Senbet LW. Options, option repricing in managerial compensation: Their effects on corporate investment risk. *Journal of Corporate Finance*, 2013.
- [5] Dash M, Kavitha V, Deepa K, Sindhu S. A study of optimal stock and options strategies. *Social Science Research Network*, 2007.
- [6] Garrett S. *An Introduction to the Mathematics of Finance. A Deterministic Approach*. Elsevier Ltd., 2013.
- [7] Griffin B. Review of collar options for cotton industry. *Chemonics International Inc.*, 2007.
- [8] Ederington L, Chaput J. Option spread and combination trading, 2002. URL: [http://optionoffice.ru/wp-content/uploads/2013/08/Option-spread-combination-trading-\(-\)-Research-paper.pdf](http://optionoffice.ru/wp-content/uploads/2013/08/Option-spread-combination-trading-(-)-Research-paper.pdf).
- [9] Carr P, Madan D. *Towards a theory of volatility trading*. Risk Books, London, 1998; 417–427.
- [10] Topaloglou N, Vladimirov H, Zenios S. Optimizing international portfolios with options and forwards. *Journal of Banking and Finance* 2011; 35: 3188–3201.
- [11] Das S, Statman M. Options and structured products in behavioral portfolios. *Journal of Economic Dynamics and Control* 2013; 37: 137–153.
- [12] Kibzun A, Ignatov A. The two-step problem of investment portfolio selection from two risk assets via the probability criterion. *Automation and Remote Control* 2015; 76: 1201–1220.
- [13] Lin C-C, Liu Y-T, Chen A-P. Hedging an option portfolio with minimum transaction lots: A fuzzy goal programming problem. *Applied Soft Computing* 2016; 47: 295–303.
- [14] Davari-Ardakani H, Aminnayeri M, Seifi A. Multistage portfolio optimization with stocks and options. *International Transactions in Operational Research* 2016; 23: 593–622.

- [15] Moshenets MK, Kritski O. Automatic system of detecting informed trading activities in european-style options. *Journal of Engineering and Applied Sciences* 2016; 11: 5727–5731.
- [16] Hajizadeh E, Mahootchi M. Optimized radial basis function neural network for improving approximate dynamic programming in pricing high dimensional options. *Neural Computing and Applications*, 2016; 1–12.
- [17] Goyal A, Saretto A. Option returns and volatility mispricing. *Social Science Research Network*, 2007.
- [18] Primbs JA. Dynamic hedging of basket options under proportional transaction costs using receding horizoncontrol. *International Journal of Control* 2009; 192: 1841–1855.
- [19] Mitchell D, Goodman J, Muthuraman K. Boundary evolution equations for American options. *Math. Finance* 2014; 24: 505–532.

# Stochastic Non-Markovian Schroedinger equation for a three-level quantum system

V. Semin<sup>1</sup>, A. Pavelev<sup>1</sup>

<sup>1</sup>Samara National Research University, 34, Moskovskoye shosse, 443086, Samara, Russia

---

## Abstract

Non-Markovian dynamics of a three-level system is studied with the help of stochastic Schrödinger equation (SSE). We derive a new form of SSE for a three-level system driven by four independent Ornstein-Uhlenbeck stochastic noises. The main advantage of the suggested SSE is the ensuring of the complete positivity of the reduced density operator. We demonstrate significant influence of the non-Markovian noises on the dynamics of the three-level quantum system.

*Keywords:* stochastic Schrödinger equation; three-level systems; non-Markovian dynamics

---

## 1. Introduction

Open quantum systems are usually described by the reduced density operator, which satisfies a master equation. All the master equations can be strictly divided into two classes: Markovian and non-Markovian [1]. The Markovian ones traditionally represent systems of the first order differential equations with the constant coefficients and they are well studied. In opposite, non-Markovian master equations forsake many open questions and they are intensively studied during the last several years [2]. Today it is clear that the non-Markovian master equations may be of two types either integro-differential or differential with variable coefficients. Both types of the non-Markovian master equations are equivalent and the differential one is used more often. Unfortunately, the general form of the non-Markovian master equation, which ensures the complete positivity of the density operator is still unknown.

Another approach to describe dynamics of open quantum systems is to use Stochastic Schrödinger equation (SSE) for the wave vector driven by the noise [3, 4]. The reduced density operator is recovered by mean of stochastic averaging over many realizations of such vectors. There is an exact correspondence between Markovian master equations and SSEs. It is obvious that the dimension of the wave vector is smaller than the dimension of the reduced density operator. This fact open a new possibilities for investigation of high-dimensional open quantum systems, such as spin chains, photosynthetic reaction centre, etc.

As in the case of non-Markovian master equations, non-Markovian SSEs are also intensively examined. The main attempts of researchers here are focused on the so called unravelling of the non-Markovian master equations [5], i.e. construction of a SSE which reproduces all the results given by the master equation. Unravelling is not always possible especially for integro-differential master equations. On the other hand, one can generalise Markovian SSEs to non-Markovian ones without direct connection with the master equation formalism. Such an approach has many advantages and one of them is ensuring of the complete positivity of the reduced density operator.

In this paper we consider the non-Markovian generalization of the SSE for a three-level quantum system. We present results of the direct simulation of the non-Markovian SSE for this model and discuss the main difference between Markovian and non-Markovian SSEs.

## 2. Model

Three level systems can be of three different types  $\Lambda$ ,  $V$  and cascade. The difference between them is forbidden transitions between the energy levels. Let us consider the three-level system of  $V$ -types for concreteness. The Hamiltonian of the  $V$  system in the photonic thermostat is [6] (we set  $\hbar = 1$ )

$$H = \omega_0 H_1 + \Omega_0 H_2 + \sum_{j=1}^{\infty} \omega_j b_j^\dagger b_j + \sum_{i=1}^{\infty} [(f_i J_+ + g_i L_+) b_i + h.c.], \quad (1)$$

where  $b_j$  and  $b_j^\dagger$  are creation and annihilation operator of the  $j$ th photon in the thermostat with frequency  $\omega_j$ ,  $\omega_0/2$  and  $\Omega_0 + \omega_0/2$  are transition frequency in the  $V$ - system,  $f_i$  and  $g_i$  are the constant of system-photon interaction and the matrices  $H_1, H_2, J_+, L_+$  are define as follow [6]

$$H_1 = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, H_2 = \frac{1}{3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, J_+ = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, K_+ = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (2)$$



### 2.1. Markovian evolution

The Markovian master equation for the reduced density operator can be written as [6]

$$\frac{\partial \rho}{\partial t} = \frac{\gamma_J}{2} [(N_J + 1)(2J_- \rho J_+ - J_+ J_- \rho - \rho J_+ J_-) + N_J(2J_+ \rho J_- - J_- J_+ \rho - \rho J_- J_+)] + (J \leftrightarrow K), \quad (3)$$

where  $\gamma_{J,K}$  are the damping constants,  $N_{J,K}$  are the average numbers of the heat photons on the corresponding transition.

The standard procedure of unravelling allows to write the SSE, corresponding to (3) in the following form

$$d|\psi\rangle = -\frac{\gamma_J}{2} ((N_J + 1)J_+ J_- + N_J J_- J_+) |\psi\rangle dt + i\sqrt{\gamma_J(N_J + 1)} J_- |\psi\rangle dW_J^1 + i\sqrt{\gamma_J N_J} J_+ |\psi\rangle dW_J^2 + (J \leftrightarrow K), \quad (4)$$

where  $W_{J,K}^{1,2}$  are independent standard Wiener processes. It is easy to verify using Ito calculus that  $\rho = \mathbb{E}(|\psi\rangle\langle\psi|)$  satisfies the master equation (3).

### 2.2. Non-Markovian evolution

To describe non-Markovian effects Barchielli in [7] suggested to replace Markovian Wiener processes in (4) by some non-Markovian noises. One of the simplest non-Markovian noises is the Ornstein-Uhlenbeck one, which satisfies the stochastic equation

$$dX = -kXdt + dW, \quad (5)$$

where  $k > 0$  is some constant. By substitution the Ornstein-Uhlenbeck processes (5) instead of the Wiener increments in (4) we derived the following non-Markovian SSE

$$d|\tilde{\psi}\rangle = -\left(\frac{\gamma_J}{2}(N_J + 1)J_+ J_- + \frac{\gamma_J}{2}N_J J_- J_+ + ik_J^1 X_J^1 \sqrt{\gamma_J(N_J + 1)}J_- + ik_J^2 X_J^2 \sqrt{\gamma_J N_J}J_+\right)|\tilde{\psi}\rangle dt + i\sqrt{\gamma_J(N_J + 1)}J_- |\tilde{\psi}\rangle dW_J^1 + i\sqrt{\gamma_J N_J}J_+ |\tilde{\psi}\rangle dW_J^2 + (J \leftrightarrow K). \quad (6)$$

Unfortunately, the above SSE is not a mean-1 martingale and we have to somehow modify the equation to satisfy the martingale property. It is straightforward to check that to be a mean-1 martingale Eq. (7) needs 4 more terms in the drift part, namely

$$d|\tilde{\psi}\rangle = -\left(\frac{\gamma_J}{2}(N_J + 1)J_+ J_- + \frac{\gamma_J}{2}N_J J_- J_+ + ik_J^1 X_J^1 \sqrt{\gamma_J(N_J + 1)}(J_- + J_+) + ik_J^2 X_J^2 \sqrt{\gamma_J N_J}(J_- + J_+)\right)|\tilde{\psi}\rangle dt + i\sqrt{\gamma_J(N_J + 1)}J_- |\tilde{\psi}\rangle dW_J^1 + i\sqrt{\gamma_J N_J}J_+ |\tilde{\psi}\rangle dW_J^2 + (J \leftrightarrow K). \quad (7)$$

Obviously, that Eq. (7) is transformed to Eq. (4) when all  $k_m^l = 0$ . Moreover, it is easy to prove that Eq. (7) is a martingale, i.e.  $\mathbb{E}(\langle\tilde{\psi}(t)|\tilde{\psi}(t)\rangle) = 1$  and  $\tilde{\rho}(t) = \mathbb{E}(|\tilde{\psi}(t)\rangle\langle\tilde{\psi}(t)|)$  is a completely positive operator by construction. Note, that operator  $\tilde{\rho}(t)$  does not satisfy the Markovian master equation (3) and, even more, we cannot construct any closed master equation for this operator due to the presence of the noise terms in the drift part of the equation (7). All the above mentioned facts demonstrate the uniqueness of Eq. (7).

## 3. Results of simulation

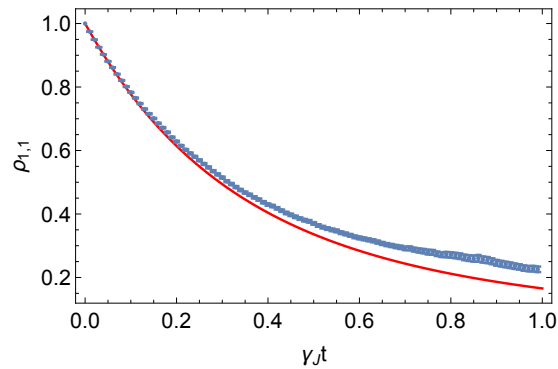
The Non-Markovian SSE (7) can be efficiently simulated. It is possible to do if we add to the three components of the wave vector  $|\tilde{\psi}\rangle$ , four stochastic equations for the Ornstein-Uhlenbeck noises. Resulting seven equations form a closed system and may be numerically solved by any suitable algorithm. The initial values for the Ornstein-Uhlenbeck processes are normally distributed random numbers with zero mean and unit standard deviation.

In this paper we use the Euler stochastic algorithm [8], which for our problem can be written in the following general form

$$\Psi_{n+1} = A\Psi_n \Delta t + \sum_{i=1}^4 B_i \Psi_n \delta W_i, \quad (8)$$

where  $A, B_i$  are constant matrices,  $\Delta t$  is the time step,  $\delta W_i$  is independent normal distributed variables  $N(0, \Delta t)$ . The vector  $\Psi = (\psi_x, \psi_y, \psi_z, X_J^1, X_J^2, X_K^1, X_K^2)^T$  has seven components. Initial conditions is  $\Psi_0 = (1, 0, 0, N(0, 1), N(0, 1), N(0, 1), N(0, 1))^T$ .

The results of simulation are presented in Fig. 1. The results was averaged over  $10^4$  realizations. The error bars are also included in the graphic. In the same pictures we draw the dynamics given by the Markovian master equation (3). One can see that the non-Markovian noise has significant effect on dynamics and cannot be neglected in general.



**Fig.1.** Evolution of the upper state. Red curve is the Markovian dynamics and blue dots are the non-Markovian dynamics. Parameters:  $\gamma_K = 2\gamma_J$ ,  $N_J = 0.2$ ,  $N_K = 0.3$ ,  $k_J^1 = k_K^1 = 0.3$ ,  $k_J^2 = k_K^2 = 0.5$ .

#### 4. Conclusion

In this paper we have derived the non-Markovian SSE for a three-level quantum system driven by the four independent Ornstein-Uhlenbeck processes. The SSE has unique properties which are hard to achieve in other approaches to non-Markovian dynamics. Especially, it is the complete positivity of the density operator  $\tilde{\rho}(t) = \mathbb{E}(|\tilde{\psi}(t)\rangle\langle\tilde{\psi}(t)|)$  for all time. The suggested SSE can be efficiently simulated using any appropriate algorithm and due to stochastic nature of the equation the solution can be easily parallelized to perform calculation on a supercomputer or GPU.

It is shown that quantum dynamics given by the non-Markovian SSE is significantly different from Markovian one and this fact should be taken into account in the explanation of future experiments with quantum ensembles.

The general features of SSE described in this paper for the three-level systems are valid for arbitrary quantum systems. Moreover, the SSE has dimension much smaller than the corresponding master equation. This means that using SSE technique one can describe high dimensional quantum systems. This may be relevant for understanding biological phenomena, such as photosynthesis.

#### References

- [1] Breuer HP, Petruccione F. The theory of open quantum systems. Oxford University Press: Oxford, 2002.
- [2] Breuer HP, Laine EM, Piilo J, Vacchini B. Non-Markovian dynamics in open quantum systems. *Rev. Mod. Phys.* 2015; 88: P. 021002.
- [3] Barchielli A, Gregoratti M. Quantum Trajectories and Measurements in Continuous Time. Springer: Berlin, 2009.
- [4] Semina I, Semin V, Petruccione F, Barchielli A. Stochastic Schrödinger Equations for Markovian and non-Markovian cases. *Open Sys. Inf. Dyn.* 2014. 21: 1440008.
- [5] Barchielli A, Pellegrini C. Jump-diffusion unravelling of a non Markovian generalized Lindblad master equation. *J. Math. Phys.* 2010. 51: 112104.
- [6] Mikhailov VA, Troshkin NV. Relaxation of a three-level atom interacting with a thermostat and an external stochastic field. *Proc. of SPIE* 2015. 9917: 991731.
- [7] Barchielli A, Pellegrini C, Petruccione F. Quantum trajectories: memory and continuous observation. *Phys. Rev. A.* 2012. 86: 063814.
- [8] Platen E, Bruti-Liberati N. Numerical Solution of Stochastic Differential Equations with Jumps in Finance. Springer-Verlag: Berlin, 2010.

# Numerical simulations of the quantum systems dynamics in the path integral approach

A. Biryukov<sup>1</sup>, M. Shleenkov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

We study the dynamics of quantum system interacting with electromagnetic field. We present density matrix and transition probability as a path integrals in energy state space without resonance and rotating wave approximations. By the use of obtained equations we develop an algorithm for numerical simulations of the dynamics of quantum system interacting with electromagnetic field. Using this approach we consider rotational dynamics of nitrogen molecules  $^{14}\text{N}_2$  and  $^{15}\text{N}_2$  which interact with a sequence of ultrashort laser pulses. Our computer simulations indicate the complex dependency of the high rotation states excitation probability upon ultrashort laser pulses sequence periods. We observe pronounced resonances, which correspond to the results of some experiments.

*Keywords:* Path integral formalism; Numerical simulation; Quantum optics; Non-resonance processes

---

## 1. Introduction

The modern development of laser radiation technologies induces theoretical and experimental investigations of the dynamics of quantum objects (such as atoms or molecules) under the action of intense electromagnetic field of different forms.

This dynamics is principally non-linear, because the probability is high of multiphoton processes (absorption and emission more the one photon) and nonresonant processes (electromagnetic field frequency is far from quantum transitions frequency). We note the recent studies of different rare gases multiphoton ionization [1, 2, 3], of multiphoton photoemission of the Au(111) surface state with 800-nm laser pulses [4], of multiphoton transitions in GaSb/GaAs quantum-dot intermediate-band solar cells [5], of three-photon electromagnetically induced absorption in a ladder-type atomic system [6].

There are certain difficulties for theoretical studies of these processes and for simulations of quantum objects dynamics that interact with laser field. Thus, different approximations are used. For example, there are two- or three-level quantum system models [7] and rotating wave approximation [8]. For high-intensity laser field the perturbation theory runs into problems. It is necessary to calculate the large number of terms. High-order perturbation theory for multilevel quantum system dynamics was considered in [9]. For theoretical researches of this processes the numerical solution of time-dependent Schrödinger equation is used [10]. For this reason different schemes of space-time discretization is realized. The discretization parameter should be small enough for simulations of a minute error.

We present original non perturbative method of transition probability calculation in path integral approach [11]. In this paper we present original approach for numerical simulations of the dynamics of a quantum system, interacting with laser radiation by path integration in energy states space.

Recent experimental [12] and theoretical [13, 14] investigations point at possibilities of selective excitations of nitrogen isotopes by a sequence of ultrashort laser pulses (a pulse train). We have developed and are applying numerical algorithm to quantum resonances problem in molecule rotational excitation by ultrashort laser pulses.

## 2. Mathematical model of quantum system interacting with electromagnetic field

We consider interaction of multilevel quantum system (such as an atom or a molecule) with electromagnetic field. The Hamiltonian  $\hat{H}_{full}$  describing our model is given as

$$\hat{H}_{full} = \hat{H}_{syst} + \hat{V}, \quad (1)$$

where  $\hat{H}_{syst}$  is Hamiltonian of the investigated quantum system. We define stationary eigenstates  $|l\rangle$  with energies  $E_l$  having the following properties:

$$\hat{H}_{syst} = \sum_{l=0}^{N-1} E_l |l\rangle\langle l|, \quad (2)$$

$$\sum_{l=0}^{N-1} |l\rangle\langle l| = 1, \quad \langle l'|l\rangle = \delta_{l'l}; \quad (3)$$

$\hat{V}$  – the interaction operator.

Our main goal is to define the probability  $P(l_f, t|l_{in}, 0)$  of investigated quantum system transition from eigenstate  $|l_{in}\rangle$  at the moment  $t = 0$  to the one  $|l_f\rangle$  at the moment  $t > 0$ .

We describe the investigated system by statistical operator  $\hat{\rho}(t)$ . The evolution equation of  $\hat{\rho}(t)$  in Dirac (interaction) picture [15] is as follows:

$$\hat{\rho}(t) = \hat{U}_D(t)\hat{\rho}(0)\hat{U}_D^\dagger(t), \quad (4)$$

where  $\hat{\rho}(0)$  — statistical operator at initial time moment  $t = 0$ ,

$$\hat{U}_D(t) = T \exp\left[-\frac{i}{\hbar} \int_0^t \hat{V}_D(\tau) d\tau\right] \quad (5)$$

– the evolution operator in Dirac picture,

$$\hat{V}_D(\tau) = \exp\left[\frac{i}{\hbar} \hat{H}_{\text{sys}} \tau\right] \hat{V}(\tau) \exp\left[-\frac{i}{\hbar} \hat{H}_{\text{sys}} \tau\right] \quad (6)$$

– the operator of quantum system and electromagnetic field interaction in Dirac picture.

Eq. (4) in energy representation on the base of eigenvectors Eq. (3) is

$$\rho_{l_f m_f}(t) = \sum_{l_{in}, m_{in}} \langle l_f | \hat{U}_D(t) | l_{in} \rangle \rho_{l_{in} m_{in}} \langle m_{in} | \hat{U}_D^\dagger(t) | m_f \rangle, \quad (7)$$

where

$$\rho_{l_f m_f}(t) = \langle l_f | \hat{\rho}(t) | m_f \rangle, \quad \rho_{l_{in} m_{in}} = \langle l_{in} | \hat{\rho}(0) | m_{in}' \rangle \quad (8)$$

are density matrix elements in energy representation at time moment  $t = 0$ .

The probability of a quantum state observation is to define as diagonal matrix element. At initial time moment  $t = 0$  it is equal to  $\rho_{l_{in} l_{in}}(t = 0) = \rho_{l_{in}}$ . At final time moment  $t$  it is equal to  $\rho_{l_f l_f}(t) = \rho_{l_f}(t)$ .

Eq. (7) describes evolution of the probability of a quantum state observation:

$$\rho_{l_f}(t) = \sum_{l_{in}} \langle l_f | \hat{U}(t) | l_{in} \rangle \langle l_{in} | \hat{U}^\dagger | l_f \rangle \rho_{l_{in}}, \quad (9)$$

where  $l_{in}, l_f = 1, 2, \dots$

The quantum transition probability from state  $|l_{in}\rangle$  or  $\rho_{l_{in}}(0) = \delta_{l_{in} n_{in}}$  at time moment  $t = 0$  to state  $|l_f\rangle$  or  $\rho_{l_f}(t) = \rho_{l_f m_f}(t) \delta_{l_f m_f}$  at time moment  $t > 0$  is to describe as follows

$$P(l_f, t|l_{in}, t) = \langle l_f | \hat{U}_D(t) | l_{in} \rangle \langle l_f | \hat{U}_D(t) | l_{in} \rangle \quad (10)$$

By the use of eq. (10) we present eq. (9) in the following

$$\rho_{l_f}(t) = \sum_{l_{in}} P(l_f, t|l_{in}, 0) \rho_{l_{in}}(0) \quad (11)$$

For numerical calculation  $\rho_{l_f m_f}(t)$ ,  $\rho_{l_f}(t)$ ,  $P(l_f, t|l_{in}, 0)$  by the use of eq. (7), eq. (9) and eq. (10) we need to know matrix elements  $\langle l_f | \hat{U}(t) | l_{in} \rangle$  of evolution operator  $\hat{U}(t)$ .

For that reason we use evolution operator  $\hat{U}$  group properties and express the evolution operator  $\langle l_f | \hat{U}_D(t) | l_{in} \rangle$  as

$$\hat{U}_D(t) = \prod_{k=1}^{K+1} \hat{U}_D(t_k, t_{k-1}), \quad (12)$$

as long as  $t_k > t_{k-1}$  and where

$$\hat{U}_D(t_k, t_{k-1}) = \exp\left[-\frac{i}{\hbar} \int_{t_{k-1}}^{t_k} \hat{V}_D(\tau) d\tau\right], \quad (13)$$

here we introduce the notations  $t_{K+1} = t$ ,  $l_{K+1} = l_f$ ,  $t_0 = 0$ ,  $l_0 = l_{in}$ ,  $\sum_{k=1}^{K+1} (t_k - t_{k-1}) = t$ .

By the use of eq. (12) and completeness condition Eq. (3) of eigenvectors  $|l_k\rangle$  basis the kernel  $\langle l_f | \hat{U}_D(t) | l_{in} \rangle$  can be expressed as

$$\langle l_f | \hat{U}_D(t) | l_{in} \rangle = \sum_{l_1, \dots, l_K=0}^{N-1} \prod_{k=1}^{K+1} \langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle, \quad (14)$$

where

$$\langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \langle l_k | \exp \left[ -\frac{i}{\hbar} \int_{t_{k-1}}^{t_k} \hat{V}_D(\tau) d\tau \right] | l_{k-1} \rangle \quad (15)$$

We consider the evolution operator kernel Eq. (15) as a series and for the time interval  $(t_k - t_{k-1}) \rightarrow 0$  it is

$$\langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \langle l_k | l_{k-1} \rangle - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} \langle l_k | \hat{V}_D(\tau) | l_{k-1} \rangle d\tau. \quad (16)$$

Using eq. (6), interaction operator matrix element in Dirac picture  $\langle l_k | \hat{V}_D(\tau) | l_{k-1} \rangle$  is expressed

$$\langle l_k | \hat{V}_D(\tau) | l_{k-1} \rangle = V_{l_k l_{k-1}}(\tau) \exp[i\omega_{l_k l_{k-1}} \tau]. \quad (17)$$

where  $V_{l_k l_{k-1}}(\tau) = \langle l_k | \hat{V}(\tau) | l_{k-1} \rangle$  – interaction operator matrix element,  $\omega_{l_k l_{k-1}} = (E_{l_k} - E_{l_{k-1}})/\hbar$  – frequency of quantum transition between eigenstates with eigenvalues (energies)  $E_{l_k}$  and  $E_{l_{k-1}}$ .

It is possible to prove that for small time interval  $(t_k - t_{k-1}) \rightarrow 0$  the evolution operator kernel  $\langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle$  eq. (16) can be expressed as

$$\langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \int_0^1 \exp[iS[l_k, l_{k-1}; \xi_{k-1}]] d\xi_{k-1}, \quad (18)$$

where  $S[l_k, l_{k-1}; \xi_{k-1}]$  – dimensionless (in  $\hbar$  units) action in energy representation during time interval  $(t_k - t_{k-1})$

$$S[l_k, l_{k-1}; \xi_{k-1}] = 2\pi(l_k - l_{k-1})\xi_{k-1} - \int_{t_{k-1}}^{t_k} \frac{V_{l_k l_{k-1}}(\tau)}{\hbar} 2 \cos[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau] d\tau, \quad (19)$$

where  $V_{l_k l_{k-1}}(\tau) = \langle l_k | \hat{V}(\tau) | l_{k-1} \rangle$  – interaction operator matrix element.

For this proof, by using eq. (15) we transform eq. (18) into eq. (16).

By the use of eq. (15) we present eq. (14) in the following

$$\begin{aligned} & \langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \\ & = \int_0^1 \exp[2\pi i(l_k - l_{k-1})\xi_{k-1}] \exp\left[-\frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) 2 \cos[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau] d\tau\right] d\xi_{k-1} \end{aligned} \quad (20)$$

If  $(t_k - t_{k-1}) \rightarrow 0$  then we write

$$\begin{aligned} & \exp\left[-\frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) 2 \cos[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau] d\tau\right] \simeq \\ & \simeq 1 - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) 2 \cos[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau] d\tau \end{aligned} \quad (21)$$

where

$$2 \cos[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau] = e^{-i[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau]} + e^{+i[2\pi(l_k - l_{k-1})\xi_{k-1} - \omega_{l_k l_{k-1}} \tau]}. \quad (22)$$

Using eq. (21) and eq. (22) we present eq. (20) in the following

$$\begin{aligned} & \langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \\ & = \int_0^1 \exp[2\pi i(l_k - l_{k-1})\xi_{k-1}] d\xi_{k-1} - \\ & - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) \int_0^1 (\exp[4\pi i(l_k - l_{k-1})\xi_{k-1} - i\omega_{l_k l_{k-1}} \tau] + \exp[i\omega_{l_k l_{k-1}} \tau]) d\xi_{k-1} d\tau = \\ & = \int_0^1 \exp[2\pi i(l_k - l_{k-1})\xi_{k-1}] d\xi_{k-1} - \\ & - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) \int_0^1 \{\exp[i\omega_{l_k l_{k-1}} \tau] + \exp[4\pi i(l_k - l_{k-1})\xi_{k-1}] \cdot \exp[-i\omega_{l_k l_{k-1}} \tau]\} d\tau d\xi_{k-1} \end{aligned} \quad (23)$$

We note that

$$\int_0^1 \exp[4\pi i(l_k - l_{k-1})\xi_{k-1}] d\xi_{k-1} = \delta_{l_k l_{k-1}}, \quad (24)$$

if  $n = 1, 2, \dots$  is integer.

Using eq. (24) we transform eq. (23) to the following

$$\langle l_k | \hat{U}_D(t_k, t_{k-1}) | l_{k-1} \rangle = \delta_{l_k l_{k-1}} - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) \exp[i\omega_{l_k l_{k-1}} \tau] d\tau - \frac{i}{\hbar} \int_{t_{k-1}}^{t_k} V_{l_k l_{k-1}}(\tau) \exp[-i\omega_{l_k l_{k-1}} \tau] d\tau \quad (25)$$

By the use of eq. (3) and  $V_{l_k l_{k-1}} = 0$  for  $l_k = l_{k-1}$  we prove that eq. (25) is the same as eq. (16)

We note that using Eq. (14), Eq. (18), Eq. (12) quantum transition amplitude  $U_D(l_f, t | l_{in}, 0)$  for any  $t$  can be expressed as path integral in energy eigenstates space

$$\langle l_f | \hat{U}_D(t) | l_{in} \rangle = U_D(l_f, t | l_{in}, 0) = \lim_{K \rightarrow \infty} \sum_{l_1, \dots, l_K=0}^{N-1} \int_0^1 \dots \int_0^1 \exp[iS[l_f, l_K, \xi_K; \dots; l_k, l_{k-1}, \xi_{k-1}; \dots; l_1, l_{in}, \xi_0]] d\xi_0 \dots d\xi_K, \quad (26)$$

where

$$S[l_f, l_K, \xi_K; \dots; l_k, l_{k-1}, \xi_{k-1}; \dots; l_1, l_{in}, \xi_0] = \sum_{k=1}^{K+1} S[l_k, l_{k-1}, \xi_{k-1}] \quad (27)$$

– dimensionless action. It is a functional, which is defined on a path set in discrete variables  $l_k$  space of size  $N$  (quantum system levels number) and continuous c-number variables  $\xi_k$  space  $[0, 1]$ .

The quantum transition amplitude eq. (26) with eq. (27) and eq. (19) describes transition of quantum system under electromagnetic field influence. It is possible to use for high-intensity and an arbitrary structure of field in space and time. Parameters  $\omega_{l_k l_{k-1}}$  and  $V_{l_k l_{k-1}}$  must be defined for investigated model.

However analytical calculation eq. (26) can not be realized on practice. Then we develop numerical approach to amplitude eq. (26) calculation as well as for eq. (11), (10), (7).

### 3. Algorithm of numerical simulation of quantum system dynamics

We consider algorithm for numerical calculation of quantum transition amplitude  $U(l_f, t | l_{in}, 0)$  and probability  $P(l_f, t | l_{in}, 0)$ . Using eq. (14) the quantum transition amplitude calculation was made by recurrence relation

$$U(l_K, t_K | l_{in}, 0) = \sum_{l_{k-1}} U(l_K, t_K | l_{k-1}, t_{k-1}) U(l_{k-1}, t_{k-1} | l_{in}, 0), \quad (28)$$

where we introduce

$$\begin{aligned} U(l_K, t_K | l_{in}, 0) &= \langle l_K | \hat{U}_D(t_K) | l_{in} \rangle, \\ U(l_K, t_K | l_{k-1}, t_{k-1}) &= \langle l_K | \hat{U}_D(t_K, t_{k-1}) | l_{k-1} \rangle, \\ U(l_{k-1}, t_{k-1} | l_{in}, 0) &= \langle l_{k-1} | \hat{U}_D(t_{k-1}) | l_{in} \rangle. \end{aligned}$$

We define transition amplitude  $U(l_0, t_0 | l_{in}, 0)$  for first iteration with  $t_0 = \Delta\tau_0$ . By the use of eq. (28) and eq. (18) we obtain transition amplitude for an arbitrary time moment  $t = \sum_{k=0}^K \Delta\tau_k$ .

For any  $t_k, t_{k-1}$  the transition amplitude  $U(l_k, t_k | l_{k-1}, t_{k-1})$  is a complex number. Then for numerical calculation we need to express real and imaginary part of amplitude:

$$U(l_k, t_k | l_{k-1}, t_{k-1}) = \Re[U(l_k, t_k | l_{k-1}, t_{k-1})] + i \Im[U(l_k, t_k | l_{k-1}, t_{k-1})] \quad (29)$$

For these parts eq. (28) transform into two equations:

$$\Re[U(l_k, t_k | l_{in}, 0)] = \sum_{l_{k-1}=0}^{N-1} (\Re[U(l_k, t_k | l_{k-1}, t_{k-1})] \Re[U(l_{k-1}, t_{k-1} | l_{in}, 0)] - \Im[U(l_k, t_k | l_{k-1}, t_{k-1})] \Im[U(l_{k-1}, t_{k-1} | l_{in}, 0)]) \quad (30)$$

$$\Im[U(l_k, t_k | l_{in}, 0)] = \sum_{l_{k-1}=0}^{N-1} (\Im[U(l_k, t_k | l_{k-1}, t_{k-1})] \Re[U(l_{k-1}, t_{k-1} | l_{in}, 0)] + \Re[U(l_k, t_k | l_{k-1}, t_{k-1})] \Im[U(l_{k-1}, t_{k-1} | l_{in}, 0)]) \quad (31)$$

We present eq. (30) and eq. (31) in matrix form

$$= \sum_{l_{k-1}=0}^{N-1} \int_0^1 \begin{pmatrix} \Re[U(l_k, t_k | l_{k-1}, t_{k-1})] & -\Im[U(l_k, t_k | l_{k-1}, t_{k-1})] \\ \Im[U(l_k, t_k | l_{k-1}, t_{k-1})] & \Re[U(l_k, t_k | l_{k-1}, t_{k-1})] \end{pmatrix} \begin{pmatrix} \Re[\tilde{U}(l_k, t_k | l_{in}, 0)] \\ \Im[\tilde{U}(l_k, t_k | l_{in}, 0)] \end{pmatrix} = \begin{pmatrix} \Re[U(l_{k-1}, t_{k-1} | l_{in}, 0)] \\ \Im[U(l_{k-1}, t_{k-1} | l_{in}, 0)] \end{pmatrix}, \quad (32)$$

The initial condition for pure quantum state  $|l_{in}\rangle$  is as follows

$$\begin{pmatrix} \Re[U(l_0, 0 | l_{in}, 0)] \\ \Im[U(l_0, 0 | l_{in}, 0)] \end{pmatrix} = \begin{pmatrix} \delta_{l_0 l_{in}} \\ 0 \end{pmatrix}. \quad (33)$$

Quantum transition probability  $P(l_k, t_k | l_{in}, 0)$  of investigated system from the state  $|l_{in}\rangle$  at moment  $t = 0$  to the state  $|l_k\rangle$  at moment  $t_k$  can be expressed as

$$P(l_k, t_k | l_{in}, 0) = (\Re[U(l_k, t_k | l_{in}, 0)])^2 + (\Im[U(l_k, t_k | l_{in}, 0)])^2, \quad (34)$$

The transition probability  $P(l_k, t_k | l_{in}, 0)$  must be normalized for each time moments  $t_k$

$$\sum_{l_k=0}^{N-1} P(l_k, t_k | l_{in}, 0) = 1. \quad (35)$$

For this we calculate  $\Re[U(l_k, t_k | l_{in}, 0)]$  and  $\Im[U(l_k, t_k | l_{in}, 0)]$  using eq. (34) and product them on normalizing factor  $A$ :

$$\begin{pmatrix} \Re[U(l_k, t_k | l_{in}, 0)] \\ \Im[U(l_k, t_k | l_{in}, 0)] \end{pmatrix} = A \begin{pmatrix} \Re[U(l_k, t_k | l_{in}, 0)] \\ \Im[U(l_k, t_k | l_{in}, 0)] \end{pmatrix}. \quad (36)$$

The normalizing factor  $A$  is calculated by the following formula:

$$A = \left( \sum_{l_k=0}^{N-1} (\Re[\tilde{U}(l_k, t_k | l_{in}, 0)]^2 + \Im[\tilde{U}(l_k, t_k | l_{in}, 0)]^2) \right)^{-1/2}. \quad (37)$$

Using Eq. (30)–(37) we calculate the amplitude  $U(l_f, t | l_{in}, 0)$ , the transition probability  $P(l_f, t | l_{in}, 0)$  and the probability of quantum state observation for any  $t$ .

#### 4. Rotational dynamics of $^{14}\text{N}_2$ and $^{15}\text{N}_2$ interacting with laser pulses sequences

Recent results of experimental observation of  $^{14}\text{N}_2$  and  $^{15}\text{N}_2$  high rotational states excitation were published in [12]. Detailed discussions of the results were in [14, 13].

In the experiments the groups of  $^{14}\text{N}_2$  and  $^{15}\text{N}_2$  molecules were investigated. At the initial moment the distribution of rotational population is thermal and corresponds to  $T = 6.3$  K. Molecules interact with a sequence of ultrashort laser pulses with period from 6.5 ps to 9.5 ps. Each laser pulse has duration equal 500 fs. Laser radiation intensity reaches the value  $I = 5 * 10^{12}$  W/cm<sup>2</sup>. The relative populations were measured of the rotational levels of  $^{14}\text{N}_2$  and  $^{15}\text{N}_2$  and the functional dependence of the populations on the pulse train period was obtained.

The results of these experiments show that there are quantum nonlinear resonances i.e. the nonlinear increase of rotational excitation efficiency under specific values of the pulse train period. The most efficient population transfer up the rotational ladder occurs around 8.4 ps for  $^{14}\text{N}_2$  and 9 ps for  $^{15}\text{N}_2$ .

We analyse these experiments using the method developed by us which is based on path integral formulation in energy states space.

We calculate the energy  $E_l$  of investigated molecules rotational levels for quantum rigid rotor model [16]

$$-\frac{\hbar^2}{2I} \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta \frac{\partial}{\partial \theta}) Y_l(\theta) = E_l Y_l(\theta), \quad (38)$$

where  $I = \mu R^2$  – moment of inertia,

$\mu$  – molecule reduced mass,

$R$  – atom distances,

$Y_l(\theta) = Y_l^0(\theta, \phi)$ , where  $Y_l^m(\theta, \phi)$  – spherical harmonics.

Eq. (38) defines the rotational energy spectrum of a diatomic molecule

$$E_l = \frac{\hbar^2}{2I} l(l+1), \quad (39)$$

where  $l$  – azimuthal quantum number.

It is known, that nonpolar molecule dipole moment is equal to zero. However, strong laser fields induce the molecular dipole by exerting an angle-dependent torque.

The interaction is described by the potential [17, 18]

$$V(\tau) = -\frac{1}{4}\Delta\alpha E^2(\tau) \cos^2 \theta, \quad (40)$$

where  $\Delta\alpha$  describes the molecular polarizability,  $\theta$  is the angle between the molecular axis and the field polarization.

Matrix elements of interaction operator are

$$V_{l'l}(\tau) = -\frac{1}{4}\Delta\alpha E^2(\tau)\langle l' | \cos^2 \theta | l \rangle, \quad (41)$$

where

$$\langle l' | \cos^2 \theta | l \rangle = 2\pi \int_0^\pi Y_{l'}^*(\theta) \cos^2 \theta Y_l(\theta) \sin \theta d\theta. \quad (42)$$

Matrix elements  $\langle l' | \cos^2 \theta | l \rangle$  were numerically calculated by Eq. (41) and Eq. (42).

The investigated molecules parameters are [19]:  $\Delta\alpha = 1.97 * 10^{-40}$  C\*m<sup>2</sup>/V,  $I = 1.4 * 10^{-46}$  kg\*m<sup>2</sup> for <sup>14</sup>N<sub>2</sub>,  $I = 1.5 * 10^{-46}$  kg\*m<sup>2</sup> for <sup>15</sup>N<sub>2</sub>.

We consider a sequence of ultrashort laser pulses which was used in [12]. The electric field value is as follows

$$E(\tau) = \sum_{n=-3}^3 J_n(A) E_0 \exp\left[-\frac{(\tau - n\tau_{per})^2}{\tau_{pul}^2}\right], \quad (43)$$

where  $J_n(A)$  is Bessel function of the first kind,

$A = 2.5$  is the spectral phase modulation amplitude,

$E_0 \approx 6 * 10^9$  V/m is electric field value,

$\tau_{pul} \approx 500$  fs is each laser pulse duration,

$7.98$  ps  $\leq \tau_{per} \leq 9.38$  ps is pulse train period.

We are considering the model of N<sub>2</sub> with  $N = 8$  rotational levels ( $l = 0, 1, \dots, 7$ ). This model is a good approximation, because in experiments [12] higher rotational states are practically not excited.

The initial distribution of rotational population is thermal and corresponds to  $T = 6.3$  K:

$$P_{l_m} = \frac{1}{Z} \exp\left[-\frac{E_{l_m}}{k_B T}\right], \quad (44)$$

where

$$Z = \sum_{l_m=0}^{N-1} \exp\left[-\frac{E_{l_m}}{k_B T}\right] \quad (45)$$

— particle function,

$k$  — Boltzmann factor,

$T$  — absolute temperature,

$N$  — rotational states number in the theoretical model.

By the use of Eq. (44)-(45) and numerical simulation algorithm we calculate the probability of excitation from the initial state (Boltzmann distribution) to different rotational states having interacted with a sequence of ultrashort laser pulses as a function of pulse train period. The absolute error of our probability calculation was not more than  $10^{-3}$ . The results of our numerical simulations are given in fig. 1, fig. 2 and agree well with experimental data as for <sup>14</sup>N<sub>2</sub> both for <sup>15</sup>N<sub>2</sub> molecules.

In fig. 1 we present the population of <sup>14</sup>N<sub>2</sub> molecules on different rotational quantum level  $l$  after interaction with pulse train. For the pulse train period equal to 2.79 ps, 5.58 ps and 8.38 ps for <sup>14</sup>N<sub>2</sub> the population is efficiently transferred from the initial (thermal distribution) states  $l = 0, 1, 2$  to the higher states  $l = 3, 4, 5, 6, 7$ . The resonance train period value  $\tau = 8.38$  ps was observed in experiment [12].

In fig. 2 we present normalized probability of <sup>14</sup>N<sub>2</sub> molecules rotational state observation after they have interacted with 7 laser pulses with period  $\tau = 8.38$  ps and different values of laser pulses maximum intensity  $0.5I_0$ ,  $I_0$  and  $2I_0$ , where  $I_0 = 5 * 10^{12}$  W/cm<sup>2</sup>. We note the population of high rotational state is depend on intensity of laser pulses non-linearly.

## 5. Conclusion

In this paper we present new method of calculating the transition probability of a quantum system interacting with electromagnetic field by the path integral formalism. We construct the amplitude and probability of quantum transition as path integrals



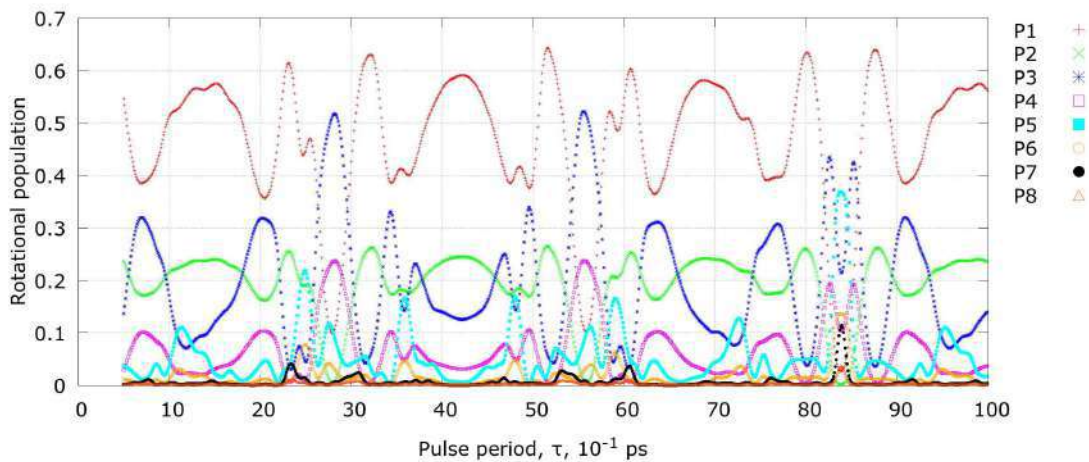


Figure 1: Rotational population of  $^{14}\text{N}_2$  molecules on different rotational quantum level  $l$  after interaction with pulse train with period  $\tau$

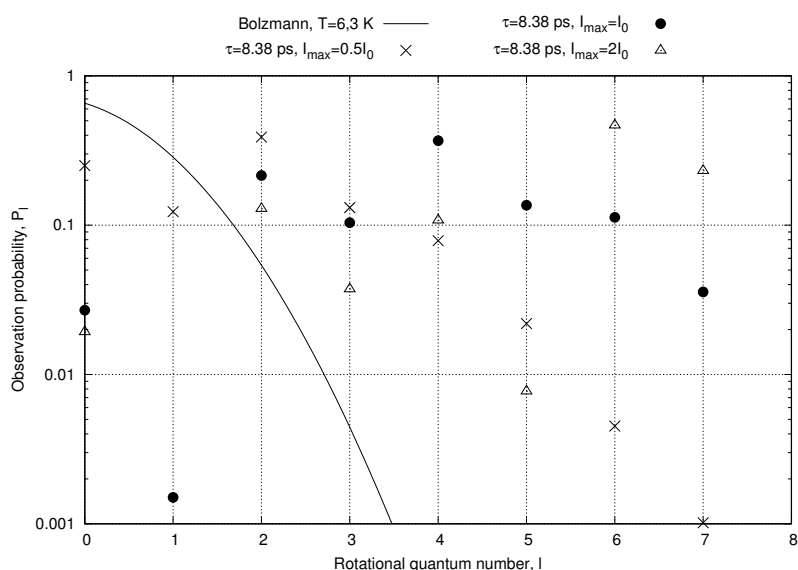


Figure 2: The distribution of observation probabilities of  $^{14}\text{N}_2$  molecules on different rotational quantum level  $l$  after interaction with pulse train ( $\tau = 8.38$  ps)

in energy states space. The algorithm of path integral calculation was developed. This approach enables us to perform computer simulations of molecule dynamics induced by a laser field.

By the deduced formulas we describe quantum resonances in dynamics of nitrogen molecules, that interact with a sequence of ultrashort laser pulses. The obtained results are in good agreement with the experimental data [12] and the theoretical investigations [14, 13] by Schrödinger equation numerical solution.

The approach developed is applicable to nonperturbative studies of different multiphoton and nonresonant processes.

## References

- [1] Gerken N., Klumpp S., Sorokin A. A., Tiedtke K., Richter M., Burk V., Mertens K., Juranic P., Martins M., Time-dependent multiphoton ionization of xenon in the soft-x-ray regime, *Phys. Rev. Lett.* 112 (2014) 213002.
- [2] Guichard R., Richter M., Rost J.-M., Saalman U., Sorokin A. A., Tiedtke K., Multiple ionization of neon by soft x-rays at ultrahigh intensity, *J. Phys. B: At. Mol. Opt. Phys.* 46 (2013) 164025.
- [3] Richter M., Amusia M. Y., Bobashev S. V., Feigl T., Juranic P. N., Martins M., Sorokin A. A., Tiedtke K., Extreme ultraviolet laser excites atomic giant resonance, *Phys. Rev. Lett.* 102 (2014) 163002.
- [4] Sirotti F., Beaulieu N., Bendounan A., Silly M. G., Chauvet C., Malinowski G., Fratesi G., Vniard V., Onida G., Multiphoton k-resolved photoemission from gold surface states with 800-nm femtosecond laser pulses, *Phys. Rev. B* 90 (2014) 035401.
- [5] Hwang J., Lee K., Teran A., Forrest S., Phillips J. D., Multiphoton sub-band-gap photoconductivity and critical transition temperature in type-II gasb quantum-dot intermediate-band solar cells, *Phys. Rev. App.* 1 (2014) 051003.
- [6] Moon H. S., Jeong T., Three-photon electromagnetically induced absorption in a ladder-type atomic system, *Phys. Rev. A* 89 (2014) 033822.
- [7] Cho S., Moon H., Chough Y., Bae M., Kim N., Quantum coherence and population transfer in a driven cascade three-level artificial atom, *Phys. Rev. A* 89 (2014) 053814.

- [8] Spiegelberg J., Sjöqvist E., Validity of the rotating-wave approximation in nonadiabatic holonomic quantum computation, *Phys. Rev. A* 88 (2013) 054301.
- [9] Biryukov A. A., Danilyuk B. V., Rabi oscillations in many-level quantum system, *Proc. SPIE* 7024 (2008) 702405.
- [10] Fleischer S., Khodorkovsky Y., Prior Y., Averbukh I. S., Controlling the sense of molecular rotation, *New J. Phys.* 11 (2009) 105039.
- [11] Biryukov A., Shleenkov M., The influence functional approach to the quantum systems dynamics, *PoS(QFTHEP 2013)* 076 (2013) 1.
- [12] Zhdanovich S., Bloomquist C., Floss J., Averbukh I. S., Hepburn J. W., Milner V., Quantum resonances in selective rotational excitation of molecules with a sequence of ultrashort laser pulses, *Phys. Rev. Lett.* 109 (2012) 043003.
- [13] Floss J., Averbukh I. S., Quantum resonance, anderson localization, and selective manipulations in molecular mixtures by ultrashort laser pulses, *Phys. Rev. A* 86 (2012) 021401.
- [14] Floss J., Fishman S., Averbukh I. S., Anderson localization in laser-kicked molecules, *Phys. Rev. A* 88 (2013) 023426.
- [15] Dirac P. A. M., *Principles of Quantum Mechanics*, Oxford University Press, 1982 (fourth edition).
- [16] Landau L. D., Lifshitz L. M., *Quantum Mechanics. Non-Relativistic Theory*, Butterworth-Heinemann, 1976.
- [17] Zon B. A., Katsnelson B. G., Nonresonant scattering of intense light by a molecule, *JETP* 69 (1975) 1166–1178.
- [18] Underwood J. G., Sussman B. J., Stolow A., Field-free three dimensional molecular axis alignment, *Phys. Rev. Lett.* 94 (2005) 143002.
- [19] Irikura K., Experimental vibrational zero-point energies: Diatomic molecules, *J. Phys. Chem. Ref. Data* 36 (2007) 389.
- [20] Feynman R. P., Space-time approach to non-relativistic quantum mechanics, *Rev. Mod. Phys.* 20 (1948) 367.
- [21] Feynman R. P., Hibbs A. R., *Quantum Mechanics and Path Integrals*, McGraw-Hill Companies, 1965.
- [22] Dirac P. A. M., The lagrangian in quantum mechanics, *Physikalische Zeitschrift der Sowjetunion* 3 (1933) 64–72.
- [23] Bornyakov V., Ilgenfritz E.-M., Martemyanov B., Mitryushkin V., Muller-Preussker M., Topology across the finite temperature transition studied by overimproved cooling in gluodynamics and qcd, *Phys. Rev. D* 87 (2013) 114508.
- [24] Valgushev S. N., Lushevskaya E. V., Pavlovsky O. V., Polikarpov M. I., Ulybyshev M. V., The influence of defects on the conductivity of graphene within the effective theory approach, *JETP Lett.* 98 (2013) 445.
- [25] Bichkov A. B., Mityureva A. A., Smirnov V. V., Short-pulse photoexcitation process in the hydrogen atom, *Phys. Rev. A* 79 (2009) 013402.
- [26] Bichkov A. B., Mityureva A. A., Smirnov V. V., Path-integral-based evaluation of the probability of hydrogen atom ionization by short photo-pulse, *J. Phys. B: At. Mol. Opt. Phys.* 44 (2011) 135601.
- [27] Kleinert H., Zatloukal V., Green function of the double-fractional fokker-planck equation: Path integral and stochastic differential equations, *Phys. Rev. E* 88 (2013) 052106.
- [28] Feynman R. P., Vernon Jr. F. L., The theory of a general quantum system interacting with a linear dissipative system, *Annals of Physics* 24 (1963) 118.

# Mathematical modeling of incentive mechanisms in projects for the development of new production

O.V. Pavlov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The incentive problem of executors of the new products development project at the industrial enterprise is considered in this article. Mastering of a new product leads to the learning effect, which implies reduction of time spent on performing repetitive tasks by workers, resulting in a dynamic change in the economic performance of production. The project of the new products development is considered as a managed hierarchical dynamic system, consisting of a project management board (principal) and teams of agents. The interaction of project participants is formalized as a hierarchical dynamic game. To solve the problem, a numerical algorithm is developed based on a sequential solution of two optimal control problems that are solved using the Bellman dynamic programming method.

*Keywords:* new products development project; learning effect; hierarchical dynamic game

---

## 1. Introduction

In the new products development project at industrial enterprises, employees have to master new types of work and equipment, which is associated with the acquisition of new professional skills. In the process of mastering, the learning effect manifests itself, which is that the time spent by workers on performing repetitive tasks is reduced [1, 2]. The learning effect leads to a dynamic change in the economic performance of production: the volume of output per unit of time, labor intensity and production costs.

Every time the cumulative volume of production doubles, the productivity of workers increases by 10-15 percent [1]. The cumulative volume of production means the number of products manufactured from the beginning of production as a cumulative result.

The new production development project at an industrial enterprise is regarded as a managed hierarchical dynamic system consisting of the manager of the project (principal) and executors (agents). The state of the hierarchical dynamic system in each period of time depends on its state and the actions of the participants in the previous period.

Production activity in the project of the new production development is characterized by the mismatching interests of the principal and agents, which leads to a decrease in economic efficiency. The solution of these contradictions is possible with coordinated management mechanisms that encourage agents to choose actions that are beneficial to the principal.

Dynamic models of interaction of unequal players are considered in the theory of Stackelberg's dynamic games [3] and the theory of dynamic games of Germeyer [4]. Applied models of the dynamic games theory in the field of economics and management are given in [5, 6].

In this dynamic game model there are dynamics of decision making and dynamics of the managed system. The inequality of participants is fixed by the moves order. The first move is made by the principal, who chooses his own strategy – the unit payment rate, - and reports it to the agents. The principal, knowing the agent's goal functions, maximizes his goal function, taking into account the optimal responses of agents. It is assumed that agents are not linked to each other and perform independently. Dynamic game model is considered in a discrete form, which reflects the nature of production activity.

## 2. Statement of the dynamic problem of proportional incentive

### 2.1. The decision-making model of the principal

A two-level dynamic production system consisting of the principal and  $n$  independent agents is considered. Agents produce parts, from which the final product is assembled. Labor costs and financial incentives for agents depend only on their own actions.

The dynamics of the new product production is described by a discrete equation:

$$x_t = x_{t-1} + u_t, \quad t = 1, T, \quad (1)$$

where  $x_t$  - cumulative production volume of a new product for the  $t$ -th time period,  $t$  - number of the time period,  $u_t$  - production volume of new product in the period  $t$ ,  $T$  - amount of time periods.

In the initial period, the number of products produced is as follows:

$$x_0 = X_0, \quad (2)$$

In the final period, the cumulative volume of finished products should be equal to the specified volume:

$$x_T = X_0 + R, \quad (3)$$

where  $R$  – specified number of finished products.

The production volume of products in each period  $t$  is imposed by the following restrictions:

$$Q^{\min} \leq u_t \leq Q^{\max}, \quad t = 1, T, \quad (4)$$

where  $Q_i^{\min}$  – minimum production volume, taking into account technological and logistic requirements,  $Q^{\max}$  – maximum production volume limited by the production capacity of equipment.

The labor costs of manufacturing the product in period  $t$  are defined as the multiplication of the product labour intensity  $c_{pt}$  and production volume in this period  $u_t$ :

$$C_{pt} = c_{pt}u_t, \quad t = 1, T. \quad (5)$$

Due to the learning effect, the product labor intensity decreases depending on the cumulative production volume [1]:

$$c_{pt} = a_p x_{t-1}^{-b_p}, \quad (6)$$

where  $a_p$  – labor costs of agents for the production of the first product,  $b_p$  – speed of a product labor intensity reduction with increase in the cumulative production volume.

Let us substitute the expression (6) in the formula (5) and find the labor costs for manufacturing the finished product in  $t$  period:

$$C_{pt} = a_p x_{t-1}^{-b_p} u_t, \quad t = 1, T. \quad (7)$$

The production volumes of the finished product  $u_t$  are chosen by the principal while planning production activities based on its goal function.

Several options are considered as a goal function of the principal.

1. Minimization of the discounted cumulative labor costs of all agents producing a new product:

$$J_p = \sum_{t=1}^T \frac{a_p x_{t-1}^{-b_p} u_t}{(1+r_p)^t} \rightarrow \min, \quad (8)$$

where  $r_p$  – discount rate of the principal.

2. Maximization of discounted profit from the production of a new product:

$$J_p = \sum_{t=1}^T \frac{P_{t-1} u_t - a_p x_{t-1}^{-b_p} u_t}{(1+r_p)^t} \rightarrow \max, \quad (9)$$

where  $P_{t-1}$  - price of the new product.

3. Maximization of the production volume of a new product:

$$J_p = x_T \rightarrow \max. \quad (10)$$

4. Minimization of implementation time of the new product development project:

$$J_p = T \rightarrow \min. \quad (11)$$

Let us formulate the dynamic task of planning the production volumes of a new product for the principal.

The dynamic planning task consists of finding optimal production volumes  $u_t^{opt}$ ,  $t = 1, n$  satisfying the constraint (4), which transfer the dynamic production system (1) from the initial state (2) to the final state (3) and deliver the extremum of one of the goal functions of the principal (8) - (11).

To solve the formulated optimal control problem, Bellman's dynamic programming method [7], implemented in the Free pascal programming environment, was used. Formulation and solutions for dynamic planning problems are given in [8].

As a result of solving the problem of dynamic planning, the principal determines the optimal production volumes of a new product  $u_t^{opt}$ ,  $t = 1, n$ . To implement the project, it is necessary that the production volumes of the product and parts match. But the choice of the actual production volumes of parts, from which the finished product is assembled, is made by agents upon their own interests. The principal influences the production process through the mechanism of material incentives, encouraging agents economically to fulfill the planned production volumes.

## 2.2. The decision-making model of an agent

The dynamics of the production activity of the  $i$ -th agent who manufactures the parts for a new product is described by a discrete equation:

$$y_t = y_{t-1} + v_t, \quad t = 1, T, \quad (12)$$

where  $y_t$  - cumulative production volume by the agent for the  $t$ -th time period,  $v_t$  - production volume by the agent at the period  $t$ .

The choice of the production volume  $v_t$  at the period  $t$  is the agent's management.

In the initial period, the number of parts produced by agent is known:

$$y_0 = Y_0, \quad (13)$$

in the final period, the cumulative volume of the parts produced by the agent should be equal to the specified volume:

$$y_T = Y_0 + R, \quad (14)$$

where  $R$  – specified number of parts, which coincides with the number of finished products.

The production volume of parts in each period  $t$  is imposed by the following restrictions:

$$Q_i^{\min} \leq v_t \leq Q^{\max}, \quad t = 1, T, \quad (15)$$

where  $Q_i^{\min}$  – minimum parts production volume, considering technological and logistic requirements,  $Q^{\max}$  – maximum parts production volume limited by the production capacity of equipment.

Restrictions on the parts production volume coincide with the restriction on the final product production volume.

The agent labor costs in monetary terms in the period  $t$  are defined as the multiplication of part labor intensity  $c_{at}$ , the cost of the norm-hour at the enterprise  $s$ , and the parts production volume in this period  $v_t$  :

$$C_{at} = sc_{at}v_t, \quad t = 1, T. \tag{16}$$

Due to the learning effect, the labor intensity of parts decreases depending on the cumulative production volume [1]:

$$c_{at} = a_a y_{t-1}^{-b_a}, \tag{17}$$

where  $a_a$  – agent’s labor costs for the first part production,  $b_a$  – speed of a part labor intensity reduction with increase in the cumulative production volume.

Substituting the expression (17) in the formula (16) the agent’s labor costs in the period  $t$  can be found:

$$C_{at} = sa_a y_{t-1}^{-b_a} v_t, \quad t = 1, T. \tag{18}$$

Principal uses a dynamic proportional incentive system for the project implementation:

$$\sigma_t(\alpha_t, v_t) = \alpha_t v_t, \quad t = 1, T, \tag{19}$$

where  $\alpha_t$  – agent’s payment rate for a manufactured part in the periods  $t=1, T$  (parameters of the incentive function).

Parameters of the incentive function are principal management, through which the principal affects the economic interests of the agent, stimulating him to choose the planned production volumes.

The amount of financial incentives for an agent should not exceed a limited payroll budget  $F$ :

$$\sum_{t=1}^T \alpha_t v_t \leq F, \quad i = 1, n. \tag{20}$$

The goal function of the agent is to maximize the discounted income:

$$J_a = \sum_{t=1}^T \frac{\sigma_t(\alpha_t, v_t) - C_{at}}{(1+r_a)^t} \rightarrow \max, \tag{21}$$

where  $r_a$  – agent’s discount rate.

The agent income is the difference between the financial incentives and his labor costs, expressed in monetary terms.

Taking into account (18) and (19), the agent goal function (21) will be as follows:

$$J_a = \sum_{t=1}^T \frac{\alpha_t v_t - sa_a y_{t-1}^{-b_a} v_t}{(1+r_a)^t} \rightarrow \max. \tag{22}$$

The stated problem is the problem of discrete system optimal control for an agent. The solution of the stated problem is an optimal control  $v_t^{opt}$ ,  $t=1, n$ , satisfying constraint (15), which transfers the discrete system (12) from the initial state (13) to the final state (14) and maximizes the agent’s discounted income (22). Alongside, the solution of the agent’s optimization task depends on the payment rates for the production unit  $\alpha_t$ ,  $t=1, n$ , which are given by the principal.

To solve the formulated optimal control problem, Bellman's dynamic programming method [7], implemented in the Free pascal programming environment, was used.

### 2.3. Algorithm for solving the dynamic incentive problem

Let us formulate the problem of agent dynamic incentive:

$$\begin{aligned} J_p &= \sum_{t=1}^T g_p(t, u_t, x_{t-1}) \rightarrow \max(\min), \\ x_t &= x_{t-1} + u_t, \quad t = 1, T, \\ Q^{\min} &\leq u_t \leq Q^{\max}, \quad t = 1, T, \\ x_0 &= X_0, \\ x_T &= X_0 + R, \\ J_a &= \sum_{t=1}^T \frac{\alpha_t v_t - sa_a y_{t-1}^{-b_a} v_t}{(1+r_a)^t} \rightarrow \max, \\ \sum_{t=1}^T \alpha_t v_t &\leq F, \\ y_t &= y_{t-1} + v_t, \quad t = 1, T, \\ Q^{\min} &\leq v_t \leq Q^{\max}, \quad t = 1, T, \\ y_0 &= Y_0, \\ y_T &= Y_0 + R. \end{aligned} \tag{23}$$

where  $g_p$  – specific form of the goal function of the principal, is determined by one of the expressions (8) - (11).

The formulated problem (23) represents a dynamic game, the solution of which determines the conditions for coordination between the principal and the agent. The solution of the dynamic game will be the parameters of the incentive system and the production volumes of the parts  $\alpha_t^*, v_t^* = u_t^*$ ,  $t=1, T$ , that deliver the extremum to the goal functions of the principal and the agent.

Let us formulate an algorithm for solving the dynamic problem of proportional incentive.

1. The optimal control problem for the principal is solved using Bellman's dynamic programming method, the new product optimal planned volumes  $u_t$ ,  $t=1, T$  are found.

2. The principal management is being set – the parameters of the incentive function  $\alpha_t$ ,  $t=1, T$  should meet the condition of not exceeding the agent's payroll budget:

$$\sum_{t=1}^T \alpha_t u_t(\alpha_t) \leq F.$$

3. For the given parameters of the incentive function  $\alpha_t, t=1, T$  the optimum control problem for the agent is solved using the Bellman dynamic programming method, and the agent's optimal response is determined as an actual part production volumes  $v_t(\alpha_t), t=1, T$ .

4. The coincidence condition of the planned and actual production volumes that the agent chooses is checked:

$$\sum_{t=1}^T (u_t - v_t(\alpha_t))^2 \leq \varepsilon,$$

where  $\varepsilon$  - predetermined small value. If the condition is satisfied, then the dynamic proportional incentive problem is solved. If not, then the parameters of the incentive function  $\alpha_t, t=1, T$  are changed and the step 2 is repeated. The principal goal functions (8) - (11) considered above do not depend on the function of financial incentives. Let's formulate the goal functions of the principal, taking into account the expenses of the principal for the agent incentives.

1. Minimization of discounted costs for agent incentives:

$$J_p = \sum_{t=1}^T \frac{\alpha_t v_t}{(1+r_p)^t} \rightarrow \min. \quad (24)$$

2. Maximization of discounted income from the agent's production activity:

$$J_p = \sum_{t=1}^T \frac{(p_t - \alpha_t) v_t}{(1+r_p)^t} \rightarrow \max, \quad (25)$$

where  $p_t$  - part price produced by the agent.

In this case, the task of the agent dynamic incentive is as follows:

$$\begin{aligned} J_p &= \sum_{t=1}^T g_p(t, v_t, \alpha_t) \rightarrow \max(\min), \\ J_a &= \sum_{t=1}^T \frac{\alpha_t v_t - s a_a y_{t-1}^{-b_a} v_t}{(1+r_a)^t} \rightarrow \max. \\ y_t &= y_{t-1} + v_t, \quad t=1, T, \\ Q^{\min} &\leq v_t \leq Q^{\max}, \quad t=1, T, \\ y_0 &= Y_0, \\ y_T &= Y_0 + R. \end{aligned} \quad (26)$$

where  $g_p$  - specific form of the goal function of the principal, is determined by one of the expressions (24) - (26).

The solution of the dynamic game will be the parameters of the incentive function and the parts production volumes  $\alpha_t^*, v_t^*, t=1, T$ , that deliver the extremum to the goal functions of the principal and the agent. Let us formulate an algorithm for solving the dynamic problem of proportional incentive in the case when the goal function of the principal depends on the incentive costs (26):

1. The parameters of the principal incentive function  $\alpha_t, t=1, T$  are given.

2. For the given parameters of the incentive function  $\alpha_t, t=1, T$  the optimum control problem for the agent is solved using the Bellman dynamic programming method and the agent's optimal response is determined - the actual part production volumes  $v_t(\alpha_t), t=1, T$ .

3. The found agent's response  $v_t(\alpha_t)$  is substituted into the optimal control problem for the principal, which is solved by Bellman's dynamic programming method. Thus the optimal parameters of the incentive function  $\alpha_t(v_t(\alpha_t)), t=1, T$  are found.

4. The condition of coincidence of the parameters of the incentive function at the given iteration and the previous one is checked:

$$\sum_{t=1}^T (\alpha_t(v_t(\alpha_t)) - \alpha_t)^2 \leq \varepsilon,$$

where  $\varepsilon$  - predetermined small value. If the condition is satisfied, then the dynamic proportional incentive problem is solved. If not, the parameters of the incentive function must be changed and the step 2 repeated.

### 3. Results and Discussion

The problem of dynamic incentive with the following initial data is considered.

$$\begin{aligned} J_p &= \sum_{t=1}^{12} \frac{42,64 x_{t-1}^{-0,7} u_t}{(1+r_p)^t} \rightarrow \min \\ x_t &= x_{t-1} + u_t, \quad t=1,12, \\ 0 &\leq u_t \leq 40, \quad t=1,12, \\ x_0 &= 1, \quad x_T = 241, \\ J_a &= \sum_{t=1}^T \frac{\alpha_t v_t - 3837,6 y_{t-1}^{-0,1} v_t}{(1+r_a)^t} \rightarrow \max, \\ \sum_{t=1}^T \alpha_t v_t &\leq 960000, \\ y_t &= y_{t-1} + v_t, \quad t=1,12, \\ 0 &\leq v_t \leq 40, \quad t=1,12, \\ y_0 &= 1, \quad y_T = 241. \end{aligned}$$

The numerical solution of the problem was obtained using the proposed algorithm.

The planned trajectory corresponds to the trajectory, which minimizes labor costs of agents and coincides with the rate of a product mastering  $b_p = -0,7$ . At a constant parameters of an incentive function the agent chooses an actual trajectory of the production cumulative volume, which corresponds to an agent's learning rate  $b_a = -0,1$ . Figure 1 shows the planned and actual trajectory of the cumulative volume of production of a new product.

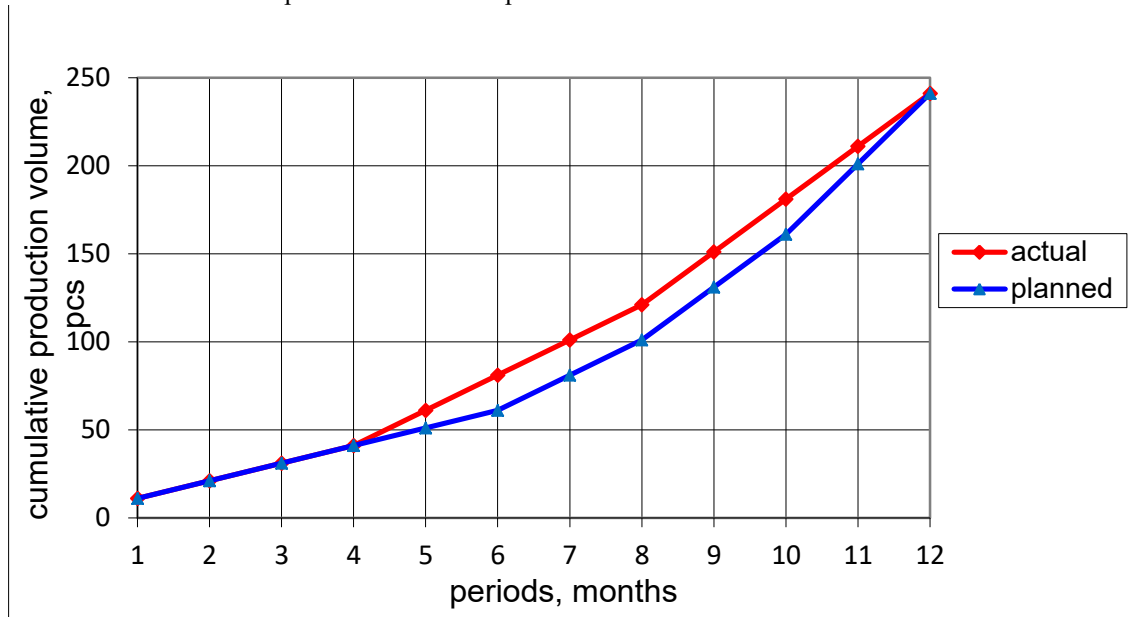


Fig. 1. The planned and actual trajectory of the cumulative volume of production.

In the process of numerical research the following results were obtained:

1. Constant rate of payment does not affect the choice of the agent of the actual cumulative production trajectory. In this case, the trajectory is determined only by the speed of the agent learning.
2. Coordination of the interests of the principal and the agent is achieved using a variable wage rate, which depends on the cumulative volume of production.
3. Applying a payment rate in the form of a linearly increasing function from the cumulative production volume  $\alpha_t = ky_t + d$  makes an agent select the trajectories with a higher learning rate. The agent moves from the "slow" trajectory to the "fast", more "convex" one. The larger value of the control parameter  $k$  corresponds to a more "convex" trajectory selected by the agent (Fig. 2).
4. Applying a payment rate in the form of a linearly decreasing function from the cumulative production volume leads to the agent selection of trajectories with a lower learning rate. The agent moves from "fast" trajectory to "slow", less "convex" one. The larger modulo value of the control parameter  $k$  corresponds to the agent's chose of a less "convex" trajectory (Fig. 3).

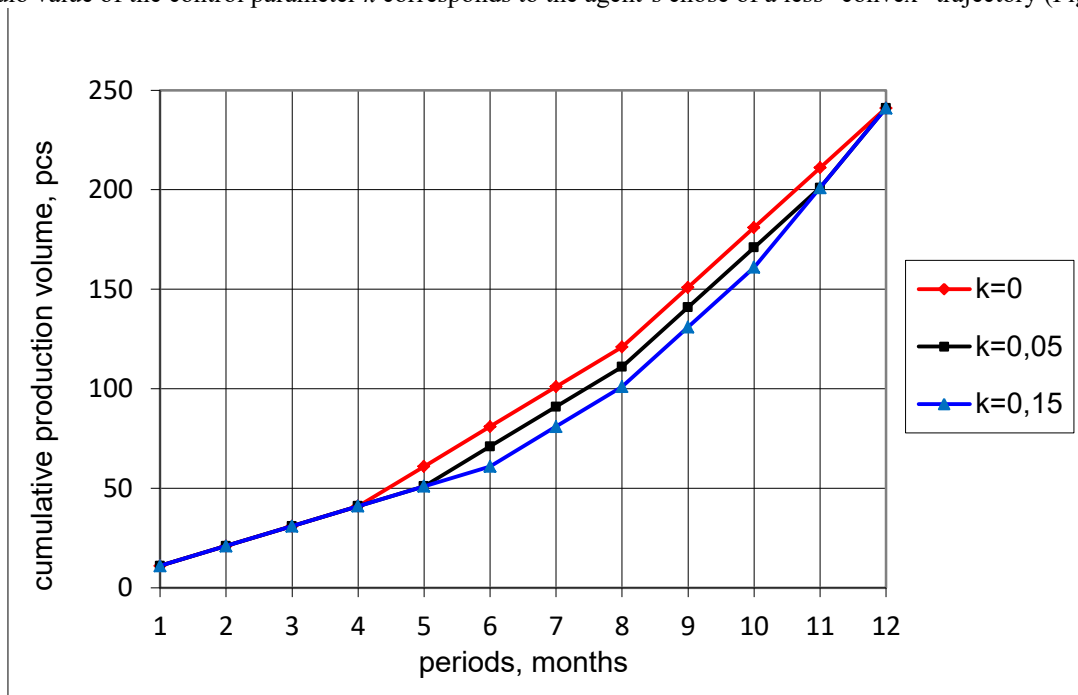


Fig. 2. Influence of parameter  $k$  on the actual trajectory of the cumulative volume of production.

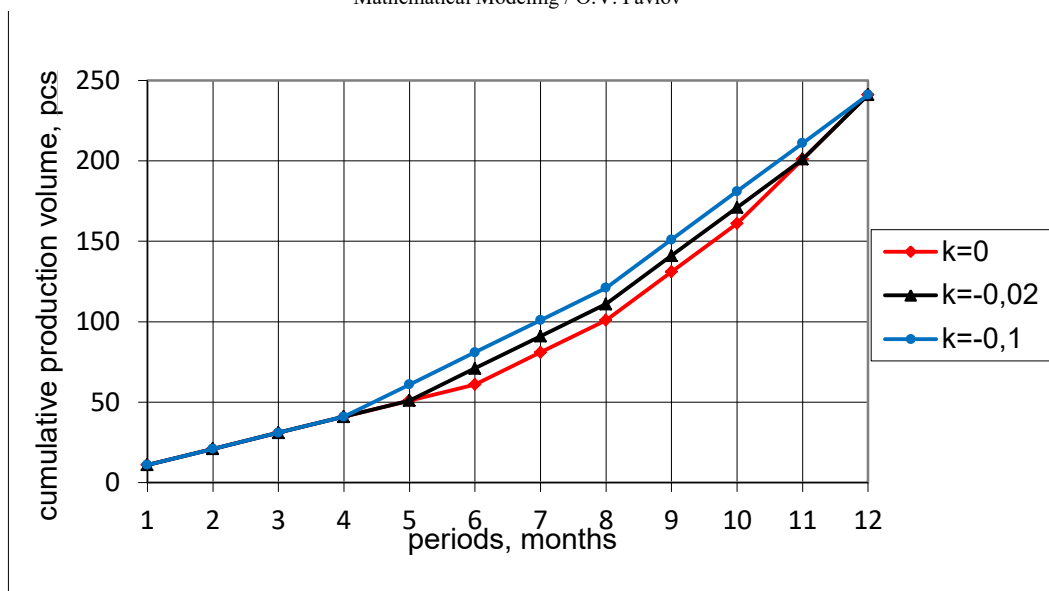


Fig. 3. Influence of parameter  $k$  on the actual trajectory of the cumulative volume of production.

#### 4. Conclusion

Dynamic decision-making models for the principal and agent in the project for the production of a new product have been developed. The problems of agent dynamic incentive for various goal functions of the principal are formulated. Two options are considered: the goal functions include the costs of material incentives for the agent and do not include them.

For both variants, a numerical algorithm is proposed, based on a sequential solution of two optimal control problems, which are solved using Bellman's dynamic programming method.

A numerical example of the problem solution for the principal goal function, which does not depend on the agent incentive costs, is given. It is shown that the application of the agent salary rate in the form of a linear function, that depends on the cumulative production volume, ensures that the agent selects the planned trajectory of the principal.

#### Acknowledgements

The reported study was funded by RFBR and Samara region according to the research project № 17-46-630606.

#### References

- [1] Wright TP. Factors affecting the cost of airplanes. *Journal of the aeronautical sciences* 1936; 3(4): 122–128.
- [2] Novikov DA. Models of learning in the work process. *Large-scale Systems Control* 2007; 19: 5–22.
- [3] Gorelik VA, Gorelov MA, Kononenko AF. *Analysis of conflict situations in management systems*. M.: Radio and communication, 1991.
- [4] Basar T, Olsder GJ. *Dynamic Noncooperative Game Theory*. Philadelphia: SIAM, 1999.
- [5] Ougolnitsky GA. *Management of sustainable development of active systems*. Rostov-on-Don: Publishing of Southern Federal University, 2016; 940 p.
- [6] Dockner E, Jorgensen S, Long NV, Sorger G. *Differential games in economics and management Science*. Cambridge: Cambridge University Press, 2000.
- [7] Bellman R. *Dynamic programming*. M.: Foreign Literature Publishing House, 1960.
- [8] Pavlov OV. Dynamic optimization of production activities of the enterprise taking into account learning curve effect. *Vestnik of Samara State Economics University* 2015; 3(125): 88–92.



# Nonlinear eigenvalue problems in fracture mechanics: eigenspectra and eigenfunctions

A.A. Peksheva<sup>1</sup>, L.V. Stepanova<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

## Abstract

The study is aimed at analytical determination of eigenfunctions of the nonlinear eigenvalue problems following from the crack problems in power law materials under mode III loading and mixed mode (mode I and mode II) loading. The study is based on the perturbation theory technique (the small artificial parameter method) allowing us to find the analytical solution for the eigenfunctions in the closed form in the case of mode III crack problems and to derive the analytical approximations for mixed mode (Mode I and Mode II) crack problems. The method of analytical determination of eigenfunctions of the nonlinear eigenvalue problem is presented.

*Keywords:* crack tip field; antiplane shear; series expansion method; nonlinear eigenvalue problem; eigensolution; eigenspectrum; eigenfunction; closed-form solution

---

## 1. Introduction

Hutchinson [1,2] and Rice and Rosengren [3] derived the classical Hutchinson-Rice-Rosengren (HRR) stress field in plane stress and plane strain for a crack in power-law hardening materials. They solved the governing nonlinear differential equations for the stress function (describing a nonlinear eigenvalue problem) by a numerical procedure. This solution exclusively describes the dominant singular crack-tip field. Up to now for plane stress and plane strain neither higher order eigensolutions are known nor an analytical solutions for the dominant field is available in the literature [4]. As it is noted in [4] the corresponding antiplane shear problem of a notch with traction free faces in an nonlinear hardening material first was analyzed by Neuber [5] and Rice [6,7] by the use of the hodograph transformation. The brief review of classical results for antiplane deformation of cracked bodies can be found in [4]. Since then researchers have tried to derive the analytical solutions for the antiplane shear problem as well for mode I and mode II crack problems [4 - 25]. Thus in [4] the higher order fields at a notch or a crack tip in the power-law hardening material under mode III loading (longitudinal shear) are studied. The authors derived a closed form solution for the eigenvalues, determining the asymptotic behavior of the fields analytically applying the perturbation technique. It is shown that the eigenvalues of the nonlinear eigenvalue problem solely depend on the eigenvalues of the corresponding linear problem and on the hardening exponent. It is noted that it is valid for all three combinations of homogeneous boundary conditions. A method is derived for constructing the higher order eigensolutions from dominant singular solutions.

The asymptotic stress and strain fields near the crack tip under antiplane shear in an elastic power-law hardening material are developed in [8]. Using an asymptotic expansion and separation of variables for the stress function, a series solution for all of the hardening exponents can be obtained. The stress exponents for the higher order terms are analytically determined; the angular distributions which are governed solely by plastic strains are also analytically obtained. Good agreement with the finite element solutions confirms the proposed approach. It is further demonstrated that the first three terms, controlled by two parameters, can be used to characterize the crack tip stress and strain fields with various hardening exponents. In [8] a series solution with assumed separation of variables form for the stress and strain fields near the crack tip in an elastic power-law hardening material under antiplane shear has been developed. The leading order term is analytically obtained by solving a nonlinear eigenvalue problem. The higher order fields are governed by either linear homogeneous eigenvalue equations or linear nonhomogeneous governing equations. The stress exponents of higher order fields for any hardening exponent are analytically determined. The governing equations for higher order terms which are controlled solely by the plastic strains can also be obtained analytically. However, the governing equations governed by elastic and plastic strains need to be solved numerically. With the analytically determined stress exponents, distinct regions resulting from different strain hardening exponents where the higher order terms up to the fourth order attributed to the plastic strains or elastic and plastic strains can be identified. It has been demonstrated that a truncated three term solution with two parameters accurately characterizes the crack tip stress and strain fields. The paper [9] considers the mechanical fields near the tip of a crack deformed by an anti-plane shear at infinity for a class of nonlinear elastic materials. For brittle materials rupture occurs when a maximal stretch is reached. Taking into account of this critical value, the crack is replaced by a totally damaged zone of finite thickness named a quasicrack. Inside this domain, the stress is identically zero and the shape of the boundary between damaged and undamaged body is found analytically. C. Stolz has determined [9] the shape of the damaged zone under anti-plane shear condition for hyperelastic brittle material. The analytical results are a generalization of preceding results obtained in [10] for brittle materials. The thickness of the damaged zone is determined by the critical strain energy at rupture and the loading. C. Stolz has extended [9] the theory for a more complex constitutive law and recovered results obtained many years ago. The case of power law and its extension on a class of non linear elastic law is discussed with and without brittle damage. With brittle damage one obtains for mode III loading, the geometry of the quasi-crack proposed by Neuber for hardening law has been found. This result is extended to some cases of softening especially for a generalization of the special material introduced in [11]. In [12] the stress and strain fields near the tip of a steady-state growing crack are examined for elastic-viscous materials. A solution to this problem has been originally

derived by Hui and Riedel [13], with some paradoxes such as the non-dependence of the far fields with respect to the crack growth rate. A two-scale match asymptotic analysis is suggested in [12] to overcome these paradoxes. The scale factor is completely determined by the material properties. The inner scale may be considered as a boundary layer, where the stress field completely described by a serial Fourier analysis. The unit value fits with the Hui and Riedel solution [13].

In [14] it is noted that there exist many nonlinear eigenvalue problems in science and engineering. Nonlinear eigenvalue problems are much more difficult to solve than linear ones. Many nonlinear eigenvalue problems have multiple eigenvalues and eigenfunctions. However, even by means of numerical techniques, it is difficult to find all multiple solutions of a nonlinear differential equation. There are some analytic techniques for nonlinear eigenvalue problems, which are based on either perturbation techniques [15 - 20], or traditional non-perturbation methods such as the Adomian decomposition method [21 - 24], Lyapunov artificial small parameter method [25], and so on. It is well known that perturbation techniques are too strongly dependent upon small physical parameters. Besides, convergence radius of perturbation series is often small, so that perturbation approximations are valid in general only for problems with weak nonlinearity. In [14] a general analytic approach for nonlinear eigenvalue problems is described. Two physical problems are used as examples to show the validity of this approach for eigenvalue problems with either periodic or non-periodic eigenfunctions. Unlike perturbation techniques, this approach is independent of any small physical parameters. Besides, different from all other analytic techniques, it provides a simple way to ensure the convergence of series of eigenvalues and eigenfunctions so that one can always get accurate enough approximations. Finally, unlike all other analytic techniques, this approach provides great freedom to choose an auxiliary linear operator so as to approximate the eigenfunction more effectively by means of better base functions. This approach provides us a new way to investigate eigenvalue problems with strong nonlinearity. In [14] an analytic approach to get series solutions of nonlinear eigenvalue problems is described by means of two examples. This analytic approach is valid for nonlinear eigenvalue problems with either periodic or non-periodic eigenfunctions, and thus is rather general. All of the series solutions agree well with exact or numerical results, and this fact shows the validity of the analytic approach realized in [14]. The author of [14] shows that the analytic approach proposed has some obvious advantages. First of all, unlike perturbation techniques, it is independent of any small physical parameters: it is valid no matter whether or not there exist any small physical parameters in governing equations and/or boundary conditions. Second, different from other traditional techniques, it provides us a simple way to ensure the convergence of series solution of eigenvalue and eigenfunction, so that one can always get accurate enough approximations. Thus, this approach can be applied to solve eigenvalue problems with strong nonlinearity. Third, unlike all other analytic techniques, this approach provides us great freedom to choose an auxiliary linear operator so as to approximate the eigenfunction more effectively by means of better base functions. Therefore, this approach can be widely applied to solve strongly nonlinear eigenvalue problems in science and engineering, no matter whether the corresponding eigenfunction is periodic or not. Analytical approaches to nonlinear eigenvalue problems following from fracture mechanics analysis as well as the perturbation theory methods in general attract many researches in the past and nowadays [9,12, 26-36].

The present paper is aimed at analytical determination of eigenfunctions of the nonlinear eigenvalue problems arising from the antiplane shear crack problems in power-law materials. The goal of the study is to develop the analytical approach for determination of the eigenfunctions of the nonlinear eigenvalue problems by the perturbation theory methods. The paper continues the perturbation theory method applied for the mode III crack problems in [4, 33-35].

## 2. Fundamental equations. Statement of the problem

Singular fields and higher order fields in the vicinity of the crack in a power-law material under longitudinal shear are investigated in many works [1-4]. The very neat approach for the nonlinear eigenvalue problem has been proposed in [4] where singular fields and higher order fields near a sharp notch in a power-law material under longitudinal shear are analyzed. In [4] using the perturbation theory method the whole set of eigenvalues is determined. A closed form solution for the eigenvalues determining the asymptotic behavior of the fields is analytically derived by applying the perturbation method. However nowadays along with the eigenvalues and along with the eigenspectrum of the problem it is important to know the eigenfunctions corresponding to the eigenvalues derived. In the present paper the closed form solution for the eigenfunctions for the crack tip fields is obtained. It is shown that the asymptotic analysis and methods of summability allow us to derive the analytical solution for the eigenfunctions of the nonlinear eigenvalue problem. The constitutive behavior shall be given by the power law of the Ramberg-Osgood type

$$\varepsilon_{rz} = 3B\sigma_e^{n-1}\sigma_{rz}/2, \quad \varepsilon_{\theta z} = 3B\sigma_e^{n-1}\sigma_{\theta z}/2, \quad \sigma_e = \sqrt{\sigma_{rz}^2 + \sigma_{\theta z}^2} \quad (1)$$

where  $\sigma_e$  is the effective stress, and  $B$  and the hardening exponent  $n$  are materials constants determined experimentally. The equilibrium equation and the compatibility equation in polar coordinates are written as

$$r\sigma_{rz,r} + \sigma_{\theta z,\theta} + \sigma_{rz} = 0, \quad \varepsilon_{r,\theta} = (r\varepsilon_{\theta z})_{,r}. \quad (2)$$

Introducing the stress function  $\chi(r, \theta)$  such as  $\sigma_{rz} = r^{-1}\chi_{,\theta}$ ,  $\sigma_{\theta z} = -\chi_{,r}$ , the equilibrium equation is satisfied identically. The asymptotic solution is searched in the separable form

$$\chi(r, \theta) = r^s f(\theta). \quad (3)$$

Introducing the asymptotic presentation (3) into (1) and (2) one can obtain the nonlinear ordinary differential equation (NODE)

$$f_e^2 f'' + (n-1)f'^2 (f'' + s^2 f) + [(s-1)n+1]sf_e^2 f = 0, \quad f_e = \sqrt{(f')^2 + (sf)^2}. \quad (4)$$

The solution of equation (4) should satisfy the conventional traction-free boundary conditions on the crack faces:

$$f(\theta = \pm\pi) = 0. \quad (5)$$

In conjunction with the boundary conditions (5) the nonlinear ordinary differential equation (4) describes a nonlinear eigenvalue problem where the unknown eigenvalue  $s$  and the eigenfunction  $f(\theta)$  depend on the boundary conditions and the hardening exponent  $n$ . The unknown eigenvalue  $s$  and the eigenfunction  $f(\theta)$  should be found as a part of the solution. In [4] the subtle approach allowing us to find the closed form solution for the eigenvalue has been used. Thus, hereafter we will consider that all the eigenvalues are known.

### 3. Eigenvalues and eigenfunctions of the antiplane shear problem

An analytical expression for the eigenfunctions of the nonlinear equation (4) can be derived by applying the perturbation technique. For this purpose, the eigenvalue is represented in the form [4]  $s = s_0 + \varepsilon$ , where  $s_0$  is the eigenvalue of the "undisturbed" linear problem and  $\varepsilon$  is the deviation on account of the nonlinearity. Furthermore, the hardening exponent  $n$  and the stress function  $f(\theta)$  are represented as power series

$$n = n_0 + \varepsilon n_1 + \varepsilon^2 n_2 + \varepsilon^3 n_3 + \varepsilon^4 n_4 + \varepsilon^5 n_5 + \dots = \sum_{j=0}^{\infty} \varepsilon^j n_j, \quad (6)$$

$$f(\theta) = f_0(\theta) + \varepsilon f_1(\theta) + \varepsilon^2 f_2(\theta) + \varepsilon^3 f_3(\theta) + \varepsilon^4 f_4(\theta) + \varepsilon^5 f_5(\theta) + \dots = \sum_{j=0}^{\infty} \varepsilon^j f_j(\theta)$$

where  $n_0$  and  $f_0(\theta)$  are referred to the linear "undisturbed" problem. Introducing the asymptotic expansions (6) and collecting terms of equal power in  $\varepsilon$ , the following set of linear differential equations is obtained

$$\varepsilon^0: f_0'' + s_0^2 f_0 = 0 \quad (7)$$

$$\varepsilon^1: f_1'' + s_0^2 f_1 = -[2 + n_1(s_0 - 1)]s_0 f_0 \quad (8)$$

$$\varepsilon^2: f_2'' + s_0^2 f_2 = -[n_1 + n_2(s_0 - 1)]s_0 f_0 - [1 + n_1(s_0 - 1)]f_0 - 2s_0 n_1 f_0'^2 / g_0 \quad (9)$$

$$\varepsilon^3: f_3'' + s_0^2 f_3 = -[n_1 + n_2(s_0 - 1)]s_0 f_0 g_1 / g_0 - [1 + n_1(s_0 - 1)]f_0 g_1 / g_0 - 2s_0 n_2 f_0'^2 / g_0 - s_0 [n_2 + n_3(s_0 - 1)]f_0 - [n_1 + n_2(s_0 - 1)]f_0 - (f_2'' + s_0^2 f_2)(g_1 + n_1 f_0'^2) / g_0 - n_1 f_0'^2 / g_0 \quad (10)$$

where  $g_0 = f_0'^2 + s_0^2 f_0^2$ ,  $g_1 = 2f_0' f_1' + 2s_0 f_0'^2 + 2s_0^2 f_0 f_1$ . The boundary conditions are the conventional traction free conditions on the crack surfaces:  $f_k(\theta = \pm\pi) = 0$ . It implies that all the functions  $f_k(\theta)$  have to satisfy the same condition. It is known that when the boundary value problem for the homogeneous differential equation has a nontrivial solution, the corresponding boundary value problem for inhomogeneous differential equation has a solution if and only if the inhomogeneous part satisfies the solvability condition [20]. The solvability condition permits to find the coefficients  $n_k$  in equations (6). The coefficients  $n_k$  have been found and the closed form solution was presented in [4]:

$$n = 1 + \frac{s_0}{s_0 - 1} \sum_{j=0}^{\infty} \left( \frac{-(2s_0 - 1)\varepsilon}{s_0(s_0 - 1)} \right)^j - \frac{1}{s_0 - 1} \sum_{j=0}^{\infty} \left( \frac{-\varepsilon}{s_0 - 1} \right)^j = \frac{2s_0 - 1}{(2s_0 - 1)s - s_0} - \frac{s}{s - 1}.$$

The asymptotic expansion for the hardening exponent when the HRR-type problem ( $s_0 = 1/2$ ) is considered takes the form

$$n = 1 - \frac{1}{s_0 - 1} \sum_{j=0}^{\infty} \left( \frac{-\varepsilon}{s_0 - 1} \right)^j = -\frac{s}{s - 1}.$$

It allows us to find the whole spectrum of the eigenvalues

$$s = \frac{(n+1)s_0 + (n-1)(2s_0 - 1)}{2n(2s_0 - 1)} + \frac{\sqrt{((n+1)s_0 + (n-1)(2s_0 - 1))^2 - 4n^2 s_0^2 (2s_0 - 1)}}{2n(2s_0 - 1)}.$$

Our aim is to study the possibility to derive the closed form solution for eigenfunctions. To obtain the closed form solution one can solve analytically the system of linear ordinary equations and two point boundary problems for these equations. Further one can analyze the structure of the solution and reveal the general features and inherent properties of the approximate solutions. For this purpose one can analyze the structure of the solutions of each boundary value problem obtained. In the case of a linear material the eigenfunctions and eigenvalues can be easily determined:

$$f_0(\theta) = 2\cos(\theta/2). \quad (11)$$

The solution of equation (8) satisfying the traction free boundary conditions  $f_1(\theta = \pm\pi) = 0$  can be written as

$$f_1(\theta) = -n_1 \cos(\theta/2) \quad (12)$$

The solution of equation (9) satisfying the traction free boundary conditions  $f_2(\theta = \pm\pi) = 0$  can be expressed as

$$f_2(\theta) = -\frac{1}{16} \frac{-n_1^2 + n_1^2 \cos^2 \theta + 8n_2 \cos \theta + 8n_2}{\cos(\theta/2)}. \quad (13)$$

The solution of the two point boundary value problem for equation (10) with the boundary conditions  $f_3(\theta = \pm\pi) = 0$  with respect to function  $f_3(\theta)$  has the form

$$f_3(\theta) = \frac{1}{32} \frac{n_1^3 \sin^2 \theta + n_1^3 \cos \theta + 3n_1^3 \cos^2(\theta) - n_1^3 \cos \theta \sin^2 \theta - 8n_1 n_2 \cos^2 \theta - 2n_1^3 \cos^3 \theta - 4n_1 n_2 \sin^2 \theta + 8n_1 n_2 - 6n_1^3}{\cos(\theta/2)} - \frac{1}{32} n_1 \frac{-2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta + 2n_1^2}{\cos(\theta/2)} + [-n_3 + n_1 n_2 + (n_2 - n_1^2) n_1] \cos(\theta/2). \quad (14)$$

Similarly one can find the solution of the boundary value problem for the subsequent function  $f_4(\theta)$ :

$$f_4(\theta) = -\frac{n_1}{64} \frac{n_1^3 \sin^2 \theta + n_1^3 \cos \theta + 3n_1^3 \cos^2(\theta) - n_1^3 \cos \theta \sin^2 \theta - 8n_1 n_2 \cos^2 \theta - 2n_1^3 \cos^3 \theta - 4n_1 n_2 \sin^2 \theta + 8n_1 n_2 - 6n_1^3}{\cos(\theta/2)} + \frac{1}{256} \frac{-24n_1^2 n_2 \cos^3 \theta - 4n_1^4 \sin^2 \theta + 24n_1^2 n_2 \cos \theta - 32n_1 n_3 \sin^2 \theta + 24n_1^2 n_2 - 24n_1^2 n_2 \sin^2 \theta \cos \theta + 64n_1 n_3 + 32n_2^2 - 72n_2 n_1^2}{\cos(\theta/2)} + \frac{1}{96} \frac{7n_1^4 - n_1^4 \cos^4 \theta - 5n_1^4 \cos \theta + 5n_1^4 \cos^3 \theta}{\cos(\theta/2)} - \frac{1}{32} \frac{4n_1^4 \cos^2 \theta + 8n_2^2 \cos^2 \theta - 3n_1^4 \cos \theta \sin^2 \theta + 4n_2^2 \sin^2 \theta + 16n_1 n_3 \cos^2 \theta}{\cos(\theta/2)} - \frac{1}{64} \frac{n_1^4 \sin^2 \theta \cos^2 \theta}{\cos(\theta/2)} - \frac{1}{2^9} \frac{(2n_1^2 - 2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta)^2}{\cos(\theta/2)} - \frac{1}{4} (-n_2 + n_1^2) \frac{-2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta + 2n_1^2}{\cos(\theta/2)} + [-n_3 + n_1 n_3 + (n_2 - n_1^2) n_2 + (n_3 - 2n_2 n_1 + n_1^3) n_1] \cos(\theta/2). \quad (15)$$

The solution of the boundary value problem for the function  $f_5(\theta)$  can be expressed as

$$f_5(\theta) = \frac{1}{32} \frac{(-n_3 + n_1 n_2 + (n_2 - n_1^2) n_1)(2n_1^2 - 2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta)}{\cos(\theta/2)} - \frac{n_1}{256} \frac{12n_1^2 \sin^2 \theta \cos \theta - 64n_1 n_3 \cos^2 \theta + 56n_1^2 n_2 \cos^2 \theta + 64n_2 n_3 + 32n_2^2 - 72n_1^2 n_2 + (16 \cdot 7/3)n_1^4 - 16n_2^2 \sin^2 \theta + 24n_1^2 n_2 \cos \theta}{\cos(\theta/2)} - \frac{n_1}{256} \frac{32n_1 n_3 \sin^2 \theta + 16n_1^4 \cos^2 \theta + 32n_2^2 \cos^2 \theta + (80/3)n_1^4 \sin^4 \theta + 5n_1^4 \cos^3 \theta - (16/3)n_1^4 \cos^4 \theta + 4n_1^4 \sin^2 \theta}{\cos(\theta/2)} - \frac{n_1}{256} \frac{4n_1^4 \sin^2 \theta \cos^2 \theta + 12n_1^2 n_2 \cos^3 \theta + 12n_1^2 n_2 \sin^2 \theta \cos^2 \theta + (80/3)n_1^4 \cos \theta}{\cos(\theta/2)} - \frac{n_1}{2^{11}} \frac{(2n_1^2 - 2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta)^2}{(\cos(\theta/2))^3} - \frac{1}{2^{10}} (2n_1^2 - 2n_1^2 \cos^2 \theta - n_1^2 \sin^2 \theta) \times \times n_1 \frac{n_1^3 \sin^2 \theta + n_1^3 \cos \theta + 3n_1^3 \cos^2 \theta - n_1^3 \sin^2 \theta \cos \theta - 8n_1 n_2 \cos^2 \theta - n_1^3 \cos^3 \theta - 4n_1 n_2 \sin^2 \theta - 3n_1^3 + 8n_1 n_2}{(\cos(\theta/2))^3} + \frac{n_1^5}{2^9} \frac{\cos^2 \theta \sin^2 \theta - \cos^5 \theta - \sin^4 \theta \cos \theta + \sin^2 \theta}{\cos^2(\theta/2)} + \frac{n_1^5}{2^9 \cdot 3} \frac{47 \cos \theta + 19 \cos^4 \theta - 43}{\cos^2(\theta/2)} + \frac{1}{2^6} \frac{-n_1^3 n_2 \sin^2 \theta + n_1^3 n_2 \sin^2 \theta \cos \theta - n_1^3 n_2 \sin^2 \theta \cos^2 \theta}{\cos^2(\theta/2)} - \frac{1}{2^8} \frac{n_1^3 n_2 \sin^4 \theta + 3n_1^5 \sin^2 \theta \cos \theta}{\cos^2(\theta/2)} + \frac{1}{2^6} \frac{n_1 n_2^2 \sin^2 \theta + 3n_1^3 n_3 \sin^2 \theta - 9n_1^2 n_3 - 9n_1 n_2^2 - 3n_1^2 n_3 \cos^3 \theta - 3n_1 n_2^2 \sin^2 \theta \cos \theta - 3n_1^2 n_3 \sin^2 \theta \cos \theta}{\cos^2(\theta/2)} + \frac{1}{2^6} \frac{n_1^5 \cos^2 \theta + 3n_1^2 n_3 \cos \theta + 3n_2^2 n_1 \cos \theta + 9n_2^2 n_1 \cos^2 \theta + 9n_1^2 n_3 \cos^2 \theta - 3n_2^2 n_1 \cos^3 \theta}{\cos^2(\theta/2)} + \frac{1}{2^3} \frac{n_1 n_4 + n_2 n_3 - n_1^3 n_2 \cos^2 \theta - n_1 n_4 \cos^2 \theta}{\cos^2(\theta/2)} + \frac{1}{2^4} \frac{5n_1^3 n_2 \cos^3 \theta - 5n_2 n_1^3 \cos \theta - n_1^3 n_2 \cos^4 \theta + 7n_1^3 n_2}{\cos^2(\theta/2)} - \frac{1}{2^4} \frac{n_2 n_3 \sin^2 \theta + n_1 n_4 \sin^2 \theta}{\cos^2(\theta/2)} + \frac{1}{2^7} \frac{n_1^5 \sin^4 \theta - n_1^5 \sin^2 \theta \cos^3 \theta}{\cos^2(\theta/2)} + [(n_3 - 2n_1 n_2 + n_1^3) n_2 + (n_2 - n_1^2) n_3 + n_1 n_4 - n_5 + (n_4 - 2n_1 n_3 - n_2^2 + 3n_2 n_1^2 - n_1^4) n_1] \cos(\theta/2) + \frac{1}{2^6} (-n_2 + n_1^2) \frac{(n_1^3 \sin^2 \theta + n_1^3 \cos \theta + 3n_1^3 \cos^2 \theta - n_1^3 \sin^2 \theta \cos \theta - 8n_1 n_2 \cos^2 \theta - n_1^3 \cos^3 \theta - n_1 n_2 \sin^2 \theta - 3n_1^3 + n_1 n_2)}{\cos(\theta/2)}. \quad (16)$$

The sequence of the solutions obtained has been analyzed carefully. The following perturbation series expansions based on the generalized multinomial theorem [36] are used

$$\left( \sum_{i=0}^{\infty} \varepsilon^i a_i \right)^m = (a_0 + \varepsilon a_1 + \varepsilon^2 a_2 + \varepsilon^3 a_3 + \dots + \varepsilon^n a_n + \dots)^m = \lambda_0 + \varepsilon \lambda_1 + \varepsilon^2 \lambda_2 + \varepsilon^3 \lambda_3 + \dots + \varepsilon^n \lambda_n + \dots \quad (17)$$

where  $\lambda_n$  is the coefficient of the term  $\varepsilon^n$ . The asymptotic expansion of  $(a_0 + \varepsilon a_1 + \varepsilon^2 a_2 + \varepsilon^3 a_3 + \dots + \varepsilon^n a_n + \dots)^m$  can be determined as follows [36]:

$$\begin{aligned}
 & (a_0 + \varepsilon a_1 + \varepsilon^2 a_2 + \varepsilon^3 a_3 + \dots + \varepsilon^n a_n + \dots)^m = \\
 & = \sum_{k_0=0}^{\infty} \sum_{k_1=0}^{k_0} \dots \sum_{k_{n-1}=0}^{k_{n-2}} \dots \binom{m}{k_0} \binom{k_0}{k_1} \dots \binom{k_{n-1}}{k_n} \dots a_0^{m-k_0} (\varepsilon a_1)^{k_0-k_1} \dots (\varepsilon^n a_n)^{k_{n-1}-k_n} \dots = \\
 & = \sum_{k_0=0}^{\infty} \sum_{k_1=0}^{k_0} \dots \sum_{k_{n-1}=0}^{k_{n-2}} \dots \binom{m}{k_0} \binom{k_0}{k_1} \dots \binom{k_{n-1}}{k_n} \dots a_0^{m-k_0} a_1^{k_0-k_1} \dots a_n^{k_{n-1}-k_n} \dots \varepsilon^{0+(k_0-k_1)+2(k_1-k_2)+\dots+n(k_{n-1}-k_n)} = \\
 & = \sum_{k_0=0}^{\infty} \sum_{k_1=0}^{k_0} \dots \sum_{k_{n-1}=0}^{k_{n-2}} \dots \binom{m}{k_0} \binom{k_0}{k_1} \dots \binom{k_{n-1}}{k_n} \dots a_0^{m-k_0} a_1^{k_0-k_1} \dots a_n^{k_{n-1}-k_n} \dots \varepsilon^{k_0+k_1+k_2+\dots+k_n+\dots} = \\
 & = \lambda_0 + \varepsilon \lambda_1 + \varepsilon^2 \lambda_2 + \varepsilon^3 \lambda_3 + \dots + \varepsilon^n \lambda_n + \dots
 \end{aligned} \tag{18}$$

The coefficient  $\lambda_n$  can be expressed as

$$\lambda_n = \sum_{k,n} \binom{m}{k_0} \binom{k_0}{k_1} \dots \binom{k_{n-2}}{k_{n-1}} a_0^{m-k_0} a_1^{k_0-k_1} \dots a_{n-1}^{k_{n-2}-k_{n-1}}, \quad \sum_{i=0}^{n-1} k_i = n, \quad k_i \geq k_{i+1} \geq 0 \quad (i=0,1,2,\dots,n-2). \tag{19}$$

The solutions (12) – (16) have been thoroughly analyzed and then the perturbation expansions (17) – (19) are used. This gave the opportunity to derive the general form of the eigenfunction  $f_k(\theta)$ . Having obtained the solutions of the linear ordinary differential equations one can find the sum of the second series expansion in (6)  $k = (n-1)/(n+1)$ :

$$f(\theta) = ((n+1)/n) \sqrt{\sqrt{(1-k^2 \sin^2 \theta) + \cos \theta}} \left[ \sqrt{(1-k^2 \sin^2 \theta) - k \cos \theta} \right]^k / (2(1-k)^k).$$

#### 4. Mode I and mixed mode loadings of the cracked specimens

The objective of this part of the paper is to study the stress singularities at the vicinity of the mixed mode (Mode I and Mode II) crack under plane stress conditions by the approach described above. The governing equations for the power law constitutive relations are transformed to eigenvalue problems of ordinary differential equations (ODEs) based on the assumption that the stress fields are asymptotic near the mixed-mode crack tip. The asymptotic and numerical methods are further developed in the present work to analyze eigenvalue problems of ODEs. Consider a stationary crack in a power-law material under plane stress conditions. Applied loading is accounted as mixed-mode I/II loading. Polar coordinates are introduced and centered at the crack tip. With reference to the polar coordinates the equilibrium equations can be written as

$$r\sigma_{rr,r} + \sigma_{r\theta,\theta} + \sigma_{rr} - \sigma_{\theta\theta} = 0, \quad r\sigma_{r\theta,r} + \sigma_{\theta\theta,\theta} + 2\sigma_{r\theta} = 0. \tag{20}$$

The compatibility condition has the following form

$$2(r\varepsilon_{r\theta,\theta})_{,r} = \varepsilon_{rr,\theta\theta} - r\varepsilon_{rr,r} + r(r\varepsilon_{\theta\theta})_{,rr} \tag{21}$$

For a material subjected to a power law hardening the constitutive equations for plane stress conditions can be written as follows

$$\varepsilon_{rr} = B\sigma_e^{n-1} (2\sigma_{rr} - \sigma_{\theta\theta}) / 2, \quad \varepsilon_{\theta\theta} = B\sigma_e^{n-1} (2\sigma_{\theta\theta} - \sigma_{rr}) / 2, \quad \varepsilon_{r\theta} = 3B\sigma_e^{n-1} \sigma_{r\theta} / 2 \tag{22}$$

where  $\sigma_e = \sqrt{\sigma_{rr}^2 + \sigma_{\theta\theta}^2 - \sigma_{rr}\sigma_{\theta\theta} + 3\sigma_{r\theta}^2}$  is the von Mises equivalent stress;  $B, n$  are the material constants. It should be noted that in the case considered the analogy between nonlinear elastic behavior and creep holds. That implies that all relations and solutions obtained for a nonlinear elastic (plastic) material with the constitutive equations (3) can be transferred to creep processes with the constitutive relations of Norton's creep law simply by replacing the strains by strain rates. The solution of Eqs. (1) – (3) should satisfy the traditional traction free boundary conditions on the crack surfaces  $\sigma_{r\theta}(r, \theta = \pm\pi) = 0$ ,  $\sigma_{\theta\theta}(r, \theta = \pm\pi) = 0$ . The mixed-mode loading can be characterized in terms of the mixity parameter  $M^p$  which is defined as

$$M^p = (2/\pi) \arctg \left| \lim_{r \rightarrow 0} \sigma_{\theta\theta}(r, \theta = 0) / \sigma_{r\theta}(r, \theta = 0) \right|. \tag{23}$$

The mixity parameter  $M^p$  equals 0 for pure mode II; 1 for pure mode I, and  $0 < M^p < 1$  for different mixities of modes I and II. Thus, for combine-mode fracture the mixity parameter  $M^p$  completely specifies the near-crack-tip fields for a given value of the hardening exponent  $n$ . By postulating the Airy stress function  $\chi(r, \theta)$  expressed in the polar coordinate system, the stress components state are expressed as:  $\sigma_{\theta\theta} = \chi_{,rr}$ ,  $\sigma_{rr} = \chi_{,r}/r - \chi_{,\theta\theta}/r^2$ ,  $\sigma_{r\theta} = -(\chi_{,\theta}/r)_{,r}$ . As for the asymptotic stress field at the crack tip  $r \rightarrow 0$ , one can postulate the following Airy stress function

$$\chi(r, \theta) = Kr^{\lambda+1} f(\theta) \tag{24}$$

where  $K$  is an indeterminate coefficient,  $\lambda$  is indeterminate exponent and  $f(\theta)$  is an indeterminate function of the polar angle, respectively. In view of the asymptotic presentation (5) the asymptotic stress field at the crack tip is derived as follows  $\sigma_{ij}(r, \theta) = Kr^{\lambda-1} \sigma_{ij}(\theta)$  or

$$\sigma_{rr}(r, \theta) = Kr^{\lambda-1} [(\lambda+1)f(\theta) + f''(\theta)], \quad \sigma_{\theta\theta}(r, \theta) = Kr^{\lambda-1} (\lambda+1)\lambda f(\theta), \quad \sigma_{r\theta}(r, \theta) = -Kr^{\lambda-1} \lambda f'(\theta) \tag{25}$$

where  $\lambda-1$  denotes the exponent representing the singularity of the stress field, and will be called the stress singularity exponent hereafter. According to (3) the asymptotic strain field as  $r \rightarrow 0$  takes the form  $\varepsilon_{ij}(r, \theta) = BK^n r^{(\lambda-1)n} \varepsilon_{ij}(\theta)$  or in the expanded form

$$\begin{aligned} \varepsilon_{rr}(r, \theta) &= BK^n r^{(\lambda-1)n} f_e^{n-1} [(\lambda+1)(2-\lambda)f(\theta) + 2f''(\theta)] / 2, \\ \varepsilon_{\theta\theta}(r, \theta) &= BK^n r^{(\lambda-1)n} f_e^{n-1} [(\lambda+1)(2\lambda-1)f(\theta) - f''(\theta)] / 2, \quad \varepsilon_{r\theta}(r, \theta) = -3BK^n r^{(\lambda-1)n} f_e^{n-1} \lambda f'(\theta) / 2. \end{aligned} \tag{26}$$

The compatibility condition (2) results in the nonlinear fourth-order ordinary differential equation for the function  $f(\theta)$ :

$$\begin{aligned} & f_e^{IV} f_e^2 \left\{ (n-1)[(\lambda+1)(2-\lambda)f + 2f'']^2 / 2 + 2f_e^2 \right\} + \\ & + 6[(\lambda-1)n+1] \lambda \left\{ (n-1)f_e^2 h f' + f_e^4 f'' \right\} + (n-1)(n-3)h^2 [(\lambda+1)(2-\lambda)f + 2f''] + \\ & + (n-1)f_e^2 [(\lambda+1)(\lambda+2)f + 2f''] \left\{ [(\lambda+1)f' + f''']^2 + [(\lambda+1)f + f''](\lambda+1)f'' + \right. \\ & + (\lambda+1)^2 \lambda^2 (f'^2 + f''') - (\lambda+1)^2 \lambda f'' / 2 - [(\lambda+1)f' + f'''](\lambda+1)\lambda f' - \\ & - \frac{1}{2} [(\lambda+1)f + f''](\lambda+1)\lambda f'' + 3\lambda^2 (f'^2 + f''') \left. \right\} + 2(n-1)f_e^2 h [(\lambda+1)(2-\lambda)f' + 2f'''] + \\ & + f_e^4 (\lambda+1)(2-\lambda)f'' - (\lambda-1)n f_e^4 [(\lambda+1)(2-\lambda)f + 2f''] + \\ & + [(\lambda-1)n+1](\lambda-1)n f_e^4 [(\lambda+1)(2\lambda-1)f - f''] = 0 \end{aligned} \tag{27}$$

where the following notations are adopted

$$\begin{aligned} f_e &= \sqrt{[(\lambda+1)f + f'']^2 + (\lambda+1)^2 \lambda^2 f'^2 - [(\lambda+1)f + f''](\lambda+1)\lambda f + 3\lambda^2 f'^2}, \\ h &= [(\lambda+1)f + f''][(\lambda+1)f' + f'''] + (\lambda+1)^2 \lambda^2 f'' - [(\lambda+1)f' + f'''](\lambda+1)\lambda f / 2 - [(\lambda+1)f + f''](\lambda+1)\lambda f' / 2 + 3\lambda^2 f f''. \end{aligned} \tag{28}$$

The boundary conditions imposed on the function  $f(\theta)$  follow from the traction free boundary conditions on the crack faces:

$$f(\theta = \pm\pi) = 0, \quad f'(\theta = \pm\pi) = 0. \tag{29}$$

One of the effective methods for the solution of nonlinear eigenvalue problems is the perturbation theory technique based on the artificially introduced small parameter [14, 16, 20 - 35]. An analytical expression for the eigenvalues of the nonlinear equation (7) can be derived by applying the perturbation theory method. For this purpose the eigenvalue  $\lambda$  is split up into  $\varepsilon = \lambda - \lambda_0$  where  $\lambda_0$  refers to the “undisturbed” linear problem and  $\varepsilon$  is the deviation on account of the nonlinearity.

Furthermore, the hardening exponent  $n$  and the stress function  $f(\theta)$  are represented as power series (6). The set of the boundary value problems for  $f_k(\theta)$  is obtained:

$$\begin{aligned} \varepsilon^0: & f_0^{IV} + 2(\lambda_0^2 + 1)f_0'' + (\lambda_0^2 - 1)^2 f_0 = 0, \\ & f_0(\theta = 0) = 1, \quad f_0'(\theta = 0) = (\lambda_0 + 1) / \operatorname{tg}(M^p \pi / 2), \quad f_0(\theta = \pi) = 0, \quad f_0'(\theta = \pi) = 0, \\ & f_0(\theta = -\pi) = 0, \quad f_0'(\theta = -\pi) = 0, \quad f_0(\theta = 0) = 1, \quad f_0'(\theta = 0) = (\lambda_0 + 1) / \operatorname{tg}(M^p \pi / 2), \\ \varepsilon^1: & f_1^{IV} + 2(\lambda_0^2 + 1)f_1'' + (\lambda_0^2 - 1)^2 f_1 = -n_1 \left[ x_0 (f_0^{IV} x_0 / 2 + w_0) / (2g_0) + h_0 (x_0' g_0 - x_0 h_0 + 3\lambda_0^2 g_0 f_0') / g_0^2 \right] - \\ & - f_0'' [(\lambda_0 - 1)(4\lambda_0 - 1)n_1 + 8\lambda_0] / 2 - f_0 (\lambda_0^2 - 1) [(\lambda_0 - 1)(4\lambda_0 + 1)n_1 + 8\lambda_0] / 2, \\ & f_1(\theta = 0) = 0, \quad f_1'(\theta = 0) = 1 / \operatorname{tg}(M^p \pi / 2), \quad f_1(\theta = \pi) = 0, \quad f_1'(\theta = \pi) = 0, \\ & f_1(\theta = -\pi) = 0, \quad f_1'(\theta = -\pi) = 0, \quad f_1(\theta = 0) = 0, \quad f_1'(\theta = 0) = 1 / \operatorname{tg}(M^p \pi / 2) \\ \varepsilon^2: & f_2^{IV} + 2(\lambda_0^2 + 1)f_2'' + (\lambda_0^2 - 1)^2 f_2 = -2g_1 \left[ f_1^{IV} + 2(\lambda_0^2 + 1)f_1'' + (\lambda_0^2 - 1)^2 f_1 \right] / g_0 + \\ & - 6\lambda_0 [2 + n_1(\lambda_0 - 1)] f_1'' - (1 - 2\lambda_0) f_1'' + (\lambda_0 - 1)(1 - 2\lambda_0) f_1 + [1 + n_1(\lambda_0 - 1)] x_1 - \\ & - \lambda_0 (\lambda_0 - 1)(4\lambda_0 + 1) f_1 - (2\lambda_0 - 1)[1 + n_1(\lambda_0 - 1)] y_1 - n_1 (1 - 2\lambda_0) f_0 (f_0^{IV} x_0 + w_0) / g_0 \\ & - 6\lambda_0 [n_1 + n_2(\lambda_0 - 1)] f_1'' - 6[1 + n_1(\lambda_0 - 1)] f_0'' + f_0'' - (\lambda_0 - 1) f_0 + \\ & + [1 + n_1(\lambda_0 - 1)](1 - 2\lambda_0) f_0 - 2\lambda_0 (\lambda_0 - 1) f_0 + (2\lambda_0 - 1)[1 + n_1(\lambda_0 - 1)](4\lambda_0 + 1) f_0 - \\ & - (2\lambda_0 - 1)[n_1 + n_2(\lambda_0 - 1)] y_0 - [1 + n_1(\lambda_0 - 1)]^2 y_0 + [n_1 + n_2(\lambda_0 - 1)] x_0 - \\ & - 2g_1 \left\{ 6_0 [2 + n_1(\lambda_0 - 1)] f_0'' + (1 - 2\lambda_0) f_0'' - (\lambda_0 - 1)(1 - 2\lambda_0) f_0 - [1 + n_1(\lambda_0 - 1)] x_0 \right\} / g_0 - \\ & + [1 + n_1(\lambda_0 - 1)](1 - 2\lambda_0) f_0 - 2\lambda_0 (\lambda_0 - 1) f_0 + (2\lambda_0 - 1)[1 + n_1(\lambda_0 - 1)](4\lambda_0 + 1) f_0 - \\ & - 2g_1 \left\{ \lambda_0 (\lambda_0 - 1)(4\lambda_0 + 1) f_0 + (2\lambda_0 - 1)[1 + n_1(\lambda_0 - 1)] y_0 \right\} / g_0 - n_1 2h_1 \left[ g_0 x_0' - h_0 x_0 + 3\lambda_0^2 g_0 f_0' \right] / g_0^2 - \\ & - n_2 \left\{ x_0 g_0 (f_0^{IV} x_0 / 2 + w_0) + 2h_0 [g_0 x_0' - h_0 x_0 + 3\lambda_0^2 g_0 f_0'] \right\} / g_0^2 - \\ & - n_1 \left[ x_0 g_0 (f_1^{IV} x_0 / 2 + w_1) + x_0 g_1 (f_0^{IV} x_0 / 2 + w_0) + g_0 x_1 (f_0^{IV} x_0 + w_0) \right] / g_0^2 + \\ & - n_1 \left\{ 2h_0 [g_0 x_1' - h_0 x_1 + 3\lambda_0^2 g_0 f_1'] + 2h_0 [g_1 x_0' - h_1 x_0 + 3\lambda_0^2 g_1 f_0'] \right\} / g_0^2 - \\ & - n_1 \left\{ 6\lambda_0 [2 + n_1(\lambda_0 - 1)] g_0 h_0 f_0' - 2h_0^2 (1 - 2\lambda_0) f_0 + n_1 h_0^2 x_0 + 2g_0 (1 - 2\lambda_0) h_0 f_0' \right\} / g_0^2 \end{aligned} \tag{31}$$

$$\begin{aligned}
 f_2(\theta=0) &= 0, \quad f_2'(\theta=0) = 1/\operatorname{tg}(M^p \pi / 2), \quad f_2(\theta=\pi) = 0, \quad f_2'(\theta=\pi) = 0, \\
 f_2(\theta=-\pi) &= 0, \quad f_2'(\theta=-\pi) = 0, \quad f_2(\theta=0) = 0, \quad f_2'(\theta=0) = 1/\operatorname{tg}(M^p \pi / 2)
 \end{aligned}
 \tag{32}$$

where the following notations are used

$$\begin{aligned}
 x_k &= (\lambda_0 + 1)(2 - \lambda_0) f_k + 2f_k'', \quad y_k = (\lambda_0 + 1)(2\lambda_0 - 1) f_k - f_k'', \quad u_k = (\lambda_0 + 1) f_k + f_k'', \\
 h_1 &= u_0 (u_1' + f_0') + u_0' (u_1 + f_0) + v_0 [v_1' + (2\lambda_0 + 1) f_0'] + v_0' [v_1 + (2\lambda_0 + 1) f_0] + 3\lambda_0 f_0' (\lambda_0 f_1'' + f_0'') - \\
 &- \frac{1}{2} u_0' [v_1 + (2\lambda_0 + 1) f_0] - \frac{1}{2} u_0 [v_1' + (2\lambda_0 + 1) f_0'] - \frac{1}{2} v_0 (u_1' + f_0') - \frac{1}{2} v_0' (u_1 + f_0) + 3\lambda_0 f_0'' (\lambda_0 f_1' + f_0'), \\
 w_1 &= 2u_0' (u_1' + f_0') + u_0 [(\lambda_0 + 1) f_1'' + f_0''] + (\lambda_0 + 1) f_0'' (u_1 + f_0) + 2v_0' [v_1' + (2\lambda_0 + 1) f_0'] + \\
 &+ (v_0' - u_0 / 2) [v_1'' + (2\lambda_0 + 1) f_0''] + v_0'' [v_1 + (2\lambda_0 + 1) f_0] - (\lambda_0 + 1) f_0'' [v_1 + (2\lambda_0 + 1) f_0] / 2 - v_0 [(\lambda_0 + 1) f_1'' + f_0''] / 2 - \\
 &- v_0' (u_1' + f_0') - u_0' [v_1' + (2\lambda_0 + 1) f_0'] - \frac{1}{2} v_0'' (u_1 + f_0) + 6\lambda_0 f_0'' (\lambda_0 f_1'' + f_0'') + 3\lambda_0 f_0''' (\lambda_0 f_1' + f_0') + 3\lambda_0 f_0' (\lambda_0 f_1''' + f_0''').
 \end{aligned}$$

The solution of the fourth-order linear ordinary differential equation (30) with respect to function  $f_0(\theta)$  satisfying the traction-free boundary conditions has the form: for the crack opening mode I (for symmetric stress fields, pure mode I)  $f_0^I = \beta \cos \alpha \theta - \alpha \cos \beta \theta$ ,  $\alpha = \lambda_0 - 1$ ,  $\beta = \lambda_0 + 1$ , for the shear crack mode II (the skew-symmetric stress fields, pure mode II)  $f_0^{II} = \sin \alpha \theta - \sin \beta \theta$ , where the spectrum of the eigenvalues is determined by the characteristic equation  $\sin 2\pi \lambda_0 = 0$ , whence one can easily find  $\lambda_0 = m/2$ , where  $m$  is an integer. Thus it is shown that an infinite number of eigenvalues exists. In view of the linearity of Eq. (30) for the mixed-mode crack problem the solution is the superposition of the symmetric and antisymmetric parts of the stress field with respect to the crack plane

$$f_0(\theta) = C_1 (\beta \cos \alpha \theta - \alpha \cos \beta \theta) + C_2 (\sin \alpha \theta - \sin \beta \theta) \tag{33}$$

where  $C_1$  and  $C_2$  are unknown coefficients which have to be determined from the boundary conditions of the actual crack problem and represent the modes I and II, respectively. In view of (23) the unknown constants  $C_1$  and  $C_2$  are related to the mixity parameter  $M^p = 2 \operatorname{arctg}[(\lambda_0 + 1)C_1 / C_2] / \pi$ . The zeroth-order problem (30) has the nontrivial solution (33), hence the inhomogeneous problems for the functions  $f_1(\theta)$  and  $f_2(\theta)$  (13) will not have solutions unless a solvability condition is satisfied [16, 35]. Therefore, if  $\lambda_0$  is not an eigenvalue of the homogeneous problem (i.e. the homogeneous problem has only the trivial solution), the inhomogeneous problem has a unique solution for every continuous right hand side  $G_k(\theta)$  of the differential equation for  $f_k(\theta)$ ,  $k > 0$ . On the other hand, if  $\lambda_0$  is an eigenvalue of the homogeneous problem (i.e. the homogeneous problem has a nontrivial solution), the inhomogeneous problem does not have a solution unless [16, 27, 35]. Following the procedure described in [16, 27, 35] one can find that the compatibility condition has the form

$$\int_{-\pi}^{\pi} G_k(\theta) u(\theta) d\theta = 0. \tag{34}$$

That is,  $G_k(\theta)$  is orthogonal to the eigenfunction  $u(\theta)$ , corresponding to the eigenvalue  $\lambda_0$ . These results constitute the so-called Fredholm's theorem: for a given value  $\lambda_0$ , either the inhomogeneous problem has a unique solution for each continuous right hand side of the equation, or else the homogeneous problem has a nontrivial solution [16, 35]. To determine the solvability condition (34) we use the concept of adjoint problems [16-35]. The boundary value problem (31) is self-adjoint since the differential equation and the boundary conditions of the adjoint problem coincide with the differential equation and boundary conditions of the homogeneous problem (30). Therefore,  $u(\theta) = f_0(\theta)$ , where the function  $f_0(\theta)$  is determined by Eq. (33). According to Eq. (34) the solvability condition of the boundary value problem (34) has in the expanded form

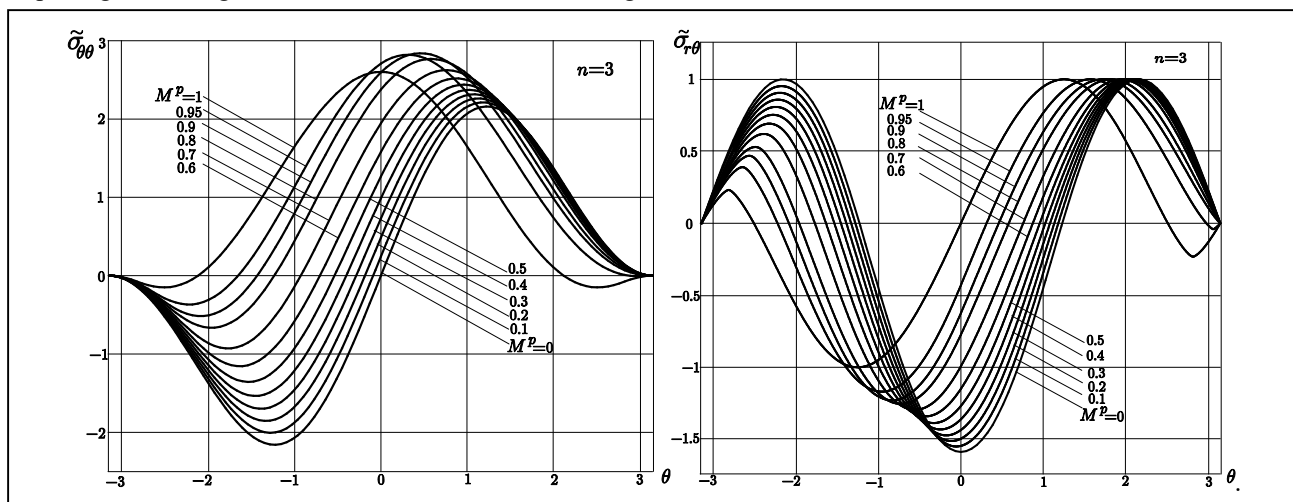
$$\begin{aligned}
 &\int_{-\pi}^{\pi} \left\{ -n_1 \left[ x_0 (f_0^{IV} x_0 / 2 + w_0) / (2g_0) + h_0 (x_0' g_0 - x_0 h_0 + 3\lambda_0^2 g_0 f_0') / g_0^2 \right] - \right. \\
 &\left. - \frac{1}{2} f_0'' [(\lambda_0 - 1)(4\lambda_0 - 1)n_1 + 8\lambda_0] - \frac{1}{2} f_0' (\lambda_0^2 - 1) [(\lambda_0 - 1)(4\lambda_0 + 1)n_1 + 8\lambda_0] \right\} f_0(\theta) d\theta = 0.
 \end{aligned}$$

The compatibility condition of the boundary value problem for the function  $f_1(\theta)$  allows us to find the coefficient  $n_1$ . Having obtained the function  $f_1(\theta)$ , one can determine the unknown function  $f_2(\theta)$ . Using the analogous reasoning, one can formulate the compatibility condition for the solution of the boundary value problem for  $f_2(\theta)$  and calculate numerically the values of the following coefficient of the asymptotic expansion  $n_2$  of the hardening exponent  $n$  for different values of the mixity parameter.

## 5. Results and Discussion

The perturbation theory method allowed us to find the closed form solution for mode III crack problem. The problem has been reduced to the nonlinear eigenvalue problem and the analytical presentational of the eigenfunction has been obtained by the small parameter method. The asymptotic analysis based on the artificial small parameter method of the perturbation theory provided a possibility to reveal the new stress singularity in the vicinity of the mixed mode crack tip. In the paper the technique

for numerical determination of the eigenvalues of the nonlinear eigenvalue problem is proposed. Numerical approach allows us to find the eigenfunctions immediately and the results of calculations are shown in Fig. 1. Using this technique the new eigenvalues resulting in the continuous radial stress components at  $\theta = 0$  are found. It is shown that the method proposed gives the eigenvalues corresponding to the HRR problem in particular cases of mode I and mode II crack problems. The theoretical significance of the present paper is that from the method described here one can clearly know all the mathematically possible distributions of stress singularities at the crack tip under mixed-mode loading. It should be noted that it is important to develop asymptotic analysis methods and their applications for nonlinear eigenvalue problems in solid mechanics [27-37] and, in particular, in nonlinear fracture mechanics and continuum damage mechanics [37] for enunciating newer and better approaches for imparting knowledge on reliable determination of fatigue and fracture behavior. In nonlinear fracture mechanics the



eigenfunction expansion method is one of the most commonly encountered approaches [25-37]. The method leads to nonlinear eigenvalue problems which stipulate the possible distributions of stress singularity at the crack tip and the determination of the whole eigenspectrum requires invoking developed asymptotic and computational techniques and their combinations.

Fig. 1. Eigenfunctions: solution of the nonlinear eigenvalue problem for near crack-tip stress field under mixed-mode loading.

## 6. Conclusion

Using an asymptotic expansion and separation of variables for the stress function a series solution for all hardening exponents is obtained. In the present work the closed form solution for the eigenfunctions for the crack tip fields under antiplane shear is obtained. It is shown that the perturbation method allows us to derive the analytical solution for the eigenfunctions. The approach developed here and the closed-form solution obtained can be used for Mode I, Mode II and mixed mode crack problems for determining the eigenfunctions. It should be noted either that the class of nonlinear eigenvalue problems arising in nonlinear fracture mechanics is essential in connection with creating the multiscale models of fracture with multi-singularities with different orders at the crack point. The singularity representation scheme has to be considered where the local damage at the different scales will be modeled by different orders of the stress singularities. Different stress singularities can be related to different loading type and severity of material damage. In accordance to these models it is necessary to introduce the hierarchy of the zones in the vicinity of the crack tip with dominating role of different stress asymptotic behavior and to realize the matching procedures between different stress asymptotic solutions. The accurate construction of all the intermediate zones with one or other stress asymptotics requires the knowledge of the whole spectrum of eigenvalues and these problems are still open.

## References

- [1] Hutchinson JW. Singular behaviour at the end of a tensile crack in a hardening materials. *J. Mech. Phys. Solids*. 1968; 16: 13–31.
- [2] Hutchinson JW. Plastic stress and strain fields at a crack tip. *J. Mech. Phys. Solids*. 1968; 16: 337–347.
- [3] Rice JR, Rosengren GF. Plane strain deformation near a crack tip in a power-law hardening material. *J. Mech. Phys. Solids*. 1968; 16: 1–2.
- [4] Anheuser M, Gross D. Higher order fields at crack and notch tips in power-law materials under longitudinal shear. *Archive of Applied Mechanics* 1994; 64: 508–518.
- [5] Neuber H. Theory of stress concentration for shear-strained prismatical bodies with arbitrary nonlinear stress-strain law. *Trans. ASME/E. Journal of Applied Mechanics* 1961; 28: 544–550.
- [6] Rice JR. Contained plastic deformation near cracks and notches under longitudinal shear. *International Journal of Fracture Mechanics* 1966; 2: 426–447.
- [7] Rice JR. Stresses due to a sharp notch in a work-hardening elastic-plastic material loaded by longitudinal shear. *Trans. ASME/E. Journal of Applied Mechanics* 1967; 34: 287–298.
- [8] Yang S, Yuan FG, Cai X. Higher order asymptotic elastic-plastic crack-tip fields under antiplane shear. *Engineering Fracture Mechanics* 1996; 54(3): 405–422.
- [9] Stolz C. Asymptotic fields ahead a crack for a class of non linear materials under mode III. *Mechanics of Materials* 2015; 90: 102–110.
- [10] Bui HD, Ehrlicher A. Propagation dynamique d'une zone endommagée dans un solide élastique fragile en mode III et en régime permanent. *C.R. Acad. Sci. Paris, Ser. B* 1980; 290: 273–276.
- [11] Abeyaratne R. Discontinuous deformation gradients away from the tip of a crack in anti-plane shear. *J. Elast.* 1981; 11: 373–393.
- [12] Abdelmoula R, Debruyne G. Analysis of the stress and strain fields near the crack tip of a steady-state growing crack in an elastic-viscous medium: The Hui-Riedel problem revisited by means of method of matched asymptotic expansions. *Comptes Rendus Mécanique* 2016; 344: 613–622.
- [13] Hui CY, Riedel H. The asymptotic stress and strain field near the tip of a growing crack under creep conditions. *Int. J. Fract.* 1981; 17: 409–425.
- [14] Liao S. Series solution of nonlinear eigenvalue problems by means of the homotopy analysis method. *Nonlinear Analysis: Real World Applications* 2009; 10: 2455–2470.
- [15] Cole JD. *Perturbation Methods in Applied Mathematics*. Blaisdell Publishing Company, Waltham, Massachusetts, 1968.



- [16] Nayfeh AH. Introduction to Perturbation Techniques. John Wiley & Sons, New York, 1981.
- [17] Murdock A. Perturbations: Theory and Methods. John Wiley & Sons, New York, 1991.
- [18] Bush AW. Perturbation Methods for Engineers and Scientists. CRC Press Library of Engineering Mathematics, CRC Press, Boca Raton, 1992.
- [19] Kevorkian J, Cole JD. Multiple Scales and Singular Perturbation Methods. Applied Mathematical Sciences 1995; 114.
- [20] Nayfeh AH. Perturbation Methods. John Wiley & Sons, New York, 2000.
- [21] Adomian G. Nonlinear stochastic differential equations. J. Math. Anal. Appl. 1976; 55: 441–452.
- [22] Rach R. On the Adomian method and comparisons with Picard's method. J. Math. Anal. Appl. 1984; 10: 139–159.
- [23] Adomian G, Rach R. On the solution of algebraic equations by the decomposition method. Math. Anal. Appl. 1985; 105(1): 141–166.
- [24] Adomian G. A review of the decomposition method and some recent results for nonlinear equations. Comput. Math. Appl. 1991; 21: 101–127.
- [25] Lyapunov AM. General Problem on Stability of Motion. Taylor & Francis, London, 1992.
- [26] Ehrlacher A, Markenscoff X. Duality, Symmetry and Symmetry Lost in Solid Mechanics. Press des Ponts, Paris, 2011.
- [27] Stepanova LV, Igonin SA. Asymptotics of the near-crack-tip stress field of a growing fatigue crack in damaged materials: Numerical experiment and analytical solution. Numerical Analysis and Applications 2015; 8(2): 168–181.
- [28] Stepanova LV, Adylina EM. Stress-strain state in the vicinity of a crack tip under mixed loading. Journal of Applied Mechanics and Technical Physics 2014; 55(5): 885–895.
- [29] Bui HD. Fracture Mechanics: Inverse problems and Solutions. Dordrecht: Springer, 2006.
- [31] Stepanova LV, Yakovleva EM. Mixed-mode loading of the cracked plate under plane stress conditions. PNRPU Mechanics Bulletin 2014; 3: 129–162.
- [32] Paulsen W. Asymptotic Analysis and Perturbation Theory. Boca Raton, London, New York: CRC Press, 2014.
- [33] Stepanova LV. Eigenvalues of the antiplane-shear crack problem for a power-law material. Journal of Applied Mechanics and Technical Physics 2008; 49(1): 142–147.
- [34] Stepanova LV. Eigenspectra and orders of stress singularity at a mode I crack tip for a power-law medium. Comptes Rendus – Mecanique 2008; 336(1-2): 232–237.
- [35] Stepanova L, Yakovleva E. Stress-strain state near the crack tip under mixed-mode loading: Asymptotic approach and numerical solutions of nonlinear eigenvalue problems. AIP Conference Proceedings 2016; 1785: 030030.
- [36] Qiu Z, Zheng Y. Predicting fatigue crack growth evolution via perturbation series expansion method based on the generalized multinomial theorem. Theoretical and Applied Fracture Mechanics 2016; 86: 361–369.
- [37] Stepanova LV, Fedina MYe. Self-similar solution of a tensile crack problem in a coupled formulation. Journal of Applied Mathematics and Mechanics 2008; 72(3): 360–368.

# Study of the chain transfer agent's effect on the butadiene-styrene copolymer's properties based on the Monte-Carlo method

T. Mikhailova<sup>1</sup>, E. Miftakhov<sup>2</sup>, S. Mustafina<sup>1</sup>

<sup>1</sup>Bashkir State University, 32, Validy Str., 450076, Ufa, Russia

<sup>2</sup>Ufa State Aviation Technical University, 12, K. Marx Str., 450008, Ufa, Russia

---

## Abstract

The microstructural and molecular characteristics of the butadiene-styrene copolymer are investigated depending on the feeding mode of the chain transfer agent in the paper. The study is based on mathematical simulation of the butadiene-styrene copolymerization process by the Monte Carlo method, where the tert-dodecyl mercaptan is used as the chain transfer agent. The dependences of the values of the weight-average molecular weight, molecular weight distribution and microheterogeneity index on the serial index of the reactor in the cascade are obtained.

*Keywords:* copolymerization; simulation; Monte Carlo method; butadiene; styrene; tert-dodecyl mercaptan

---

## 1. Introduction

One of the leading branches of modern petrochemistry is the synthetic rubber industry. Butadiene-styrene rubbers obtained by the method of cold emulsion-type polymerization at the temperature of 5-8 °C are the most common and popular types of synthetic rubbers. These rubbers are widely used in the production of rubber products, but the main area of their usage is the tire industry. That is due to the high technical properties of tires obtained on their basis, as well as to the availability of monomers.

## 2. The object of the study

The production of butadiene-styrene synthetic rubber is carried out in the cascade of connected polymerizers, each of which is a continuous stirred tank reactor. This process is carried out in the continuous mode with a feed of new reagents to the reactor and an extraction of reaction products from it, which provides the continuation of the reactions. But due to the peculiarities of the copolymerization process, the product obtained is heterogeneous in composition and microstructure of the macromolecules.

One of the ways to correct the parameters of the butadiene-styrene copolymer is the effect of the chain transfer agent on the macromolecules growth of the product. The chain transfer agent is continuously added to the reaction mixture at several points of the mechanism of process [1, 2]. But the optimum quantity of the chain transfer agent's feed and the selection of the feeding mode in the cascade of reactors can only be established by the experiment. In connection with this, it is relevant to study the characteristics of the butadiene-styrene copolymerization product depending on the mode of production and the composition of the reaction mixture.

## 3. Methods

In modern conditions, methods of mathematical modeling are used to study technological processes within the framework of industrial production. The obtained mathematical model allows to predict the physicochemical parameters of the product yielded under exploitation's conditions for a given kinetic scheme and parameters of process.

The kinetic scheme of butadiene-styrene copolymerization was described in paper [3], and an algorithm for simulation the synthesis of butadiene-styrene copolymer in the cascade of reactors by continuous mode has been described in papers [4, 5, 6]. The algorithm of simulation of processes is based on the Monte Carlo method. Since the process under study is continuous, the residence time distribution of the product's macromolecules is taken into account during the simulation.

The choice of this approach was conditioned by the fact that the basis of simulation is an imitation of the formation of copolymer's macromolecules, which allows to store the information about the composition and length of the chains being formed in the dynamics of the synthesis. This, in its turn, allows to determine the values of the product's characteristics at any time during the simulation.

A software was developed to simulate the synthesis of butadiene-styrene copolymer which is carried out in the cascade of stirred tank reactors on continuous mode on the basis of the created model. For the design Visual Studio programming environment was used with C # and Visual C ++ languages [7, 8].

## 4. Results and Discussion

The developed software was used to investigate the effect of the chain transfer agent's feed mode on the characteristics and microstructure of the obtained product. For this purpose, series of computational experiments were performed under the following conditions:

- the load on the cascade by monomers: 3.5 t/h (100 w.p., butadiene – 70 w.p., styrene – 30 w.p.),

- dosage of initiator (pinane hydroperoxide): 0.054 w.p.,
- ratio water / monomers – 220:100,
- working volume of polymerizer – 10.8 m<sup>3</sup>,
- volumetric flow rate – 9.5982 m<sup>3</sup>/h,
- residence time of the reaction mixture in polymerizer – 1.125 h.

In this case, we will use the following chain transfer agent's feed mode: 3 points (1<sup>st</sup> reactor – 0.125 w.p., 3<sup>rd</sup> reactor – 0.027 w.p., 6<sup>th</sup> reactor – 0.027 w.p.) and 2 points of the cascade (1<sup>st</sup> reactor – 0.125 w.p., 6<sup>th</sup> reactor – 0.027 w.p.).

Fig. 1 depicts the dependence of the weight-average molecular weight of the butadiene-styrene copolymer on the index of the reactor in the cascade. The additional feed of the chain transfer agent in the third point of the cascade promotes to slow growth of the values of the product's weight average molecular weight. The molecular weight distribution (MWD) is characterized by an increase in the low molecular weight's fractions and a decrease in the high molecular weight's fractions of the formed copolymer (Fig. 2).

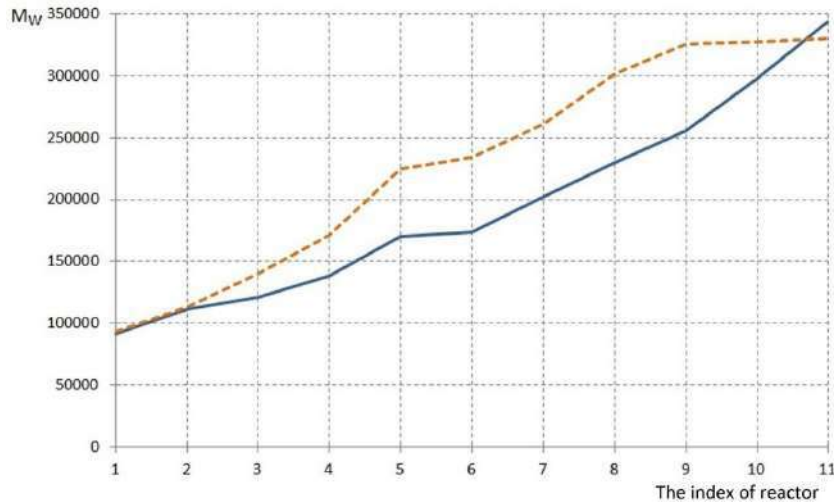


Fig. 1. Changing the weight-average molecular weight of the formed copolymer depending on the index of reactor in the cascade: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

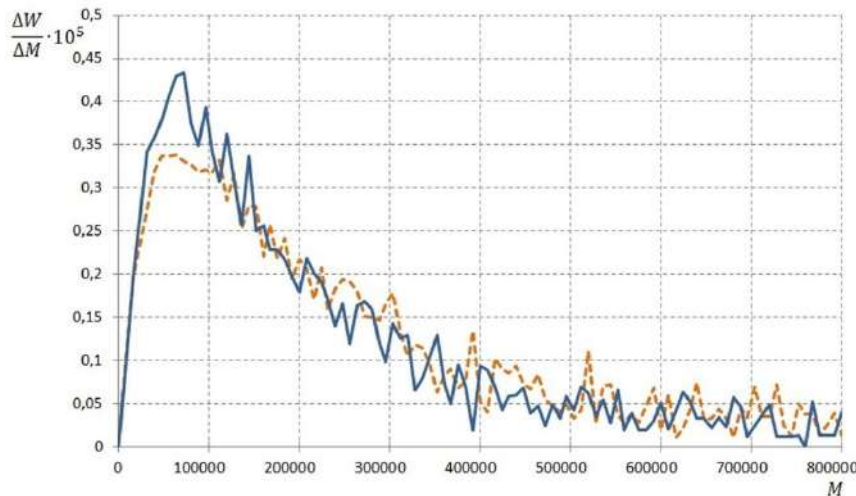


Fig. 2. Differential curve of the molecular weight distribution of the styrene-butadiene copolymer: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

The constructed model allows to investigate the sequence of the combination of monomeric units in the formed copolymer's chains. At the same time, the microstructure of macromolecules is usually characterized not by fractions of different sequences of units, but by parameters that represent some of their combinations. This parameter is the microheterogeneity index for the binary copolymer.

If the chain of the binary copolymer can be represented as a sequence of dyads of butadiene-butadiene (BB), butadiene-styrene (BS), styrene-butadiene (SB), styrene-styrene (SS), whose fractions are denoted  $P_{BB}$ ,  $P_{BS}$ ,  $P_{SB}$ ,  $P_{SS}$ , then the microheterogeneity index can be calculated from the following formula:

$$K_M = \frac{P_{BS}}{P_B P_S}, \quad (1)$$

where the fractions of butadiene and styrene in the chains are calculated according to the formulas:

$$\begin{aligned} P_B &= P_{BB} + P_{BS}, \\ P_S &= P_{SS} + P_{SB}. \end{aligned} \quad (2)$$

In Fig. 3 the dotted line shows the dependence of the microheterogeneity index of the copolymer on the index of the reactor in the cascade. Values of the microheterogeneity index vary from 0.98 in the first reactor to 0.87 in the last reactor of the cascade. It characterizes the final product as a statistical copolymer with a tendency to form long blocks. The additional feed of the chain transfer agent to the third point of the cascade helps to narrow the range of variation of the microheterogeneity index and decrease the probability of formation of long blocks: the microheterogeneity index varies from 0.98 in the first reactor to 0.94 in the last reactor of the cascade.

It can be noted that the range of the change in the fraction of butadiene-butadiene homodyads from 0.78 to 0.38 corresponds to the two-point feed mode of the chain transfer agent to versus the range of the change from 0.78 to 0.52 for the three-point feed mode. The fraction of styrene-styrene homodyads varies from 0.02 to 0.2 at the two-point feed mode of the chain transfer agent versus the change from 0.02 to 0.09 at the three-point feed mode. The range of the change in the fraction of butadiene-styrene heterodyads at different feed mode of the chain transfer agent varies insignificantly: from 0.2 to 0.42 in the two-point feed mode and from 0.2 to 0.39 in the three-point feed mode. A significant change in the fraction of dyads in the last reactors of the cascade is associated with the total consumption of the chain transfer agent (Fig. 4-5).

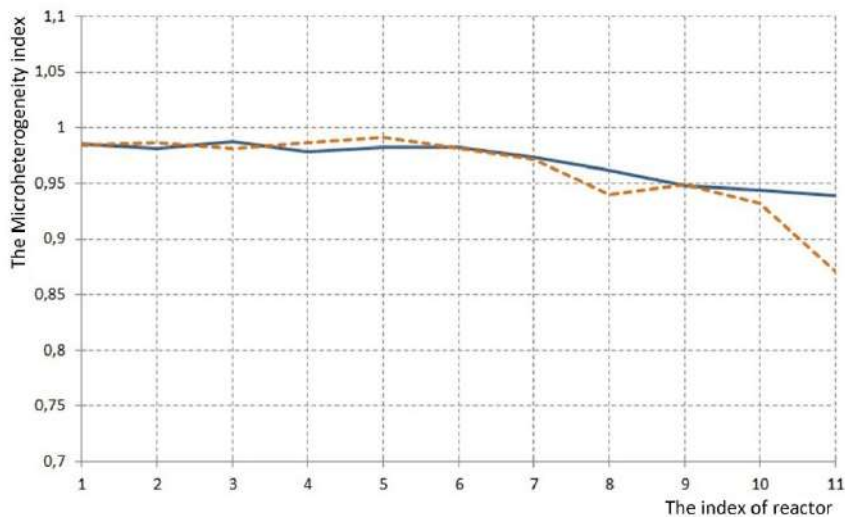


Fig. 3. Changing the microheterogeneity index of the formed copolymer depending on the index of reactor in the cascade: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

## 5. Conclusion

Simulation of the synthesis of the butadiene-styrene copolymer makes it possible to study the characteristics of the obtained product on the basis of the Monte Carlo method. Since simulation is based on imitating the growth of copolymer's macromolecules and tracking the given processes, it contributes to the accumulation of information on the composition and length of the formed chains in the dynamics of synthesis. This makes it possible to predict and analyze the microstructure of the product. It is established that the fractional feed mode of the chain transfer agent supply leads to the narrowing of the range of the copolymer's microheterogeneity index. The high molecular weight fractions of the copolymer increase during the course of the process, which results in the rigidity of the product obtained on the basis of the copolymer. At the same time, the increase in the content of styrene homodyads in macromolecules contributes to the decrease in the elasticity of the product.

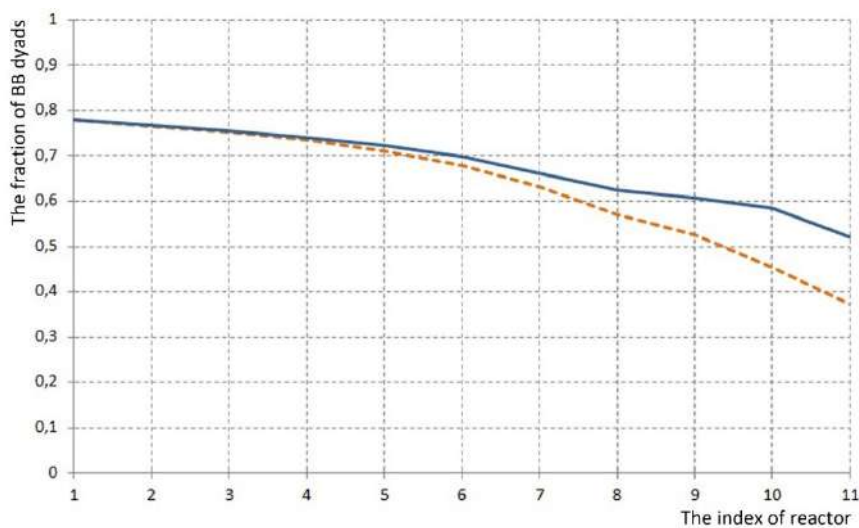


Fig. 4. Changing the values of the butadiene-butadiene homodyads fraction in copolymer chains depending on the index of reactor in the cascade: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

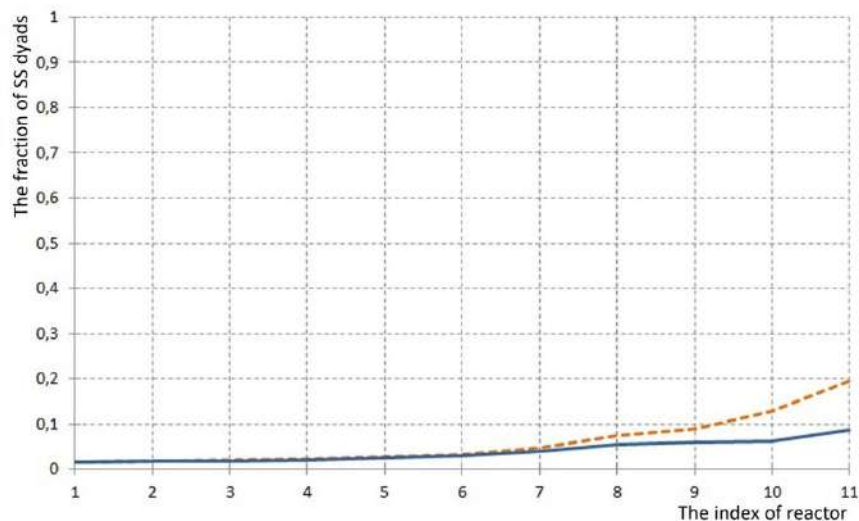


Fig. 5. Changing the values of the styrene-styrene homodyads fraction in copolymer chains depending on the index of reactor in the cascade: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

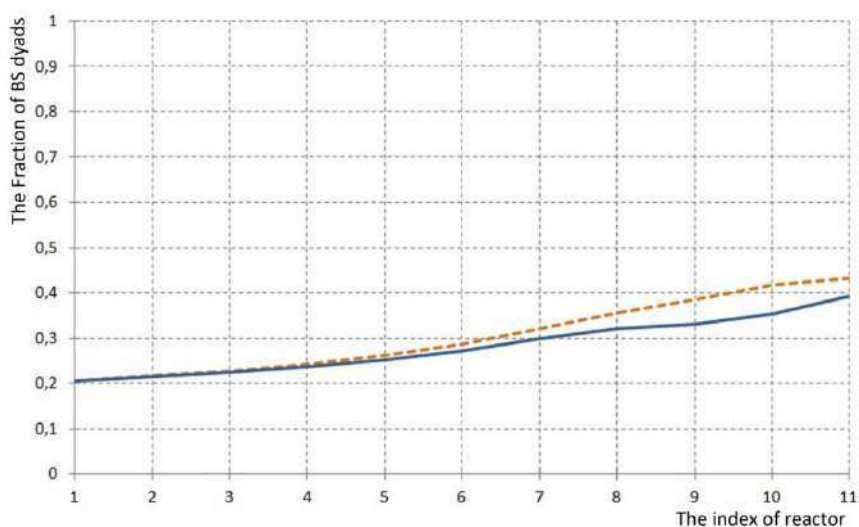


Fig. 6. Changing the values of the butadiene-styrene heterodyads fraction in copolymer chains depending on the index of reactor in the cascade: the dotted line – two-point feed mode, the solid line – three-point feed mode of chain transfer agent.

## Acknowledgements

The study was funded by RFBR according to the research projects №16-31-00162 and №17-47-020068 and project No.13.5143.2017/ BCH, carried out by the University in the framework of the State Task of the Ministry of Education of the Russian Federation.

## References

- [1] Kirpichnikov PA, Beresnev VV, Popova LM. Album of technological schemes of the main industries of the synthetic rubber. Leningrad: Chemistry, 1986; 224 p. (in Russian)
- [2] Averko-Antonovich LA, Averko-Antonovich YuO, Davletbaeva IM, Kirpichnikov PA. Chemistry and technology of synthetic rubber. Moscow: Chemistry, 2008; 357 p. (in Russian)
- [3] Mustafina S, Miftakhov E, Mikhailova T. Solving the direct problem of butadiene-styrene copolymerization. International Journal of Chemical Sciences 2014; 12(2): 564–572.
- [4] Mustafina S, Mikhailova T, Miftakhov E. Mathematical Study of the butadiene-styrene copolymerization product by the Monte-Carlo method. International Journal of Chemical Sciences 2015; 13(2): 849–856.
- [5] Mikhailova T, Miftakhov E, Mustafina S. Mathematical Simulation of the Styrene-Butadiene Rubber's Production in the Cascade of Reactors by the Monte-Carlo Method. International Journal of Chemical Sciences 2016; 14(4): 1865–1876.
- [6] Mikhailova T, Mustafina S, Grigoryev I. Numerical study of copolymer composition and compositional heterogeneity during the synthesis of butadiene-styrene rubber. International Journal of ChemTech Research 2017; 10(2): 1031–1036.
- [7] Mikhailova T. Computer modeling of styrene-butadiene rubber's production in the cascade of reactors by the Monte Carlo method. Control Systems and Information Technologies 2016; 4(66): 64–69.
- [8] Mikhailova TA, Miftakhov EN, Mustafina SA. Computer program «CopolMMKforCascade» for simulation of the batch and continuous processes of free-radical butadiene-styrene copolymerization. Federal Service for Intellectual Property (Rospatent). № 2016662302, 07.11.2016.

# Reconstruction of realistic three-dimensional models of biological objects from MR-images for the radiation therapy purposes

A.V. Lebedeva<sup>1</sup>, V.V. Mamontova<sup>1</sup>, S.A. Nemnyugin<sup>1</sup>, A.V. Komolkin<sup>1</sup>

<sup>1</sup>*Saint Petersburg State University, 7/9 Universitetskaya emb., 199034, St. Petersburg, Russia*

---

## Abstract

Magnetic Resonance Imaging (MRI) is one of the most widely used medical diagnostic techniques. Digital Imaging and COmmunications in Medicine (DICOM) is standard format to store results of MRI. In the paper methods of visualization of three-dimensional voxel models of human organs from data obtained using MRI are considered. These models may be used both in medical research and for planning of the radiation therapy treatment. The result of the work is a software package developed for medical physics research.

*Keywords:* Magnetic Resonance Imaging; DICOM; biological system modeling; voxel volume model; 3D rendering

---

## 1. Introduction

Efficient methods of reconstruction of realistic three-dimensional models of biological objects from medical images may be used both to improve quality of human's life and for medical purposes. For example, they may be used for creating of tool detecting cancer which is one of the leading reasons of mortality [1]. Cancer should be treated on as early stages as possible, so its early diagnostics is extremely important. One of the most widely used techniques of diagnostics is magnetic resonance imaging (MRI). MRI allows imaging in three mutually perpendicular planes. A qualified specialist should have the ability to detect abnormalities in the structure of the body without surgery by viewing the individual images. Reconstructed realistic volume model with possibility of visual transformations makes analysis more efficient. Reconstructed from real tomograms 3D models may be also used for the purposes of computer simulation of processes of radio- and hadron therapy [2-3]. Usage of such models allows getting more reliable results taking into account personal features of the patient, so it should help to develop more rigorous treatment plans. In addition, technologies of augmented reality allow associate preoperative data with the current state of the organism, or to use them in real-time in the operations.

MRI is based on the phenomenon of nuclear magnetic resonance. The patient is placed in a scanner which creates cross-sectional images of a human body or other biological object. MRI image should be analyzed and interpreted by physician. Tomographic survey results are stored in the file according to the medical industry standard DICOM 3.0 file [4]. This standard uses its own internal storage technology, so there is a need of efficient conversion of DICOM images to the volume geometrical model which may be used in diagnostics and simulation.

## 2. Reconstruction of volume models from DICOM files

DICOM format (Digital Imaging and Communications in Medicine) is most universal standard in digital medical imaging. It should be supported by software tools for medical diagnostics and simulation.

Metadata of patients and medical information are stored in DICOM format as objects with assigned attributes. The hierarchy of DICOM files is presented on Fig.1.

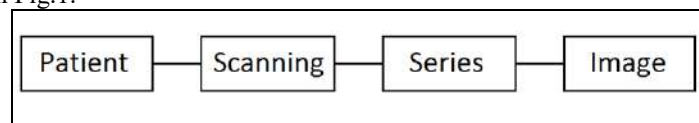


Fig. 1. The hierarchy of DICOM files.

DICOM objects and attributes have to be defined according to the DICOM Information Object Definitions (IOD). Patient IOD, for example, generally is described by name, medical record number (ID), sex, age, weight, etc. - any clinically relevant patient information. Formally the patient is a set of his attributes. DICOM includes creation of a list of standard attributes. The list is the main part of the DICOM Data Dictionary. DICOM dictionary allows to guarantee consistency of naming and handling attributes [5]. Numerical data are stored in binary format.

The primary unit of DICOM is a Data Element (DM). DM includes four mandatory and one optional elements: Group Number, Element Number, Value Length, Value, Value Representation (optional).

DICOM data also include tags. For example, if DICOM needs to search for either CT or MR studies it will search by the pattern string "CT \ MR". Information IODs register in the tags.

For decoding of DICOM files and extracting both images and metadata there are specialized DICOM viewers. Some of the DICOM viewers provide opportunity to get only two-dimensional images. Such simplified form of visualization is sufficient to formulate diagnosis. More complete and rigorous method of visualization is reconstruction of volume model on the basis of combination of 2D images corresponding to different sections. It may be highly desirable to get not only the visual appearance of the object, but also associate with geometrical volume model its physical characteristics, such as tissue density, chemical composition and so on.



Distance between successive sections in MRI images in general is greater than size of a two-dimensional pixel in the plane. Volume model is composed from voxels which are 3D generalization of plane pixels. Thus, the geometrical dimensions of the voxels in a volume model may be different in all three dimensions. In this case, the data element consists of voxels having a base corresponding to the size of the pixel which belongs to a plane and a height corresponding to the distance between section images.

### 3. Method of reconstruction of volume model

Method of volume model reconstruction consists of the following steps: processing of the DICOM-file (extracting of metadata, extracting of 2D images and patient IOD), volume model reconstruction, 3D image rendering.

#### 3.1. 2D Image Processing

MRI data of each image section have to be converted to image in graphic png format. Color of every pixel is defined by the density of body tissue in the tomogram. In case of MRI DICOM file stores signal intensity. It is necessary to take into account additional information in tags "window width" and "window center" to get an array of densities. In our work algorithm proposed in [6] is used with simple transformation function. Results are presented on Figs.2-3.

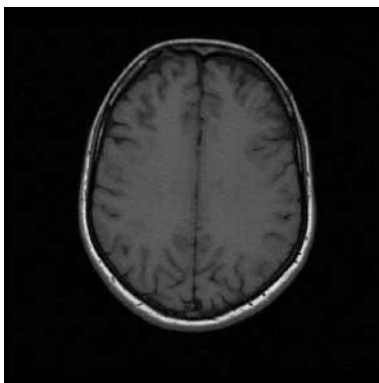


Fig. 2. Image of the brain.

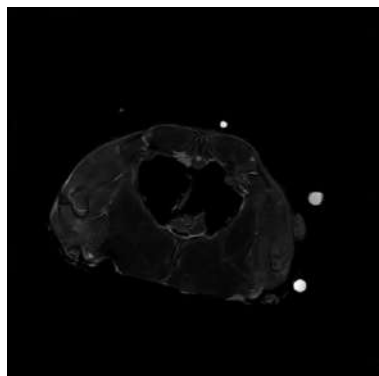


Fig. 3. Image of chicken carcass with MRI markers.

#### 3.2. 3D Image Processing

3D reconstruction implemented by third-party packages has a number of shortcomings, including the specialized formats of output files and lack of information about algorithms of volume model reconstruction, so we realized 3D model reconstruction method in our own software package.

From a set of different methods of 3D reconstruction the most common one - *voxel-based volume model* was used. In the model a voxel is not only a volumetric pixel, but it also contains a color value, generally corresponding to density of biological tissue. Reconstruction of 3D model is based on combining of section images. Distance between the images is defined by orientation of each image in space, as well as its spatial coordinates and thickness of each layer. Serious shortcoming of the voxel-based volume model is in large resulting data files. Processing of such files requires a lot of computer memory and CPU times. To reduce data to be processed in rendering it is necessary to show only those voxels that are not hid by others.

3D image processing is implemented with OpenGL Shading Language. It is high-level shading language with a syntax based on the C programming language. GLSL shaders represent a set of strings that are passed to the hardware vendor's driver for compilation from within an application using the OpenGL API's entry points.

### 4. Results and Discussion

Algorithm of the volume model reconstruction from DICOM file was implemented with options of visual transformations of rendered image. Reconstructed models represent not only 3D geometry of biological object from MRI tomogram but they also

store information on the density of tissues. Three-dimensional model of chicken carcass is presented on fig. 4. Other example is the image of the brain given on fig.5.

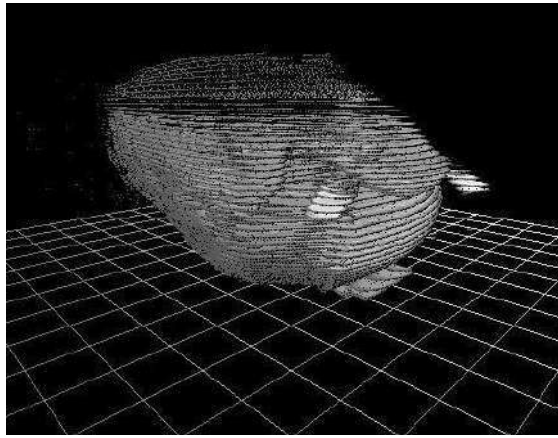


Fig. 4. 3D rendering of of chicken volum mode based on MRI.

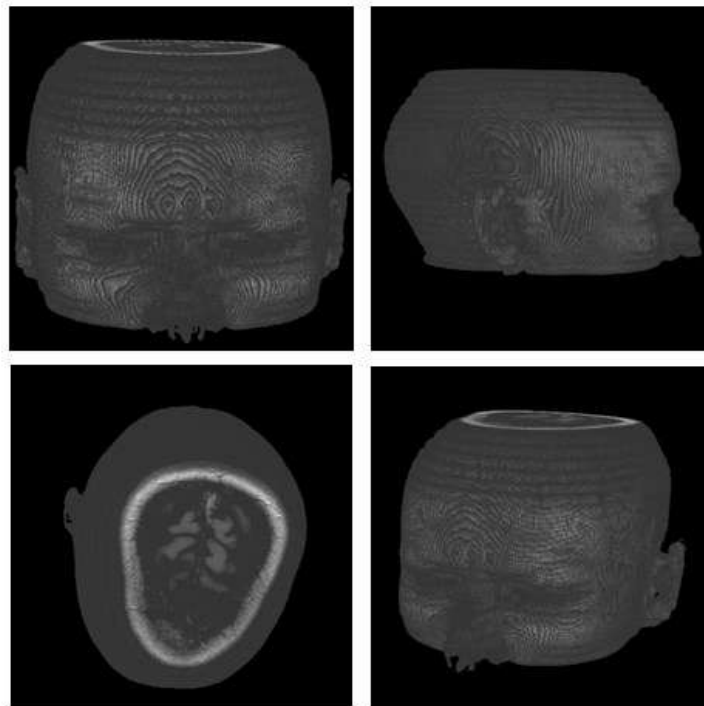


Fig. 5. 3D rendering of reconstructed volume model of the head.

## 5. Conclusion

Algorithms of the volume model reconstruction from MRI images in DICOM files were developed and implemented as software program. The program allows reading DICOM file and visualizing its content as two-dimensional image. Volume model could also be reconstructed. It may be rendered to present MRI results more completely. 3D model may be also used for simulation of interaction of the radiotherapeutic beam with tissues of human body or other biological object. It is necessary to optimize the operation of the program, for example, using HPC and parallel programming techniques. The program can be used both for medical purposes and for studies in medical physics.

## Acknowledgements

The authors acknowledge the Physics Department of the St. Petersburg State University. Research was carried out using computational resources provided by the Resource Physics Educational Centre of the Research park of Saint-Petersburg State University.

## References

- [1] Office for National Statistics. URL: <https://www.ons.gov.uk/> (25.05.2017).
- [2] Schardt D, Elsasser T. Heavy-ion tumor therapy: Physical and radiobiological benefits. *Rev. Mod. Phys.* 2010; 82: 383–425.
- [3] Kalatusha OA, Ruban OV, Nemnyugin SA. Computer Simulation Of Radiation Dose Absorption in Biological Specimens. *Math. Mod. Geom.* 2016; 4: 1: 41–50.
- [4] The DICOM Standard. URL: <http://dicom.nema.org/standard.html> (25.05.2017).
- [5] Pianykh OS. *Digital Imaging and Communications in Medicine (DICOM)*. Springer-Verlag, First edition, 2008; 383 p.
- [6] C.11.2.1.2 Window center and window width. URL: <https://www.dabsoft.ch/dicom/3/C.11.2.1.2/> (25.05.2017).



# Neural network prediction model of the pilots' errors

A.N. Danilenko<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

This paper introduces a hybrid model of the neuro-fuzzy classifier with an integrated prediction of pilots' mistakes. Experiments and studies of the network were conducted on real and test samples. The upgraded hybrid neuro-fuzzy classifier structure and the learning algorithm can solve the problem of the need for multiple individual performance measurements, the dynamics of which would make it possible to build a trend and solve the problem on small samples. Used in organizational and management activities, this principle can help in predicting the danger caused by the human factor.

*Keywords:* forecast; wrong actions of the pilot; intellectual support; hybrid neuro-fuzzy classifier; two-layer perceptron; small samples

---

## 1. Introduction

Throughout the history of aviation, pilot's error as a safety-reducing factor was the subject of attention of the flight-technical and airlines management personnel and of the researchers in the field of psychology. For a long time, claiming the pilot's error, made him guilty of the task failure, of the equipment damage and, eventually, of his death. Therefore, all measures to combat errors were aimed at professional selection, training and «education» (punishment). These measures are obviously necessary, but not sufficient, since the errors are committed by highly-qualified pilots, which suggests that the human factor accidents are not limited to the issue of professional incompetence.

Pilot's activities feature unusual for other professions spatial orientation. The pilot evaluates the aircraft position in space according to the visual reference from the ground and from electronic devices, and in bad weather and visibility conditions - only from instruments. The quality of pilot-instruments interaction is largely determined by his individual psychological characteristics.

Psychological characteristics of the person are one of the main causes of air accidents. In this case, the role of self-esteem, stress levels and their influence on the pilots' professional qualities are the most interesting aspects in the error occurrence [1].

## 2. The object of the study

The study was conducted in the air squadron in one part of the closed garrison. Based on the data of psychological testing, we have created a pilot's professional efficiency diagnostic forecast complex.

### 2.1. Mathematical formulation of the problem

Every person can be described by a finite set of attributes [2], in this case, the characteristics obtained by the psycho-diagnostics,  $A = \{A_1, A_2, \dots, A_n\}$ , where  $n = 1,55$ . Each of  $A_i$  corresponds to a universal set of  $U_i$ , consisting of linguistic variables and numerical values  $\{a_{1i}, a_{2i}, \dots, a_{nii}\}$ , where  $i = 1, n$ .

In turn, each element of the function has its own identity.

The result set  $R = \{R_1, R_2, \dots, R_k\}$ , where  $k = 1,4$ .

$R = \{\text{not suitable, partly suitable, mainly suitable, suitable}\}$

The fuzzy base of rules in general will look as follows:

$IF (A_1 = a_{i11}^1, A_2 = a_{i22}^1, \dots, A_n = a_{inn}^1) \text{ then } (\mu_{R1} = \mu_1^1, \mu_{R2} = \mu_2^1, \dots, \mu_{Rk} = \mu_k^1)$

$IF (A_1 = a_{i11}^2, A_2 = a_{i22}^2, \dots, A_n = a_{inn}^2) \text{ then } (\mu_{R1} = \mu_1^2, \mu_{R2} = \mu_2^2, \dots, \mu_{Rk} = \mu_k^2)$

...

$IF (A_1 = a_{i11}^r, A_2 = a_{i22}^r, \dots, A_n = a_{inn}^r) \text{ then } (\mu_{R1} = \mu_1^r, \mu_{R2} = \mu_2^r, \dots, \mu_{Rk} = \mu_k^r)$

where  $\mu_{Rl}$  - the degree of the rule belonging to  $R_l$  class.

## 3. Methods

Regression analysis appears to be the traditional method of forecasting in psychology. It is assumed that the values of the time series is a random time function, and the task is to identify the correct model. The choice of the form of the function is not formalized and depends entirely on the expert's experience. At the same time, a neural network acts as a universal approximator of the training data, so the use of neural networks for the prediction is very promising.

In addition, the neural network can be seen as an adaptive model, as it can develop while gaining new information. Human behavior by nature is evolutionary, and the use of static models leads to the forecast quality deterioration.

Another problem that we faced was the need for a large amount of input data for network training. It is usually assumed that the time series contains at least hundreds of values, and it is impossible for us to complete this amount of observations. However,

there are opportunities to train the neural network on small amounts of input data. In this case, the a repeated learning on the same examples is being used, as well as different methods of time series processing, allowing to extend the training set.

The peculiarity of the problem lies in the fact that self-assessment and stress levels cannot be the input vector for the prediction of the network. The input is the values vector of the professional suitability dynamics for the period from six months to two years (that means, from 3 to 12 measurements, testing being conducted no more frequently than once every two months). By the professional suitability dynamics of the candidate we mean the degree of affiliation to one of four classes: the candidate fully meets the requirements of the specialty, basically corresponds, partially meets or does not meet - which in turn is obtained by analysis of 55 psychological characteristics.

Since the information, based on which a decision on the professional suitability of the candidate is made, is the result of various psychological techniques, classified data should be inaccurate or poorly defined. Due to this fact, it is necessary to use fuzzy logic and fuzzy sets theory as an effective approach to solving this problem.

To solve all the problems above, a modified hybrid model of neuro-fuzzy classifier with an integrated forecast function has been developed.

### 3.1. Hybrid neuro-fuzzy classifier

San and Jang offered the architecture to solve the fuzzy classification problem[3]. One possible structure of a hybrid neuro-fuzzy classifier is shown in Fig. 1.

Neuro-fuzzy network consists of four layers.

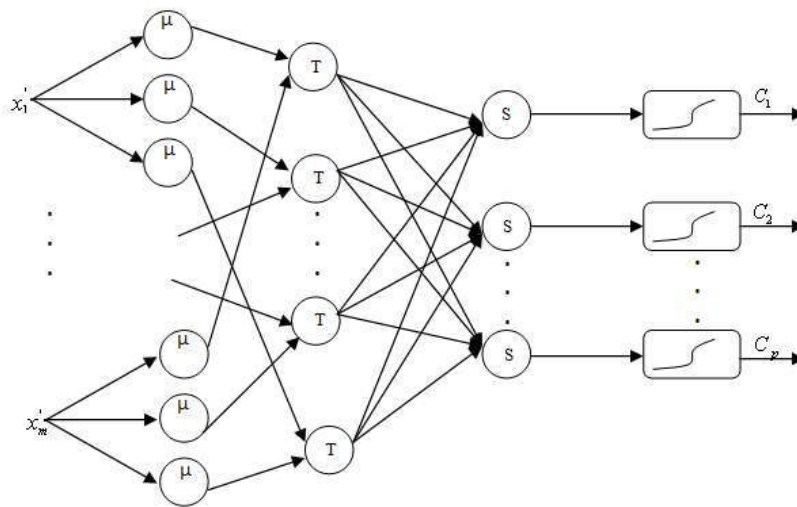


Fig. 1. Structure of the neuro-fuzzy classifier.

First-layer-elements implement the fuzzification operation, in other words, they form the degree of the membership of input data for the defined fuzzy sets  $A_{ij}$

$$\mu_{A_{ij}}(x'_j) = \exp \left[ -\frac{1}{2} \left( \frac{x'_j - c_{ij}}{\sigma_{ij}} \right)^2 \right]$$

where  $c_{ij}, \sigma_{ij}$  – the parameters of the membership bell-shaped type function.

The initial values of these parameters are set so as membership function satisfies the completeness, normality and convexity properties. Values should be equally distributed in the input vectors  $X$ . The values of these parameters can be adjusted in the process of the network education, which is based on the gradient method.

Each element of the second layer is an "I" neuron. It performs the aggregation of each database rule prerequisites truth degrees according to the T-norm operation using the following formulas:

$$\alpha_1 = \min\{A_{11}(x_1), A_{12}(x_2), \dots, A_{1n}(x_n)\}$$

$$\alpha_2 = \min\{A_{21}(x_1), A_{22}(x_2), \dots, A_{2n}(x_n)\}$$

...

$$\alpha_n = \min\{A_{n1}(x_1), A_{n2}(x_2), \dots, A_{nn}(x_n)\}$$

Third-layer elements perform the aggregation of each database rule prerequisites truth degrees according to the S-norm operation.

To solve the problem of candidates' classification for vacancies, basing on psycho-diagnostics, an input volume is quite small, with an average of 50 values. In order to speed up the network training algorithm and its simplifications, we should replace neurons of the third layer with the neurons that perform normalization and calculate the following values:

$$\beta_1 = \frac{\alpha_1}{\alpha_1 + \alpha_2 + \dots + \alpha_n}$$

$$\beta_2 = \frac{\alpha_2}{\alpha_1 + \alpha_2 + \dots + \alpha_n}$$

...

$$\beta_n = \frac{\alpha_n}{\alpha_1 + \alpha_2 + \dots + \alpha_n}$$

The elements of the fourth layer are used to calculate the conclusion values for each rule:

$$y_1' = B_1^{-1}(\alpha_1) = a_1 + \frac{1}{b_1} \ln \frac{1-\alpha_1}{\alpha_1}$$

$$y_2' = B_2^{-1}(\alpha_2) = a_2 + \frac{1}{b_2} \ln \frac{1-\alpha_2}{\alpha_2}$$

...

$$y_n' = B_n^{-1}(\alpha_n) = a_n + \frac{1}{b_n} \ln \frac{1-\alpha_n}{\alpha_n}$$

where  $a_i, b_i$  – nonlinear membership function parameters  $\mu_{B_i}(y)$  to the rule conclusion fuzzy sets.

Fuzzy network outputs are computed as follows:  $y_i' = \beta_i B_i^{-1}(\alpha_i)$ .

These outputs are interpreted as the membership degree of the object to the corresponding class. Since hybrid neuro-fuzzy classifier is represented as a multi-layer structure with a direct signal spread and the output variable value can be changed by adjusting the parameters of elements in layers, the gradient algorithms can be used to train the network.

Using this neuro-fuzzy network model the problem of classification can be solved, the results of which are the input vector for the prediction network.

### 3.2. The network structure

Then, a conventional two-layer perceptron can be added to a modified hybrid neuro-fuzzy classifier [4] with an additional neuron, which accumulates input values for classification prediction vector and, in fact, is one of Grossberg star [5] (Fig. 2). Two-layer perceptron was implemented without any changes.

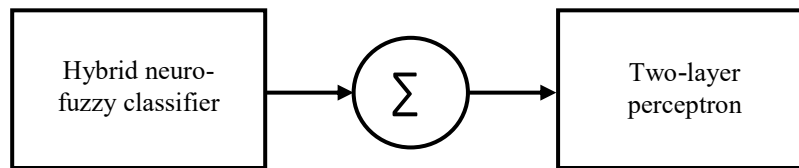


Fig. 2. The network structure.

## 4. Results and discussion

Tsukomoto algorithm was implemented in the hybrid neuro-fuzzy classifier, as well as a backpropagation method – a learning algorithm. The influence of the hybrid neuro-fuzzy classifier was also detected. The optimal structure of the network was selected: the volume of training sample - 35 samples, one network learning step  $h = 0,45$  and Gaussian fuzzification function, defuzzification method based on the r.m.s. deviations.

A study of the prediction quality using the constructed neural network was conducted on the test and the real-time series. For each time series, a structure of the network was chosen, providing the best quality of forecasting. The results are shown in Table 1.

Table 1. The frequency of the various structures use for the two-layer perceptron.

The number of inputs	Kn/s				Total
	1%	5%	10%	15%	
2	12	7	1	2	22
3	3	8	9	11	31
4	0	0	3	1	4
5	0	0	2	1	3
The number of neurons in the hidden layer	1%	5%	10%	15%	Total
2	13	15	11	10	49
3	2	0	3	4	9
4	0	0	1	1	2

The prediction is considered to be sufficiently accurate, if the prediction error is not more than 20%.

Fig. 3 shows the dependence of the prediction accuracy on the noise effects using different methods.

The "ideal" forecast - a forecast, the values deviation of which is caused only by random factors. If the prediction error is slightly different from the error of the "ideal" forecast, then it can be considered accurate.

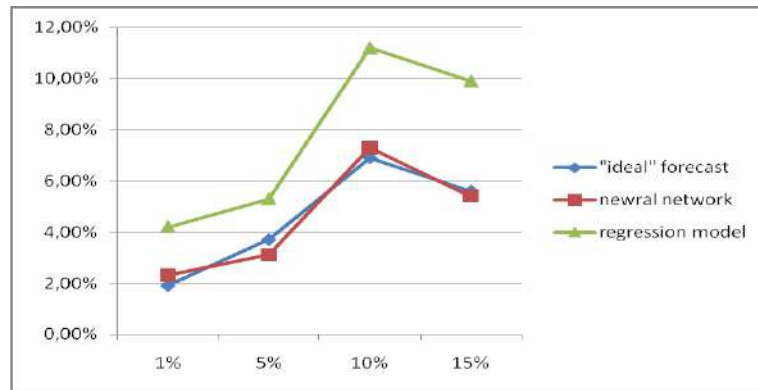


Fig. 3. Dependence of the prediction accuracy on the noise effects.

$$K1 = \frac{\sigma_{\varepsilon}^2}{\sigma_D^2} \cdot 100\% \quad \text{- noise / signal coefficient, which is a ratio of noise power to the power of the desired signal.}$$

$$K2 = \sqrt{\frac{\sum_{t=n+1}^{n+l} (Y_t^* - Y_t)^2}{\sum_{t=n+1}^{n+l} Y_t^2 + \sum_{t=n+1}^{n+l} (Y_t^*)^2}} \quad \text{- inconsistency coefficient (the second Teil coefficient), estimating the forecast accuracy.}$$

## 5. Conclusion

According to the study it can be concluded that the predictions, obtained using a neural network, have high level of accuracy and for many dynamics types seem to be significantly superior to the ones obtained using the regression model.

Moreover, the upgraded hybrid neuro-fuzzy classifier structure and the learning algorithm can solve the problem of the need for multiple individual performance measurements, the dynamics of which would make it possible to build a trend and solve the problem on small samples.

This approach allows with a certain degree of probability to calculate a predisposition to wrong actions in each case. If used in organizational and management activities, this principle can help in predicting the danger caused by the human factor.

## References

- [1] Danilenko AN, Ihsanova SG, Komakov VV. The diagnostic prediction of the professional performance of the pilot. Moscow: Mechanical engineering Flight 2012; 7: 53–60
- [2] Novak V, Perfilieva I, Mochkorzh I. Mathematical Principles of Fuzzy Logic. Moscow: Fizmatlit, 2006; 252 p.
- [3] Borisov VV, Kruglov VV, Fedulov AS. Indistinct models and networks. Moscow: The hot line –Telecom, 2007; 284 p.
- [4] Osovsky S. Neural networks for information processing. Moscow: Finance and Statistics, 2002; 344 p.

# Development of methods and algorithms for the classification of neurodegenerative diseases of the brain using MRI images

Olga Vasilchuk<sup>1</sup>, Alexey Fedorov<sup>2</sup>

<sup>1</sup>*Volga Region State University of Service, 4 Gagarin st., 445677, Togliatti, Samara region, Russia*

<sup>2</sup>*National Research University of Electronic Technology (MIET), Bld. 1, Shokin Square, 124498, Zelenograd, Moscow, Russia*

---

## Abstract

The article considers algorithms and methods for the classification of neurodegenerative diseases, in particular Alzheimer's disease and dementia. Stages are considered for carrying out such studies, including the search for raw materials, pre-processing and processing of MRI images.

*Keywords:* fMRI, neural networks, ROI extraction, Bag-of-Visual-Words, support vector machine

---

## 1. Introduction

Information technologies are presented in all areas of life, expanding opportunities and providing new tools. Medicine is no exception, as the technical complexity of the equipment used is also constantly growing. The quality and volume of information obtained from modern medical equipment makes possible statistical and other types of research based on stored data.

One of the most universal and fast tomographic methods for studying the human body is magnetic resonance imaging. It is widely used for the studies of brain, cerebral and neck vessels, temporomandibular joints, eye orbit, paranasal sinuses and oropharynx, soft tissues of the neck, spine, spinal cord, osteoarticular system, abdominal cavity organs and abdominal space, small organs Pelvis, chest, heart, arteries and veins, tumors and metastases.

At the moment, pictures taken by magnetic resonance imaging are stored in various digital formats. There are many software products with proprietary and open source code that allow specialists to view and process such images. But software products that provide the ability to conduct automated diagnostics are very few and they have many limitations.

Alzheimer's disease is the most common form of dementia. Among all reported cases of dementia Alzheimer's disease is 60-80%. Moreover, among people who have reached the age of 65, about 5% suffer from this ailment. And among people older than 85 years, the diagnosis of "Alzheimer's disease" is already 30%[1]. Alzheimer's disease belongs to diseases that impose the heaviest financial burden on society in developed countries[2]. At the moment the disease is incurable. The body of patients in the end, in any case, will lose most of its functions, which will lead to death. The newest medicines make it possible to alleviate the symptoms and to postpone the moment of complete erasure of the patient's personality. The quality of treatment depends on the stage of the disease, which was diagnosed.

All images taken with MRI are currently being processed by software, in general, not including automatic diagnostic tools. The doctor, when studying a picture of a patient, may miss the initial stage of Alzheimer's disease, if specifically does not focus on this. The presence of an effective algorithm that gives a probabilistic assessment of the possibility of a neurodegenerative disease would allow one to notice Alzheimer's disease at an early stage and begin treatment, thereby prolonging the patient's full life.

## 2. Principles of Magnetic Resonance Imaging. The Format of the Data Received from the MRI Scanner.

Magnetic resonance imaging (MRI) is a method of obtaining tomographic medical images for the study of internal organs and tissues using the phenomenon of nuclear magnetic resonance. The method is based on measuring the

electromagnetic response of atomic nuclei, most often the nuclei of hydrogen atoms, namely, their excitation by a certain combination of electromagnetic waves in a constant magnetic field of high tension.

After the research, a special file is created that contains information about the patient, research, and information for drawing the image. In fact, each file is a slice of some part of the body in any plane. The physical meaning of each pixel is the intensity of the return signal received by the scanner (simplified, tissue density of the body). The diagnostic station produces not one file, but several for one study. These files have a logical structure. Files are combined in a series and represent a set of consecutive sections of an organ. The series are combined in a stage. The stage determines the entire study. The sequence of the series in the stage is determined by the research protocol.

To visualize the data contained in the image, you need to compare the density of the texture and the color. Various transfer functions are used for this. Transfer functions are divided by type into absolute and relative ones. The absolute transfer function is constructed for all possible densities. MR-tomograph for each series generates its own set of densities. That is, for two series, the same density can correspond to different tissues of the body. Relative transfer function is built on the basis of the so-called window, which indicates which particular range of densities to draw.

Since the set of sections for MR-tomography can be represented as three-dimensional data, the concept of a voxel is introduced. Voxel is an element of a 3D image containing the value of an element in a three-dimensional space. As a voxel value, color can generally be used, but density is often used. As for the voxel shape, in general, voxels can be cubic, or a parallelepiped[3].

### 3. Databases of MRI images

There are several databases with pictures of MRI of the brain. They are designed for research in the field of automated diagnosis of various diseases. Databases storing pictures of patients with Alzheimer's disease will be considered.

#### 3.1. The BRAINnet Database

Brain Resource Ltd. Provides processed data from the Brain Resource International database available to BRAINnet for independent scientific use, freely and without restrictions on publication. The international database of brain resources is the largest accessible library of information on human brain health, obtained using standardized measures, several sources of data are available for the same individuals.

The database contains about 5000 pictures of healthy people, as well as about 1000 pictures of patients with various diseases, including Alzheimer's disease[4].

#### 3.2. The Open Access Series of Imaging Studies (OASIS)

The Open Access Research Series (OASIS) is a project aimed at making the MRI data set available to the scientific community. By compiling and freely distributing MRI data sets, the project is aimed at future discoveries in basic and clinical neurology. OASIS is provided by the Research Center for Alzheimer's Disease Research in Washington, DC, by Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the NRG Research Group at the University of Washington School of Medicine, and the BIRN[5].

Contains two sets of data:

- Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults
- Longitudinal MRI Data in Nondemented and Demented Older Adults

#### 3.3. OpenfMRI

The OpenfMRI database is a repository of human brain imaging data collected using MRI and EEG techniques. The data is collected from 2010. Initially, the project included datasets of functional MR imaging, but subsequently became open to all forms of neuroimaging data that included MRI data. The success of the platform can at least partly be explained by the simplicity of the organization of data and the absence of any obstacles to accessing the data. To obtain data that is distributed by default using the Public Domain license, no registration or license agreement is required[6].

### 3.4. Alzheimer's Disease Neuroimaging Initiative (ADNI)

ADNI is a continuous, multifaceted study designed to develop various data for the early detection and tracking of Alzheimer's disease (AD). It is divided into two main study periods (ADNI1 in ADNI2). The study aimed to enroll 400 subjects with early mild cognitive impairment (MCI), 200 subjects with early AD, and 200 normal control subjects[7].

## 4. ROI extraction

Since the processing power is limited in order to be able to process such a large amount of data, and to facilitate the task of training neural networks, the ROI extraction prestep is using. There are two common used methods for ROI extraction.

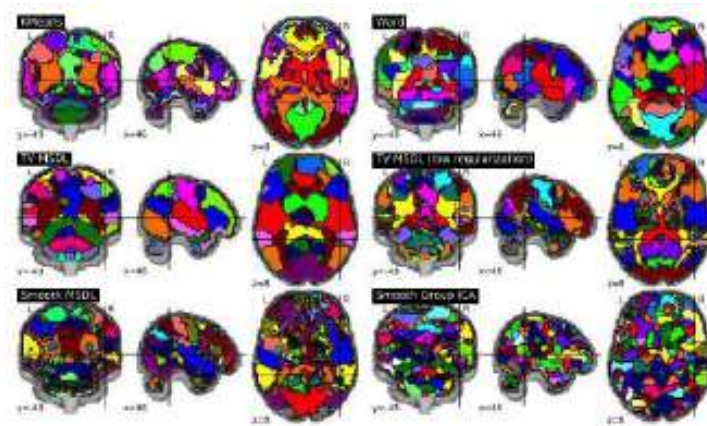


Figure 1: ROI extraction

### 4.1. Methods are based on dictionary learning

Learning dictionaries is a learning method, the purpose of which is to search for a divided representation of input data as a linear combination of basic elements, as well as the basic elements themselves. These elements form the dictionary, and no requirement to contain them orthogonal, they can also be redundant. That makes the input signals to represent a larger dimension than the specifically observed signal. This quality leads to the presence of redundant elements that provide improved flexibility of component separation and presentation. At the same time, they allow multiple representations of the same signal.

This method requires that the dictionary for learning should be composed of input data. The use of the dictionary learning method was due to the fact that signal processing usually requires the presentation of input data, in which a large number of different components are involved. Prior to this approach, the general practice was to use predefined dictionaries (such as the Fourier transform or the Wavelet transform). However, in some cases, a dictionary that is learned to customize input data can significantly improve component separation, which has value in decomposing, compressing, and analyzing data[8][9].

### 4.2. K-Mean method

The k-means method is the most popular method of clustering. He was invented in the 1950s by the mathematician Hugo Steinhaus and almost simultaneously by Stuart Lloyd. Particularly popular after McQueen's work. The action of the algorithm is such that it tends to minimize the total quadratic deviation of cluster points from the centers of these clusters:

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

$k$  - clusters number,  $S_i$  - received cluster,  $i = 1, 2, \dots, k$   $\mu_i$  - centers mass of the vector  $x_j \in S_i$ .

The algorithm is a version of the EM algorithm, which is also used to separate the Gaussian mixture. It splits the set of elements of the vector space into a known number of clusters  $k$ .

The basic idea is that at each iteration the center of mass is recalculated for each cluster obtained in the previous step, then the vectors are divided into clusters again according to which of the new centers is closer to the selected metric.

The algorithm is completed when, at some iteration, there is no change in the center of mass of the clusters. This happens for a finite number of iterations, since the number of possible partitions of a finite set is finite, and at each step the total quadratic deviation of  $V$  does not increase, so cycling is impossible[10].

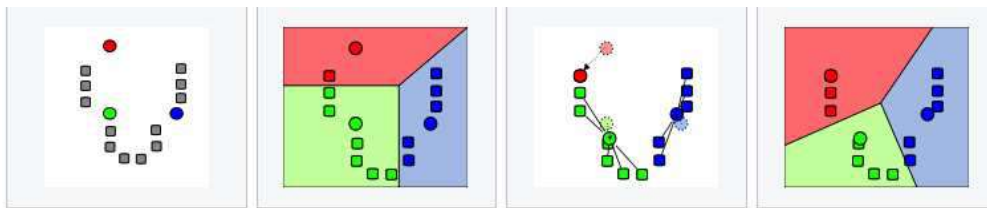


Figure 2: The K-means algorithm

## 5. Classification Methods Based on Neural Networks

### 5.1. A bag of visual words (BoVW) and the method of reference vectors

There is a study in which an approach combining several algorithms for the classification of MRI images is applied[11]. The scheme of the framework is shown in the figure 3.

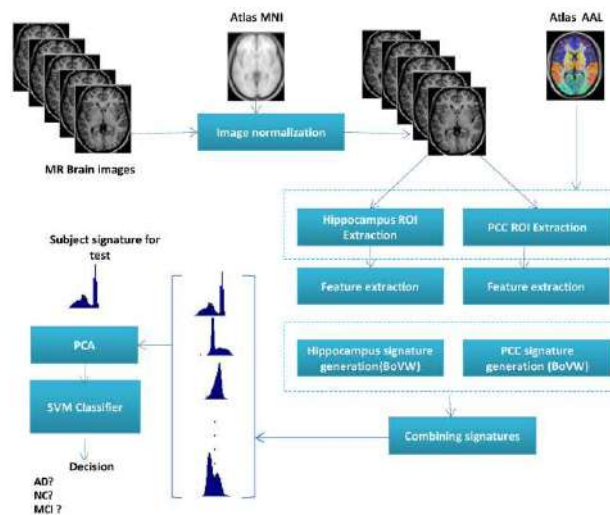


Figure 3: The possible framework overview

The method begins with the normalization of the image of the brain. Then the areas of interest (the hippocampus and the back waist crook) are extracted from normalized images, described by local visual descriptors, and processed within BoVW[12]. After decreasing the dimension, the resulting descriptors are classified using SVM.

A bag of words is a method originally created for the analysis of texts. In fact, a bag of words is a collection of word pairs - the number of its appearances in the text. In the case of images, everything is the same, with the only difference that instead of words, averaged fragments of images are used.



The support vector machine (SVM) is a set of similar learning algorithms with the teacher used for classification tasks and regression analysis. Belongs to the family of linear classifiers. A special property of the support vector method is a continuous decrease in the empirical classification error and an increase in the gap, so the method is also known as the classifier method with the maximum gap. The main idea of the method is the translation of the initial vectors into a space of higher dimension and the search for a separating hyperplane with the maximum gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane that separates the classes. The separating hyperplane is a hyperplane that maximizes the distance to two parallel hyperplanes. The algorithm works under the assumption that the greater the difference or the distance between these parallel hyperplanes, the smaller will be the average classifier error[13].

Each data object is represented as a vector (point) in  $p$ -dimensional space (an ordered set of  $p$  numbers). Each of these points belongs to only one of the two classes. The question is whether it is possible to separate points by a hyperplane of dimension  $p - 1$ . This is a typical case of linear separability. The desired hyperplanes can be many, so it is believed that maximizing the gap between classes contributes to a more confident classification. That is, it is possible to find such a hyperplane so that the distance from it to the nearest point is maximal. If such a hyperplane exists, it is called the optimal separating hyperplane, and the corresponding linear classifier is called the optimally separating classifier.

The points have the form:

$$\{(x_1, c_1), (x_2, c_2), \dots, (x_n, c_n)\} \quad (1)$$

Where  $c(i)$  takes the value 1 or -1, depending on which class the point  $x(i)$  belongs to. Each  $x(i)$  is a  $p$ -dimensional real vector, usually normalized by the values [0,1] or [-1,1].

If the points are not normalized, the point with large deviations from the average coordinates of the points will affect the classifier too much. We can treat this as a learning collection, in which the class to which it belongs is already assigned for each element. We want the algorithm of the support vector method to classify them in the same way. To do this, we construct a separating hyperplane that looks like this:

$$w \cdot x - b = 0 \quad (2)$$

The vector  $w$  is the perpendicular to the separating hyperplane. If the parameter  $b$  is zero, the hyperplane passes through the origin, which limits the solution. Since we are interested in the optimal separation, we are interested in support vectors and hyperplanes parallel to the optimal and closest to the supporting vectors of two classes. If the training sample is linearly separable, then we can choose hyperplanes in such a way that no points of the training sample lie between them and then maximize the distance between the hyperplanes. The width of the strip between them is easy to find from considerations of geometry, so our task is to minimize  $\|w\|$ .

$$c_i(w \cdot x_i - b) \geq 1, 1 \leq i \leq n \quad (3)$$

## 5.2. Convolution artificial neural networks

The neural net (CNN) – a special architecture of artificial neural networks, is part of the technology of in-depth training. Uses some features of the visual cortex, in which so-called simple cells reacting to straight lines from different angles were discovered, and complex cells whose reaction is associated with the activation of a certain set of simple cells. Thus, the idea of convolutional neural networks is the interleaving of convolutional layers and sub-sampling layers. The network structure is unidirectional (without feedbacks), essentially multilayered. For training, standard methods are used, most often the method of back propagation of the error. The function of activation of neurons (transfer function) is any, at the choice of the researcher. The operation of a convolutional neural network is usually interpreted as a transition from specific image features to more abstract details, and further to even more abstract details, up to highlighting high-level concepts. At the same time, the network is self-tuning and develops the necessary hierarchy of abstract attributes (sequences of feature cards), filtering unimportant details and highlighting the essential[14].

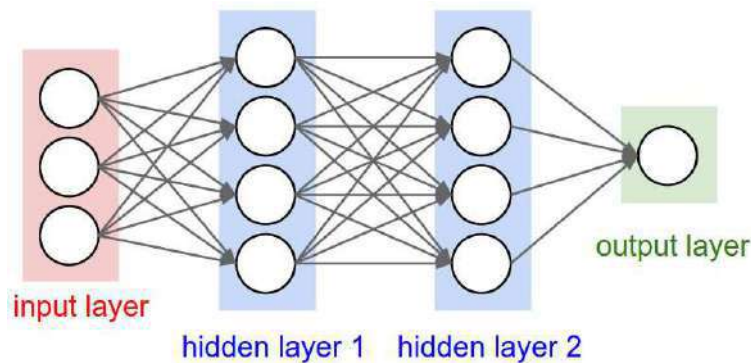


Figure 4: 3 layer neural network[15]

## 6. Conclusion

Modern methods allow them to be used for processing MRI images for the purpose of early diagnosis of brain diseases. It is necessary to develop and improve technologies that could do this automatically in the framework of other brain research. Such technology will allow to increase the statistical probability of finding diseases which require early diagnosis for effective treatment, such as Alzheimer's disease. It also allows in the long term to conduct preliminary diagnosis of diseases and detection of the patient's entry into risk groups without direct medical involvement on the basis of a combination of factors derived from various diagnostic devices and the already collected medical history.

## References

- [1] Early-onset Alzheimers Disease: Nonamnesic Subtypes and Type 2 AD 13. *PMC* 2012; 43(8): 677–685.
- [2] Zekry D, Giacobini E. The Economical Impact of Dementia. *PMC* 2005; 34(1): 35–41.
- [3] Rinck PA et al. *Magnetic Resonance in Medicine*. URL: <http://www.magnetic-resonance.org/>, 2017.
- [4] The BRAINnet Database. URL: <http://www.brainnet.net/about/brain-resource-international-database/>, 2017.
- [5] The Open Access Series of Imaging Studies (OASIS). URL: <http://www.oasis-brains.org/>, 2017.
- [6] The OpenfMRI Database. URL: <https://openfmri.org/>, 2017.
- [7] The Alzheimer's Disease Neuroimaging Initiative (ADNI). URL: <http://adni.loni.usc.edu/>, 2017.
- [8] Engan K, Aase SO, Husoy JH. *Method of optimal directions for frame design*, 1999; 5.
- [9] Censor Y, Zenios SA. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press on Demand, 1997.
- [10] MacKay D. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [11] Ahmed OB, Mizotin M, Benois-Pineau J, Allard M, Catheline G, Amar CB, Initiative ADN et al. Alzheimer's disease diagnosis on structural MR images using circular harmonic functions descriptors on hippocampus and posterior cingulate cortex. *Computerized Medical Imaging and Graphics* 2015; 44: 13–25.
- [12] Qiu G. Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition* 2002; 35(8): 1675–1686.
- [13] Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995; 20(3): 273–297.
- [14] Zhang W, Itoh K, Tanida J, Ichioka Y. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied Optics* 1990; 29(32): 4790–4797.
- [15] Convolutional Neural Networks. URL: <http://cs231n.github.io/convolutional-networks/>, 2017.

# Use of graph-based and algebraic models in lifecycle of real-time flight control software

A. Tyugashev<sup>1</sup>

<sup>1</sup>Samara State Transport University, 18 1<sup>st</sup> Bezymyanny Per., 443067, Samara, Russia

---

## Abstract

Software faults are the causes for repeating catastrophes in modern space missions. There are various problems in lifecycle of flight control software including lack of adequate models of real-time control algorithms. Real-time control algorithms have the totally distinct nature in contrast to computational algorithms. The paper presents mathematical models suitable for analysis, design and formal verification phases of lifecycle of spacecraft's flight control software. Two kinds of models - graph-based and algebraic, are being described. These models were successfully introduced in computer aided software engineering toolset for design and verification of spacecraft's real-time onboard control software.

*Keywords:* real-time control algorithm; real-time flight control software; lifecycle of the flight control software; computer aided software engineering; graph based model; algebraic model

---

## 1. Introduction

The modern spacecraft usually has various onboard systems such as Energy Supply System, Motion Control System, Onboard Control System, Autonomous Navigation System, Telemetry System, Thermal Control System, etc. In turn, each onboard system consists of a set of devices, aggregates, sensors. We can state that spacecraft is a system of systems or complex of complexes. Functioning of all these devices should be coordinated both in time and logically. The real-time mode of functioning is a very important issue entailing more complex nature of required models to be used for adequate reflecting of features of onboard apparatuses. This aspect is also quite important when we deal with the problems connected with designing and implementation of dependable control system for the spacecraft. In accordance with the Ashby's Law of Requisite Variety, variety of onboard equipment's behavior requires variety of onboard control system. Today the control logic of onboard control system is implemented in onboard Flight Control Software. Roughly speaking, there is a special control software module for each onboard device or aggregate. There are also a lot of supplemental modules involved into organization of computational process, etc. This is an illustration to a structural aspect of complexity of modern spacecraft's onboard software. Another essential aspect corresponds to behavioral aspect of complexity. We can compare the onboard apparatus of the spacecraft with the orchestra with the string and wind instruments, drums, etc. But we need a conductor to get a symphony – violin should start at the right moment, next cello should start with accurately fulfilled delay, and so on. So, we need also a special sort of real-time software which will serve as an orchestra's conductor.

For example, there are about 500 software modules concurrently running at real-time mode onboard the modern spacecrafts manufactured by Samara Rocket and Space Center 'Progress' [1-4]. These modules have a different nature and objective – system, service, computational, support, etc. The very important part of the onboard flight control software is real-time control algorithms (analog of orchestra's conductor), or so named 'programs for complex functioning' (it means cooperative functioning of various onboard spacecraft's subsystems such as Motion Control System, Telemetry System, Energy Supply System, etc.). The purpose of this part of onboard software is to run needed 'functional' program modules, and execution of the needed commands by particular onboard equipment at 'right' moments of time with the proper considering of current situation. It is clear, that the overall success of space missions has a straight dependence on the correct functioning of program for complex functioning. This is an explicit example of mission-critical software. Herewith, the cost of the errors in such algorithms made at analysis, design and development phases of lifecycle is too high. The usual way for providing of reliability and quality of flight control software is many-staged testing and debugging with utilization of specially built test beds. This process is very labor and time consuming, but unfortunately it cannot guarantee the absence of the errors [3]. Unfortunately, we face with repeating catastrophes and faults in space missions caused by software errors. First well known incident happened in 1962 with Mariner-I space probe. Probably, the most expensive one was the explosion of Ariane-5 European Space Agency rocket during its first flight in 1996. The amount of loss was estimated more than 500 millions euros. We can also mention relatively recent widely discussed failures of onboard software of Mars Polar Lander, Mars Climate Orbiter and mars rovers.

What is a reason for it? Complexity of onboard equipment entails failures of devices (which can be parried by switching to reserve equipment executed by special software module, but it requires more complex control logic). Complexity of the spacecraft tasks entails more complex behavior (especially in abnormal situations). Complex behavior also requires complex control logic. Complex Control Logic entails 'broken phone effect' between onboard devices' specialists and programmers during coding it in Onboard Flight Control Software. Nowadays, complex control logic requires complexity of Flight Control Software. Complexity of Software means higher costs of software lifecycle. Moreover, complexity of software means more errors in software itself. Is it a vicious circle?

Summarizing, we can emphasize the following modern trends and problems in spacecraft control:

1. use of onboard computers as a main control system;

2. transfer of 'decision-making point' from Earth onboard;
3. growing of size and complexity of Flight Control Software (concurrent multi-tasking, hundreds of interacting modules, millions of lines of code);
4. software-based support of spacecraft's fault tolerance feature;
5. dozens of people including non-programmers, involved in lifecycle of Flight Control Software;
6. costs of Flight Control Software's lifecycle became a very significant part of space mission's total costs, moreover - design, development and testing of the Onboard Software often is a 'critical path' in network schedule of spacecraft 's producing as a whole;
7. labor costs of control system software's creation and testing is 10 times bigger than hardware related costs [3].

The very promising way in this area is application of formal verification methods [3,5]. Unfortunately, the main efforts in area in software formal verification is oriented to computational (data transforming) algorithms and software where adequate mathematical models and methods considering semantics of the algorithms were developed and researched. We can state the inadequacy of these models and methods to nature of real-time control algorithms consisted of not elementary computations but actions related to actuators and other spacecraft's hardware. Accordingly, the development of the adequate models for this kind of software is very important. The developed models can be utilized in methods of analysis, design and verification which reduce real-time lifecycle labor costs and provide needed level of dependability of real-time control algorithms. In [6], the 'basic' algebraic based mathematical model of real-time control algorithms was presented. This is a constructive model, allowing step-by-step building of control algorithms on the basis of 'elementary' actions – so called 'functional tasks', time intervals, and logical conditions. This paper presents some extended models which supplements and clarifies the basic model for further use with various purposes.

## 2. Real-Time Control Algorithms

The object of the study is real-time control algorithms. It should provide coordinated and well synchronized functioning of onboard spacecraft's systems containing various sensors, actuators, devices.

The very important features distinguishing the control algorithms from the data transformation algorithms are the following. First, we cannot correlate the function (in mathematical understanding), and the control algorithm. Moreover, the correctness of the control algorithm cannot be defined by the contents of the computer memory at the moment of algorithm's end. The correctness of the control algorithm depends on its behavior in full time interval of functioning. Moreover, the values of conditions during execution of data transforming program are totally defined by the input data while the values of the conditions to be considered in control algorithm, are unpredictable because they are formed by the parameters of physical processes in controlled object (for example, velocity of the spacecraft). Actions executed by the control algorithm also can change not only the data in memory of the onboard computer, but influence on the state of the controlled object. When we need, for example, land the spacecraft on the Mars, we need to implement the very complex sequence of the operations with the participation of various onboard devices and mechanisms – but all of them are under control of onboard software.

Unfortunately, the major efforts in the modeling of algorithms historically were focused on data transformers, since earliest models like Turing and Post Machines, Church's recursive functions and Markov's 'normal algorithmes'. But if we want to apply the promising modern methods like formal verification or automated synthesis of the control software with the guarantee of its properness, we need the adequate semantic model for the control algorithm.

We can describe the following features of 'traditional' computational programs. Their main goal is transformation of input data to output data. The main components are the data transformers. The computational program correct if it successfully finishes (if input data is right) meanwhile output data matches specification. Structure of the traditional step-by-step sequential computational algorithm can adequately be represented by flowchart with begin and end(s) nodes. There are no time constraints and timer(s). Semantics could be adequately formalized by

1. axiomatic semantics [8] (Hoare, Floyd):  $Pre \{S\} Post$ ;
2. denotational semantics describing mapping between sets [9].

Many control algorithms are being used by 'reactive' systems. The main features of such systems could be described as follows. The reactive system should right process the input event flow. The main components are actions (in contrast to computational algorithms). The following semantic models are used for reactive systems:

1. Kripke structures [10];
2. finite state machines - automata;
3. Petri Nets.

Other important features of algorithms used in reactive systems distinguishing them both from the traditional programs and real-time control systems:

1. there are no end of algorithms, use of infinite loop of event processing instead;
2. there are no time constraints and time scale (timers) - asynchronous nature;
3. correct, if algorithm implements required model and execute right actions as reaction on pre-defined events.

We deal with Time-Driven Real-Time Control Algorithms (RTCA). This kind of algorithms has very important distinguishes from the reactive systems. Time-Driven Real-Time Algorithm should implement required schedule(s) (the term 'cyclogramme' widely used in space industry). The main components of algorithm are actions.

Other essential features can be described as follows. There are begin and ends in contrast to reactive systems. There are 'hard' time constraints. There is time scale (timers) for quantitative description of time parameters. In other words, RTCA has a

synchronous nature. This kind of algorithm correct if it execute right actions at right time (more precisely, right time moments with right considering of the current situation). Thus, the adequate semantic models should consider the stated features.

The known semantic models are

1. timed automata [11];
2. timed Petri Nets [12].

These models are based on the idea of 'state'. Unfortunately, if we try to apply this approach to real world system, we suffer because of 'state explosion' problem. This is why the author try to develop and use in CASE toolset another model based on idea of process. In some vision, this model utilizes more high level of abstraction allowing avoiding undue detalization and using in practice.

As it presented in [6], we can use the following set of tuples ('quads') for representation of semantics of real-time algorithm built from actions executed at particular time if the values of specified logical conditions are equals to 1:

$$UA = \{ \langle f_i, t_i, \tau_i, \vec{l}_i \rangle \}, i = 1, \dots, N$$

Each  $i$ -th quad in the above set describes one action executed by the real-time control algorithm;  $N$  is a number of actions executed. Here  $f_i$  is an identifier of the action,  $t_i$  – starting time of the action,  $\tau_i$  - duration. Starting time and duration defined as integers, this is adequate time model in this case because the minimal time difference recognized by the control algorithm is a 'tick' of onboard clock generator. The set of elementary actions  $F$  should be previously defined,  $f_i \in F$ . Logical vector  $\vec{l}_i$  specifies the combination of the conditions, allowing action  $f_i$  to be executed in time interval  $[t_i; t_i + \tau_i]$ . For example, logical vector can looks as follows ( $[\alpha_1=1], [\alpha_2=H], \dots, [\alpha_M=0]$ ). The 1 and 0 values recognized as true and false, and the third value 'H' means that this condition does not have an impact to execution of the action at specified time. The number of conditions actual for the control algorithm as well as the set  $L$  of the condition itself should be settled simultaneously with the set  $F$  of actions. We can interpret the logical vector in the model as an analog of the well known 'guard' conditions.

Described model have a very clear and intuitive visual representation looks like Gantt diagram. This was a reason because this model and models inherited from it were intended to use and successfully applied during development of the GRAFKONT/GEOZ [6] integrated development suite for automated design and verification of onboard flight control 'complex functioning' software.

But the 'basic' algebra of the control algorithms had the restricted descriptive power, for instance, there was no possibility to specify arbitrary time intervals between the actions, so it was necessary to introduce 'fictive' actions to taking in account the delays. And the time had the 'relative' nature only; we had no mechanism for binding the actions to the particular moment.

### 3. Methods

#### 3.1. Extended algebraic model of the real-time control algorithm

There are some known models for parallel systems utilizing higher level of abstraction than state-based models. First of all, we should mention the following approaches:

1. process Algebras (Milner's CCS [14], Hoare's CSP [15], Bergstra's ACP [16], etc.);
2. temporal Logics (LTL, CTL, RTTL, etc.) [17-19];
3. Allen's Interval Logic [20].

Unfortunately, there are serious limitations for use of named models for Flight Control Software. If we talk about process algebras, we should underline the following. They provide just 'common' means for description of parallel execution of processes without opportunity to define coordination of processes' begins and ends. Process algebras do not support logical inference (reasoning about algorithms needed for verification purposes). And there are no any tools for description of various situations (conditioning for different variants of situations – regular, abnormal, emergency, etc.) of system's functioning. But in space missions we definitely need such instruments.

In contrast to process algebras, temporal logics natively have special means for description of concurrency of processes extended in time. But initially, there are no means for quantitative aspects of synchronization in temporal logics. In addition, the semantic models are not advanced enough for our practical purposes.

Allen's Interval Logic is a very interesting and promising approach providing means for description of all possible overlay of processes extended in time. But there are no means in Allen's logic for description of quantitative aspects of synchronization. Moreover, this approach also has no means for description of 'different branches' of Real-Time Control Algorithms.

The method we developed is free from the limitations described above. It was initially developed for description of complex Real-Time Control Algorithms with considering internal logic, different variants (branches) and situations. There are advanced means for description of synchronization of parallel processes. There are means for quantitative descriptions of synchronization both for relative and absolute timing.

The proposed model uses 'constructive' approach. We can construct new control algorithms using the existing ones by application of the set of operations. The basic operations introduced by Anatoly Kalentyev had only four operations –  $CH$ ,  $CK$ ,  $\rightarrow$ , and  $\Rightarrow$ . The extended algebraic model contains the following operations:

Table 1. Operations of the extended algebraic model of real-time control algorithms.

Name	Mean	Signature
<i>CH</i>	synchronization 'begin-begin'	$(UA_1, UA_2) \rightarrow UA$
<i>CK</i>	synchronization 'end-end'	$(UA_1, UA_2) \rightarrow UA$
$\rightarrow$	direct following	$(UA_1, UA_2) \rightarrow UA$
<i>H</i>	Overlay	$(UA_1, UA_2, \text{integer}) \rightarrow UA$
<i>3A</i>	parameterized following	$(UA_1, UA_2, \text{integer}) \rightarrow UA$
@	absolute time binding	$(UA, \text{integer}) \rightarrow UA$
$\Rightarrow$	qualification by the condition	$(\text{condition}, UA) \rightarrow UA$

'Begin-begin' synchronization applicable to the two control algorithms (denoted in the table above as  $UA_1$  and  $UA_2$ ) and forms a control algorithm which includes all quads from both  $UA_1$  and  $UA_2$ , but with the correction of the start time of each action inherited from the  $UA_2$ . To make the correction, we should calculate overall starting time of the  $UA_1$  and  $UA_2$ . It can be calculated as a minimum of the starting times  $t_i$  of the actions included into the UA:  $t_{UA} = \min_{i=1..N} t_i$ . After starting times for  $UA_1$  and  $UA_2$  will be found, we should calculate the difference  $\Delta = t_{UA_2} - t_{UA_1}$ . And finally we must add the difference to the all  $t_i$  inherited from the  $UA_2$ . As a result, we will have the control algorithm where all actions from the  $UA_2$  will be shifted and the first action from  $UA_1$  and  $UA_2$  begins at the same time. The *CH* operation is transitive, associative but not communicative.

'End-end' synchronization operation *CK* has the same signature as *CH*, and its result also includes all quads from both arguments. Again, the actions inherited from  $UA_2$  should be shifted by  $\Delta$ . But the rule for calculation of the  $\Delta$  is different. The latest action of  $UA_2$  in resulting algorithm should ends at the time when ends the latest action of  $UA_1$ . So, we need calculate overall finish time  $et$  for  $UA_1$  and  $UA_2$  as follows:  $et_{UA} = \max_{i=1..N}(t_i + \tau_i)$ . And in this case  $\Delta = et_{UA_1} - et_{UA_2}$ . The *CK* operation is transitive, associative, but not communicative like *CH*.

Direct following means that in the resulting algorithm the first action inherited from  $UA_2$  starts when the latest action inherited from  $UA_1$  ends. For this, we need make shift like in the cases above, but the rule is different. We need set the starting time of the earliest action of  $UA_2$ , as  $et_{UA_1}$ . Difference  $\Delta = et_{UA_1} - t_{UA_2}$  we then need to add to all starting times of actions in resulting algorithm, which are inherited from  $UA_2$ . In contrast to basic algebra of real-time control algorithms, initially proposed by A.A. Kalentyev, the extended algebraic model includes also the following operations.

Overlay operation *H* is similar to parameterized following *3A* operation. We form the control algorithm including all actions from the arguments  $UA_1$  and  $UA_2$  like for *CH* and *CK* operations, but applying difference is defined as the additional argument of the operation – integer number. The difference between *H* and *3A* is defined by the following. In the result of *H* the first action inherited from the  $UA_2$  should starts before the end of the latest action inherited from  $UA_1$  – so, we deal with 'overlay'. In case of *3A* operation, conversely, the earliest action inherited from the second argument, should starts after the end of the latest action inherited from  $UA_1$ , and plus one.

Absolute time binding operation allows setting the particular time as the starting time of the control algorithm (starting time of the earliest action). We should find the overall starting time for the existing UA, and then calculate the difference between  $t_{UA}$  and the second integer argument of the @ operation. After this, the difference  $\Delta$  should be added to all starting times  $t_i$  of the all actions to be executed by the algorithm.

Finally, the 'qualification' operation has the UA and the condition as the arguments. We can write  $(\alpha_i=0) \Rightarrow UA_i$ , and it will mean that in all logical vectors in the UA, we need to update the corresponding component (initially all values for all conditions for all actions settles as 'H', i.e. action is to be executed imperative).

This model can be successfully utilized for the purposes of specification and verification of the real-time control algorithms. For example, author's applied it at the corresponding stages of flight control software lifecycle for spacecraft [6].

### 3.2. Graph-based model

Anyway, if we want to solve the synthesis problem of the program, we need the adequate model with the more deep detailing.

The model in previous section specifies the control algorithm at the 'semantic' level. We can see parallel with the denotational semantics of the data transformation algorithms when we nominate the function to be calculated by the program. But the same semantics can be provided by different implementations. Moreover, it is well known fact that in case of software, the quality and characteristics – including efficiency, of the programs with the same semantics, can be quite different.

We need the models for the next degree of detalization. The very popular and effective in practice models in theoretical computer science are graph-based. They are applicable also at the stages of design and analysis of efficiency of implementation of flight control software as well. However, the known models require clarification and extension.

For example, the control flow graph is a very useful and popular model. But initially it describes sequential imperative program executed by the single CPU. Understandably, there are no any essences applicable for describing phenomena of time interval.

But the nature of the complex technical system likes modern spacecraft urgently requires concurrent (multi-tasking) model of computation. It is unsurprising that the onboard software of the modern spacecraft is executed by the control of multi-tasking operating system. Such systems have an application programming interface allowing one process to be started by another (fork). In many cases, onboard real-time operating system allows starting the process with the specified time delay or at absolute time.

This is the reason for defining the extended model – 'timed logical scheme' of the real-time control algorithm. First, we take the well known program control flow graph. In case of control algorithms, nodes with the single outgoing arcs will be associated with any actions executed by the algorithm. Nodes with the two outgoing arcs will be associated with the conditions formed not by the CPU flags only, but related to parameters of spacecraft motion, state of the onboard systems, etc.

We introduce also the special ‘weighted’ arcs. The weight specified by the integer, will denote the delay before the action associated with the following node be executed. The very important issue is that any ‘action’ node can have arbitrary number of outgoing ‘weighted’ arcs additionally to one ‘usual’ unweighted. The example of the timed logical scheme is presented in Figure 1 in alongside with the visual representation of the corresponding semantic model of control algorithm.

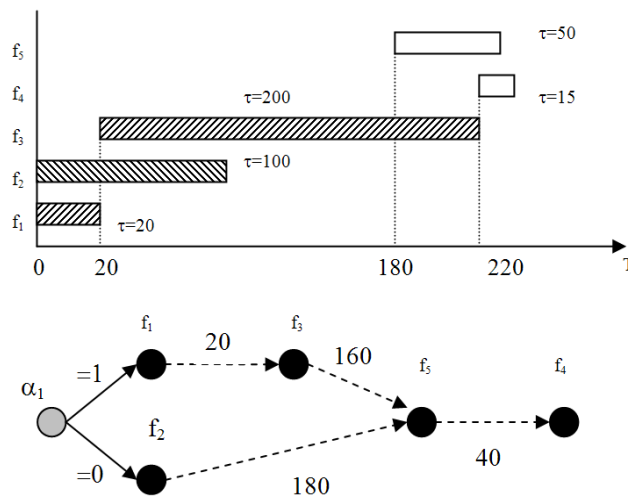


Fig. 1. Example of timed logical scheme.

The depicted UA is  $\{ \langle f_1, 0, 20, (\alpha_1=1) \rangle, \langle f_2, 0, 100, (\alpha_1=0) \rangle, \langle f_3, 20, 200, (\alpha_1=1) \rangle, \langle f_4, 180, 50, (\alpha_1=H) \rangle, \langle f_5, 220, 15, (\alpha_1=H) \rangle \}$ . Different qualification by the logical vectors can be shown by different texture or color on visual representation of the control algorithm’s semantics. Timed logical scheme implements the semantic by the fixation of the ‘key’ time moments – 0, 20, 180, 200 when the algorithm must perform actions. This ‘activations’ are divided by relatively time delays: 20, 160, 40, 180 associated with the weighed arcs in the graph-based model. We can see the ‘qualified’ branches of the timed logical scheme with the corresponding  $(\alpha_1=1)$  and  $(\alpha_1=0)$  conditions (in this case, there is only the single condition in the logical vector). Then, at timestamp 180, these branches join again.

#### 4. Results and Discussion

The presented models looks be much more adequate to the nature of the onboard spacecraft’s flight control software than approaches oriented to computational sequential algorithms. Nature of the presented models is quite corresponds to the domain of real-time control of technical complexes consisting of many subsystems, devices, aggregates, etc. The main components of control programs are actions which can be executed both by software modules and equipment. The conditions which influences the process of computation, also does not formed by the input data only, but permanently changing in accordance with the state of controlled object. The semantic model presented in the paper, initially oriented to this particularities.

These features provide possibility to potential application of these models in such area as SCADA systems, power plants, transport, etc. [7].

If we compare presented timed logical scheme of the algorithm with the timed automaton, for example, we will discover that despite there are possibility to describe time related issues, timed automata cannot be used to adequate and unambiguous descriptions of control programs. Timed logical scheme, conversely, initially was developed with this purpose and good corresponds to the factors of problem domain. Moreover, the features of real-time operating systems are taken into account.

#### 5. Conclusion

The paper presents two extended models for representation of real-time control algorithms implemented by spacecraft’s flight control software. One model is algebraic and another is graph-based. Algebraic model can be applied for ‘high level’ semantic modeling of the real-time control algorithms. This is important because known models were oriented to computational algorithms and did not take in account the nature of real-time control systems. Semantic models can be used for the accurate and unambiguous specification of flight control software and then applied for formal verification, which is reviewed nowadays as very promising method around the world.

Another presented model is graph-based. It allows analyze the efficiency issues and can be successfully used at the design stage. Constructive nature of these models and their clear visual representations allow developing of GRAFKONT/GEOZ software toolset for specifying and verification of real-time control software. The toolset was introduced at Samara Space Rocket Center.

## Acknowledgements

Author should say many thanks and acknowledgements to President of the Samara University Victor Soifer who granted him a possibility to research and develop the methods and tools for Russian space industry for many years.

## References

- [1] Kozlov DI, Anshakov GP, Mostovoy YaA, Sollogub AV. Control of Earth's Remote Sensing Spacecrafts: Computer Technologies. Moscow: Mashinostroenie, 1998; 368 p. (in Russian)
- [2] Kirilin AN, Anshakov GP, Akhmetov RN, Storozh DA. Spacecrafts Building. Samara: Agni Publishing House, 2011; 280 p. (in Russian)
- [3] Tyugashev A, Ermakov I, Ilyin I. Ways to get more reliable and safe software in Aerospace Industry. Proc. Program Semantics, Specification and Verification: Theory and Applications (PSSV), Russia: Nizhni Novgorod, 2012; 121–129.
- [4] Filatov AV, Tkachenko IS, Tyugashev AA, Sopchenko EV. Structure and algorithms of motion control system's software of the small spacecraft. Proceedings of Information Technology and Nanotechnology (ITNT-2015). CEUR Workshop Proceedings, 2015; 1490: 246–251.
- [5] Holzmann GJ, Havelund K, Joshi R, Xu R-G, Groce A. Establishing flight software reliability: testing, model checking, constraint-solving, monitoring and learning. Annals of Mathematics and Artificial Intelligence 2014; 70(4): 315–349.
- [6] Tyugashev AA. Integrated environment for designing real-time control algorithms. Journal of Computer and Systems Sciences International 2006; 45(2): 287–300.
- [7] Tyugashev A. Language and Toolset for Visual Construction of Programs for Intelligent Autonomous Spacecraft Control. IFAC-PapersOnLine. 4th IFAC Conference on Intelligent Control and Automation Sciences. France: Reims, 2016; 49(5): 120–125.
- [8] Hoare CAR. An axiomatic basis for computer programming. Communications of the ACM CACM 1969; 12(10): 576–580.
- [9] de Bakker JW. Least Fixed Points Revisited. Theoretical Computer Science 1976; 2(2): 155–181.
- [10] Schneider K. Verification of reactive systems: formal methods and algorithms. Springer, 2004.
- [11] Bengtsson J, Wang Yi. Timed Automata: Semantics, Algorithms and Tools. Lectures on Concurrency and Petri Nets. Lecture Notes in Computer Science; 3098: 87–124.
- [12] Zuberek WM. Timed Petri nets definitions, properties, and applications. Microelectronics Reliability 1991; 31(4): 627–644.
- [14] Milner R. A Calculus of Communicating Systems. Springer Verlag, 1980.
- [15] Hoare CAR. Communicating sequential processes. Communications of the ACM 1978; 21(8): 666–677.
- [16] Bergstra JA, Klop JW. ACPr: A Universal Axiom System for Process Specification. CWI Quarterly 1987; 15: 3–23.
- [17] Pnueli A. The temporal logic of programs. Proc. 18th Annual Symposium on Foundations of Computer Science (FOCS) 1977; 46–57.
- [18] Emerson EA, Halpern JY. Decision procedures and expressiveness in the temporal logic of branching time. Journal of Computer and System Sciences 1985; 30(1): 1–24.
- [19] Ostroff JS. Temporal Logic of Real-Time Systems. Advanced Software Development Series. Research Studies Press Ltd, England, 1990.
- [20] Allen JF. Maintaining knowledge about temporal intervals. Communications of the ACM 1983; 26: 832–843.



# Preface

Dmitry Savelyev<sup>1</sup>, Denis Kudryashov<sup>1</sup>

<sup>1</sup>*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

---

Session “Computer Modeling” was held at the 3rd International Conference on Information Technology and Nanotechnology - 2017 (ITNT-2017) in Samara, Russia, April 25–27, 2017 (<http://ru.itnt-conf.org/itnt17ru/>). This volume includes reports from different sections on computer modeling.

The goal of the ITNT-2017 Conference was to discuss problems of fundamental and applied research in information technology and nanotechnology, including but not limited to:

- Computer Optics;
- Diffractive Nanophotonics;
- Image Processing;
- High-performance Computing;
- Computer Vision;
- Mathematical Modeling;
- Data Science.

Scientists from Austria, Belarus, Bulgaria, Denmark, Germany, Great Britain, India, Iraq, Mexico, Moldova, Russia, Spain, USA, and Finland presented over 330 reports at the ITNT-2017 Conference. The most significant studies presented at the Conference published in *Procedia Engineering* (Elsevier, vol. 201).

We are grateful to everybody who has contributed to the session and look forward to meeting you again at future events. Further, we thank all the authors who presented their studies, as well as the reviewers and the delegates. Moreover, we sincerely thank the team of organizers for making the session successful and this publication possible.

## Guest Editors

- Roman Skidanov, Image Processing Systems Institute - Branch of the Federal Scientific Research Centre "Crystallography and Photonics", Russian Academy of Sciences, Samara, Russia;
- Vladimir Sobolev, Samara National Research University, Samara, Russia;
- Denis Kudryashov, Samara National Research University, Samara, Russia.

## Conference Organizers

- Samara National Research University
- Image Processing Systems Institute of RAS – Branch of the FSRC “Crystallography and Photonics” RAS
- Government of Samara Region
- Computer Technologies

## Chair

- Evgeniy Shakhmatov – Samara National Research University, Russia

## Vice-chairs

- Vladimir Bogatyrev – Samara National Research University, Russia
- Nikolay Kazanskiy – Image Processing Systems Institute - Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, Russia
- Eduard Kolomiets – Samara National Research University, Russia
- Alexander Kupriyanov – Samara National Research University, Russia

## Chair Program Committee

- Viktor Soifer, Samara National Research University, Russia

# Engineering of the fiber optic Bragg grating sensor of electrical parameters and software application for automatic simulation of its parameters

G.I. Leonovitch<sup>1</sup>, V.N. Zakharov<sup>1</sup>, A.I. Gorshkov<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

Nowadays one of the most effective transducer rised to the high demands based on metrological and exploitative characteristics are fiber optical. In the article there is modern state of measurement fiber optical sensors. Basic types and methods of measurement are examined. New model of fiber optic Bragg grating sensor for measurement of electric parameters is suggested. For the suggested model a program for computing the parameters of sensor is written, valid model is presented on experimental board. The results of the work and their valuating are received.

*Keywords:* fiber Bragg grating; fiber optic sensor; multisensor networks; direct current; commuted current; electrostatic field; magnetic field

## 1. Introduction

Nowadays one of the most effective transducer rised to the high demands based on metrological and exploitative characteristics are fiber optical, optomechanical, optoelectronic, transducer of the physical values with information transmission from sensor to controller by the fiber optic interconnections (with built-in fiber optic interconnections FOI).

Until quite recently the main type of sensor for measurement of mechanical deformation and temperature were strain gage sensors, piezotransducers, thermo-resistors and others. However thank for intensive development of fiber optics fiber optic sensors were elaborated and receives grate expansion possessing some advantages comparing to strain gage sensors: they possess higher sensitivity, interference protection, immunity to the influence of aggressive environments and lower cost.

Among fiber optic sensors the most perspective are quasi-distributed fiber optic Bragg grating sensor (further FOBGS), afforded to control the state of the object in most points at the same time due to the possibility of spectral and time multiplexing.

## 2. Development of the mathematical model

Bragg gratings bunch the main mode of fiber optic guide emitted in the straight direction in fiber optic guide with the main mode emitted in the opposite direction on the resonant wavelength  $\lambda_{Br}$  determined by the correlation [2]:

$$\lambda_{Br} = 2n_{eff}\Lambda,$$

where  $n_{eff}$  is efficient reflection index core of fiber of the main mode,  $\Lambda$  period of Bragg grating.

Spectral properties are the most important characteristics of Bragg gratings. The main of them are spectral location resonance its width and reflect coefficient at a maximum. Calculation of spectral characteristics of Bragg gratings usually accomplish with the use of mode coupling theory. Let's express the coefficient function of Bragg grating from wavelength by the mode coupling theory [3]:

$$r = \frac{sh^2(\gamma_B L)}{ch^2(\gamma_B L) - \frac{\sigma^2}{\kappa^2}},$$

where  $\gamma_B \equiv \sqrt{\kappa^2 - \sigma^2}$  is spectral offset from strict resonance  $\sigma$  determines by the difference of propagation constants of the main mode  $\beta = \frac{2\pi n_{eff}}{\lambda}$ :

$$\sigma(z) = \beta(z) - \beta_{Br}(z) = \frac{2\pi n_{eff}(z)}{\lambda} - \frac{\pi}{\Lambda(z)},$$

where local reflection effective value is  $n_{eff}(z) = n_{eff} + \eta \cdot \Delta n_{avr}(z)$

Coherence coefficient of grating  $\kappa(z)$  on the wave length  $\lambda$  is proportional to the mod modulating range induced reflection value  $\Delta n_{mod}(z)$ :

$$\kappa(z) = \frac{\pi \eta \Delta n_{mod}(z)}{\lambda}$$

$\eta$  – quantity of main mode power that propagates in the optic guide core.

Resonance spectral width on the half-height Bragg gratings might be expressing the following close correlation [3]:

$$\Delta\lambda_{0,5} = 2\lambda\alpha \sqrt{\left(\frac{\eta\Delta n}{2n_{eff}}\right)^2 + \left(\frac{\Lambda}{L}\right)^2}$$

where  $\alpha$  is about 1 for the deep grating (with reflection value  $r \sim 1$ ) and is about 0.5 for the small depth gratings.

For the grating with modulation period  $\Lambda = 67,06\mu m$  and with the refraction deviation index  $\Delta n = 10^{-4}$  spectrum of reflection of light signal will look as on the fig.1:

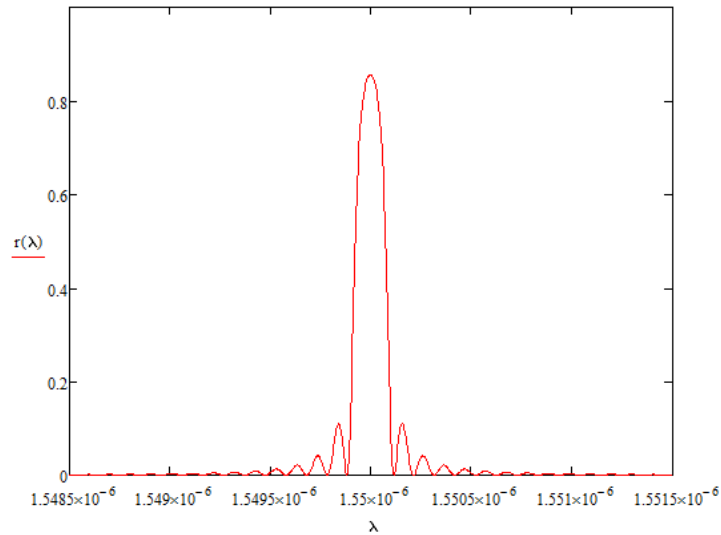


Fig. 1. Spectrum of the reflected signal.

Central wavelength of optic emission reflected by the Bragg grating depends on the effective refraction value and on the grating period. Changing of the central wavelength taking in account the influence of the temperature and mechanic strain determines this way [4]:

$$\Delta\lambda_B = 2 \left( \Lambda \frac{\partial n_{eff}}{\partial l} + n_{eff} \frac{\partial \Lambda}{\partial l} \right) + 2 \left( \Lambda \frac{\partial n_{eff}}{\partial T} + n_{eff} \frac{\partial \Lambda}{\partial T} \right),$$

where  $n_{eff}$  – is efficient reflection index core of fiber of the main mode,  $\Lambda$  period of Bragg grating.

The first component in this formula gives the value of wavelength shift depending on deformation (elongation). The second depending on the temperature. The dependence of the shift of the central wavelength of the reflected emission from the deformation also may be showed in the following way:

$$\Delta\lambda_B = \lambda_B (1 - p_e) \delta_Z,$$

where  $p_e$  is a constant of deformation optic fiber is calculated from the following formula:

$$p_e = \frac{n_{eff}^2}{2} (p_{12} - \nu(p_{11} + p_{12})),$$

where  $p_{11}$  and  $p_{12}$  Pockels coefficients in the tensor optical strains,  $\nu$  Poisson ratio. For the typical fiber  $p_{11} = 0,113$ ,  $p_{12} = 0,252$ ,  $\nu = 0,16$  and  $n_{eff} = 1,4447$ . On the bases of values of sensitivity for the wavelength  $\lambda_B = 1550 \text{ nm}$  will make  $12,36 \text{ nm}/\%$ .

By the stretching optic fiber the length of Bragg grating changes, the period of the modulation of the refraction index and there is the change of refraction indexes of core and cover of optic fiber. Formula for changing of efficient reflection index as result of stretching is designated by photoelastic effect so that

$$\Delta n_\delta = -\frac{1}{2} n_{eff}^3 \cdot p_e \cdot \delta_Z$$

it is related to anisotropy of optic fiber occurring by stretching.

The second element gives the dependence of wavelength shift from temperature. Emission wavelength reflected from Bragg grating sensors changes in dependence from temperature because of the following factors: heat expansion of optic fiber (stretches out period of Bragg grating) in other words there is a changing of grating mechanical length moreover there is a changing value of fiber refraction depending on temperature (changing of grating optical length). Whence it follows that the dependence of wavelength shift on temperature can be described the following formula [4]:

$$\Delta\lambda_B = \lambda_B (\alpha_\Lambda + \alpha_n) \Delta T,$$

where  $\alpha_\Lambda$  is temperature coefficient of linear expansion ( $\alpha_\Lambda = 0,55 \cdot 10^{-6}$  is for fused quartz),  $\alpha_n$  thermo-optic coefficient ( $\alpha_n = 8,6 \cdot 10^{-6}$  is for optic fiber with doped germanium). Due to these values of sensitivity of Bragg grating to the temperature for the wavelength  $\lambda_B = 1550 \text{ nm}$  will be  $14,1 \frac{\text{pm}}{^\circ\text{C}}$ .

The diagrams of wavelength dependence form the deformation and temperature are presented on fig. 2. The diagram of dependence from deformation is presented on the top; the diagram of dependence from temperature is below.

### 3. The software for simulation technical parameters

For the simulation of work of this type of sensors automation system was developed.

The window of this system is presented on the fig. 3. The user has various opportunities for editing parameters of simulation. After pressing the button «Добавить график» in the both diagrams new simulated data is appearing which are different from the previous by color. Therefore user has an opportunity of clearing the diagrams, all the fields and report generation according to data by pressing the button «Report».

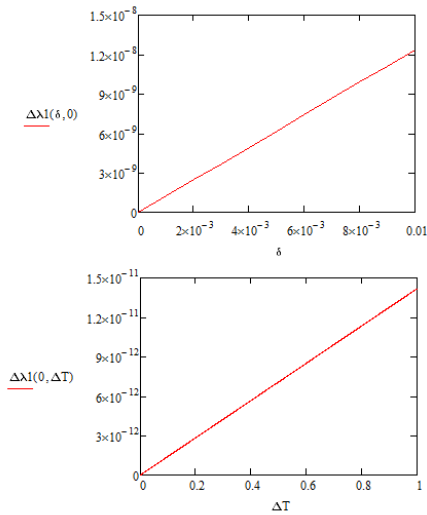


Fig. 2. The diagrams of wavelength dependence form the deformation and temperature.

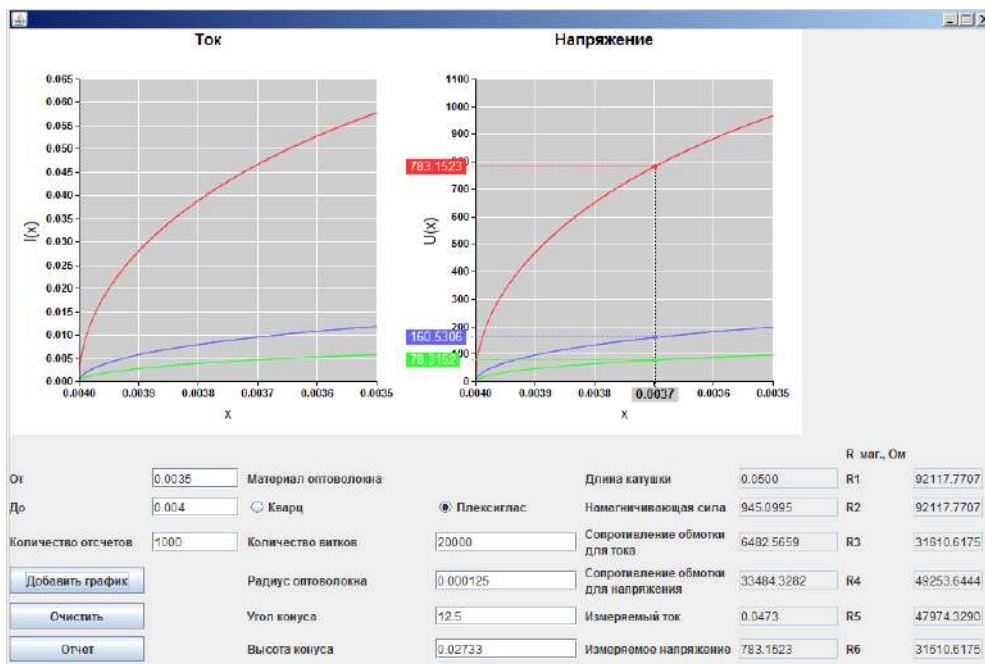


Fig. 3. Software interface.

#### 4. Laboratory tests

In the course of the works on optic fiber sensor of electrical parameters suggested mathematical model was taken. Further on the bases of this model the sensor design was developed for laboratory tests and exposure of efficiency of its work. On the fig. 4 principle diagram of sensor organization is presented.

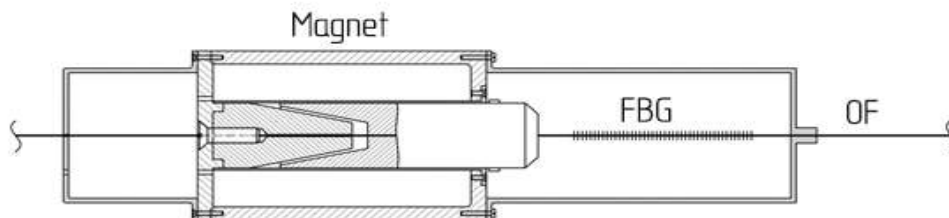


Fig. 4. Sensor organization.

In the course of the laboratory tests the experiments on the stand were performed where the sensor was assembled and the results of minimal sensitivity were received and the critical parameters of this sensor were committed (fig. 5, 6). Data received after performance of tests of this stand. Parameters of power supply: 20V, 2.5A. On the diagram we can see the surge of

wavelength changing reflecting specter from intrafibrous Bragg grating during the admission of power supply on the coil (number 7 on the fig. 5) (1524.990 – 1525.048nm).

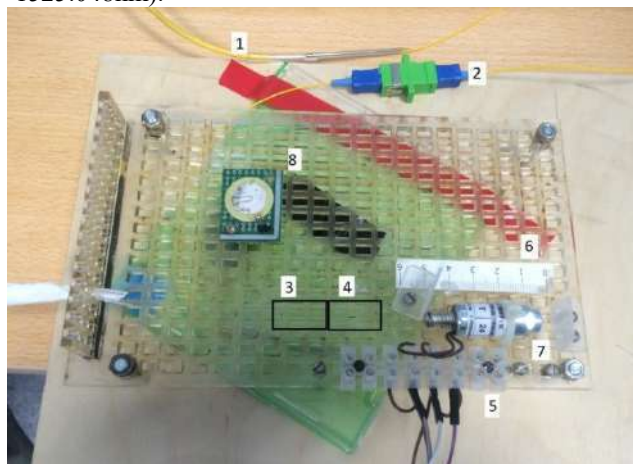


Fig. 5. The stand photo (1 optical fiber, 2 connector, 3,4 Bragg gratings, 5 electrical connectors, 6 ruler, 7 coil, 8 piezo element(is not used)).

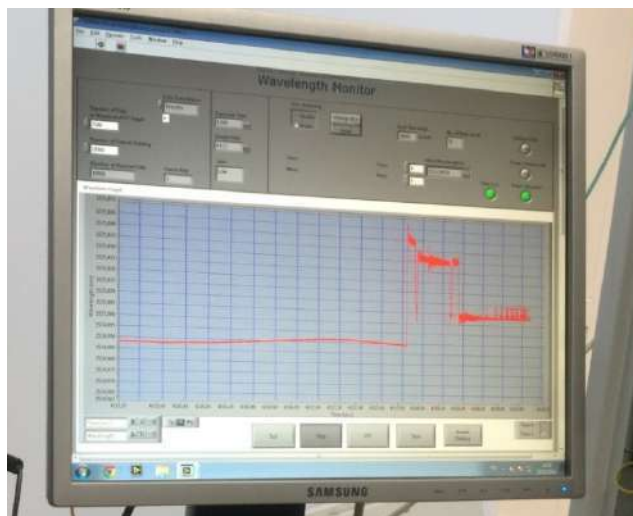


Fig. 6. Interrogator software photo.

## 5. Conclusion

During the analysis of analogs optical fiber current sensors the problems were educed: sensitivity to EM fields, amplitude separation of channels, sensitivity to acoustical influences, low correlation signal/noise. This device is suggested as a prototype which consists optical fiber with Bragg grating as a sensitive element, electromagnet for transformation electric energy into mechanic energy, optical fiber as a transfer channel and spectral analyzer with digital output for connecting to the PC. This device is free from the previously described problems. The laboratory test showed that resolution capability and sensitivity has quite high values and let use this types of sensors in the various measurement range of parameters.

## References

- [1] Vasil'yev SA, Medvedkov OI, Korolev IG, Bozhkov AS, Kurkov AS, Dianov YeM. Fiber gratings and their applications. *Quantum Electronics* 2005; 35(12): 1085–1103.
- [2] Othonos A. Fiber Bragg gratings. *Review of scientific instruments* 1997; 68(12): 4309–4341.
- [3] Medvedkov OI, Korolev IG, Vasil'yev SA. Recording of fiber Bragg gratings in a circuit with an LLOYD interferometer and modeling their spectral properties. Moscow: Fiber Optics Research Center of the RAS, 2004; 46 p. [in Russian]
- [4] Lazarev VA. A fast-acting deflection and temperature measurement system based on fiber-optic Bragg sensors. Moscow: Bauman Moscow State Technical University, 2013; 185 p. [in Russian]
- [5] Okosi T. Fiber optic sensors. Leningrad: Energoatomizdat, 1991; 256 p. [in Russian]
- [6] Gordon AV, Slivinskaya AG. Direct current solenoids. Moscow: Gosenergoizdat, 1960; 447 p. [in Russian]

# Optimization of chemical reactions by economic criteria based on kinetics of the process

K.F. Koledina<sup>1,2</sup>, S.N. Koledin<sup>1,2</sup>, I.M. Gubaydullin<sup>1,2</sup>

<sup>1</sup>*Institute of Petrochemistry and Catalysis, Russian Academy of Sciences, Prospect Oktyabrya 141, 450075, Ufa, Russia*

<sup>2</sup>*Ufa State Petroleum Technological University, Kosmonavtov St. 1, 450062, Ufa, Russia*

---

## Abstract

The paper deals with the formulation and solution of the inverse kinetic problem, methods of chemical reactions optimization by economic criteria on the basis of a process kinetic model. Yield of a target product, productivity, profit and productivity are considered as indicators.

*Keywords:* dimethylcarbonate; kinetic model; theoretical optimization; economic criteria; Hooke-Jeeves method

---

## 1. Introduction

"Green chemistry" is a scientific direction in chemistry, which can include any improvement in chemical processes that positively affects the environment. "Green chemistry" involves the use of low-toxic and non-toxic initial reagents.

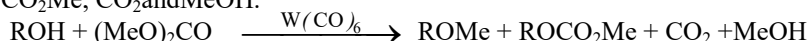
Twelve principles of green chemistry should be noted, which were developed by scientists Anastas P. and Warner J. [1] which are used by scientists:

1. It is better to prevent waste than to treat or clean up waste product after it is formed.
2. Synthesis methods should be designed to maximize the incorporation of all materials used in the process into the final product.
3. Wherever practicable, methodologies of synthesis should be designed to use and generate substances that possess little or no toxicity to human health and the environment.
4. Chemical products should be designed to preserve their efficiency and usage while reducing toxicity.
5. Better to not use at all auxiliary substances (e.g. solvents, separation agents, etc.) if there are any, they must be innocuous when used.
6. Energy requirements should be recognized for their environmental and economic impacts and should be minimized. Synthesis methods should be conducted at ambient temperature and pressure.
7. A raw material or feedstock should be renewable rather than depleting wherever technically and economically practicable.
8. Reduce amount of receiving intermediate products whenever its possible (blocking group, protection / deprotection, temporary modification).
9. Catalytic reagents (as selective as possible) are better than stoichiometric reagents.
10. Chemical products should be designed so that at the end of their function they do not persist in the environment and break down into innocuous products.
11. Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances.
12. Substances and the form of a substances used in a chemical process should be chosen to minimize potential for chemical accidents, including releases, explosions, and fires.

The reaction of alcohols with dimethyl carbonate (DMC) meets many of these principles. DMC is an effective substitute for toxic methyl halides (MeX, X = I, Br, Cl) and phosgene (3<sup>th</sup>, 4<sup>th</sup> principles). It is produced with CO<sub>2</sub> as initial reagent. For reactions, involving DMC, only a catalytic amount of transition metal complexes is required, resulting in no waste (10<sup>th</sup> principle). According to the literature, the reactivity of DMC is moderate [2-6]. In order to reduce the energy costs of the reaction (6<sup>th</sup> principle), catalysts are used. As a result, alkylmethyl esters of alcohols and alkylmethyl carbonates are formed with the process selectivity of 95-98% (2<sup>th</sup>, 6<sup>th</sup>, 9<sup>th</sup> principles).

The reaction is new, it is implemented only in the laboratory. To study its mechanism, it is necessary to develop a mathematical model and, based on the results of the constructed mathematical model, to carry out optimization.

The reaction of DMC with alcohol in the presence of the catalyst W (CO) 6 leads to the formation of four products: ROME, ROCO<sub>2</sub>Me, CO<sub>2</sub> and MeOH.



## 2. Mathematical model

A mathematical model of chemical kinetics is a system of nonlinear ordinary differential equations (SNODE) with the initial given data, i.e. the Cauchy problem (1) [7].

$$\frac{dx_i}{dt} = \sum_{j=1}^J v_{ij} w_j, i=1, \dots, I, \quad (1)$$

$$w_j = k_j^0 \cdot \exp\left(-\frac{E_j^+}{RT}\right) \cdot \prod_{i=1}^I (x_i)^{|\alpha_{ij}|} - k_{-j}^0 \cdot \exp\left(-\frac{E_j^-}{RT}\right) \cdot \prod_{i=1}^I (x_i)^{|\beta_{ij}|}$$

initial conditions: at  $t = 0, x_i(0) = x_i^0$ ;

where  $v_{ij}$  is stoichiometric coefficients;  $J$  is number of stages;  $x_i$  is concentration of substances participating in the reaction, mol/l;  $I$  is number of substances;  $w_j$  is  $j$ -th speed stage, 1 / min;  $k_j, k_{-j}$  is rate constants of direct and reverse reactions;  $E_{j+}, E_{j-}$  is activation energies of direct and reverse reactions, kJ/mol;  $R$  is universal gas constant, equal to 8.31 kJ / (mol \* K);  $T$  is temperature, K;  $\alpha_{ij}$  is negative elements of the matrix ( $v_{ij}$ );  $\beta_{ij}$  is positive elements ( $v_{ij}$ );  $k_j^0, k_{-j}^0$  is pre-exponential factors, 1 / min.

The solution of the direct problem is the SNODE solution with initial data and given kinetic parameters up to some fixed time  $t^*$ .

Such SNODE tasks of chemical kinetics are mostly stiff systems. Therefore, for their numerical solution Implicit Rosenbrock third-fourth order Runge-Kutta method is used with degree three interpolating the Maple [8] and the multi-step Gir method of variable order in Matlab.

The inverse problem is the determination of the kinetic parameters by matching the calculated kinetic curves with the experimental ones by functional (2).

$$\sum_{q=1}^Q \sum_{i=1}^I |x_{pi}^r - x_{pi}^e| \quad (2)$$

where  $x_{pi}^{\text{exp}}$  and  $x_{pi}^{\text{calc}}$  is experimental and calculated concentrations of components;  $I$  is number of substances;  $Q$  is number of measuring points.

The inverse problem was solved in the Matlab environment using the genetic algorithm and the Hook-Jeeves direct search method [9].

### 3. Results and Discussion

Fig.1. shows kinetic model for catalyst reaction of DMC with alcohols in the presence of  $W(CO)_6$ .

On the graphs of Fig.2 are shown corresponding kinetic curves and correspondence between the calculated values and the experimental data. Based on the developed kinetic model, a theoretical optimization of the process was carried out. A distinctive feature of the work is application of economic criteria at the level of laboratory experiments.

In general, the criterion for optimizing the chemical process has the form (3) [10-11]:

$$R(x, x^0, t^*, \eta, \mu, T) \rightarrow \max \quad (3)$$

Where  $R$  is optimization function,  $\eta$  is vector of the substance weights,  $\mu$  is additional cost.

For chemical reactions performed in the laboratory optimization, the following indicators of the process economic efficiency can be used as optimization criteria:

1) Productivity - the volume of output per unit time (4).

$$R: B = N \cdot C_{x_i^0} \cdot \xi_{x_i^0}(t^*, T) \cdot M_{x_i} \rightarrow \max \quad (4)$$

where  $B$  is process productivity [g / (mol \* day)];  $N$  is number of cycles per day [day<sup>-1</sup>];  $C_{x_i^0}$  is initial value of initial reagent [mole fractions];  $\xi_{x_i^0}$  is conversion of initial reagent;  $M_{x_i^0}$  is molar mass of the initial reagent [g / mol].

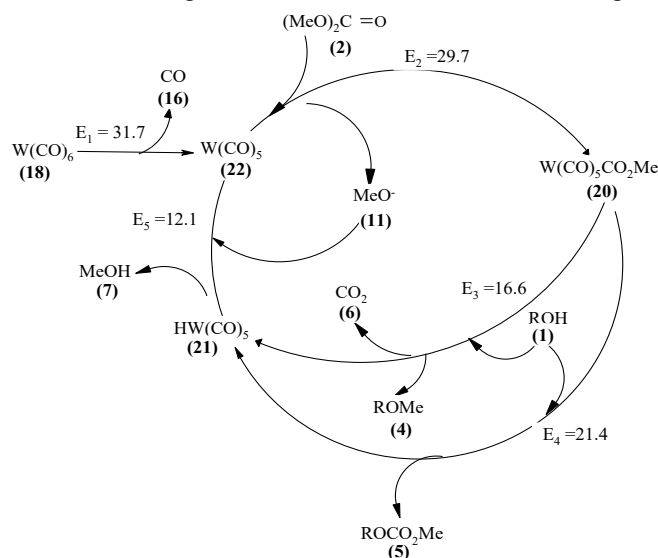


Fig. 1. Kinetic model of reaction with alcohols in the presence of tungsten hexacarbonyl.

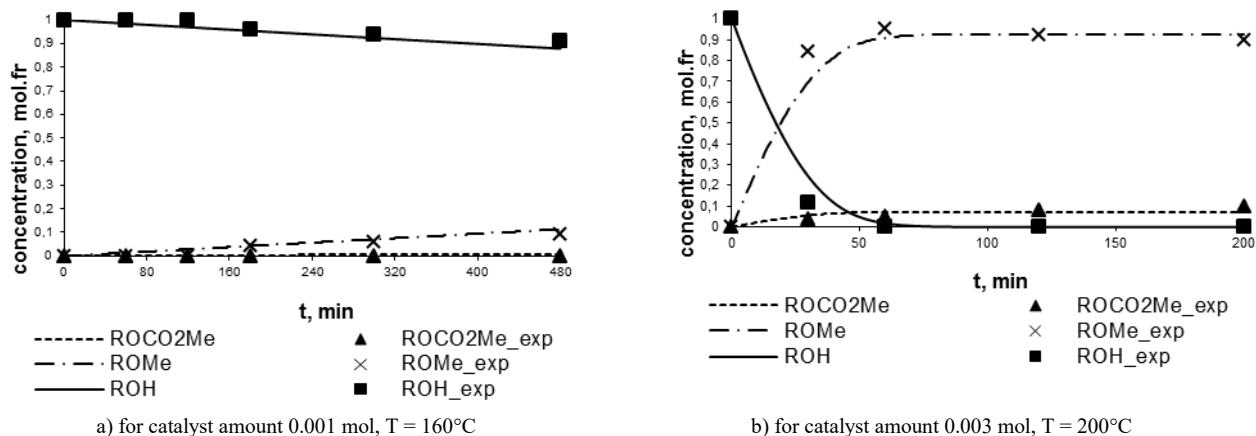


Fig.2. Graph of correspondence of experimental data (points) and calculated values (lines) of observed substrates concentration changing according to the scheme of chemical transformations in the presence of  $W(CO)_6$ .

2) Profit. The profit depends on difference between price of product and its cost, as well as on output (5).

$$R : E = \sum_{prod=1}^{Pr} x_{prod}(t^*, T) \cdot \eta_{prod} - \sum_{source=1}^{Sr} x_{source}(t^*, T) \cdot \eta_{source} - \psi(t^*, T) - A \rightarrow \max \quad (5)$$

where  $x_{prod}$  is concentration of reaction products;  $x_{source}$  is concentration of initial reagents;  $\eta$  is weight of components (normalized);  $\psi(t)$  is variable costs (normalized);  $A$  is constant costs (normalized);  $Pr$  is number of products;  $Sr$  is number of initial reagents;  $E$  is normalized profit.

3) Profitability. Profitability is defined as the ratio of the amount of profit to the volume of investment (6).

$$R : E = \frac{\sum_{prod=1}^{Pr} x_{prod}(t^*, T) \cdot \eta_{prod}}{\sum_{source=1}^{Sr} x_{source}(t^*, T) \cdot \eta_{source} + \psi(t^*, T) + A} \rightarrow \max \quad (6)$$

where  $P$  is normalized profitability.

Assuming an idle time between cycles of 1 hour (60 min.), the values of maximum process productivity and productivity at maximum yield of target product X5 are given in Table 1 for reaction with tungsten hexacarbonyl catalyst by (4).

Table 1. Indicators of DMC conversion, time, productivity under the condition of maximum product yield and maximum productivity in the temperature range.

T, °C	tungsten hexacarbonyl				
	maximum productivity			maximum yield ROME ( $\xi_{X_2}=0,26$ )	
	conversion (MeO) <sub>2</sub> CO	Reaction time, min	B, g/(mol * day)	Reaction time, min	B, g/(mol * day)
160	0,180	220	721	440	583
180	0,220	80	1763	140	1459
200	0,235	33	2836	50	2652
220	0,250	18	3597	22	3558

Thus, it can be seen that with DMC maximum possible conversion, an economically optimal solution is not achieved, because smaller conversion value corresponds to shorter reaction time, which leads to an increase in overall productivity.

The value of profit is determined by (5). The weights of substances depend on the cost of reagents on the market according to <http://www.acros.com> and <http://www.sigmaaldrich.com>. Then the graph of profit changing from time for different values of temperature has the form (Fig. 3).

Figure 3 shows the following patterns:

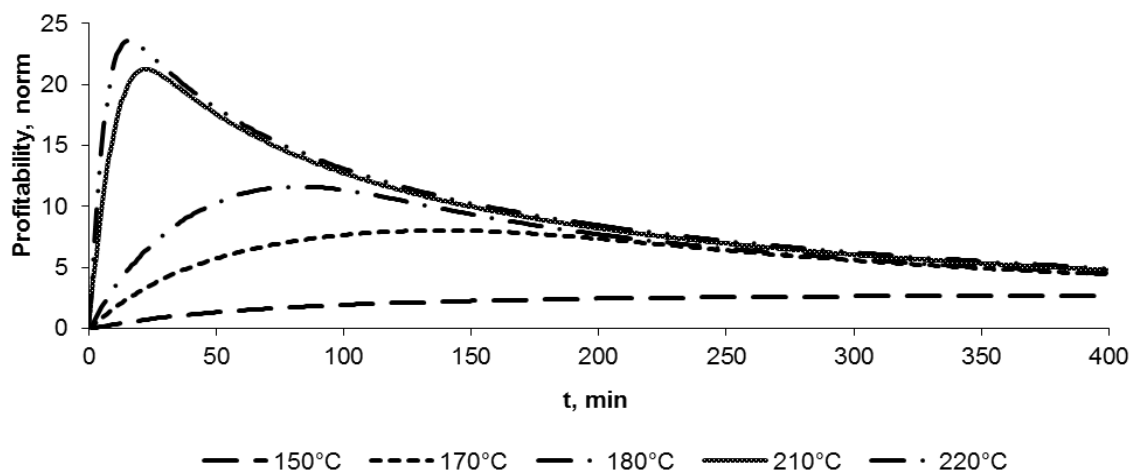
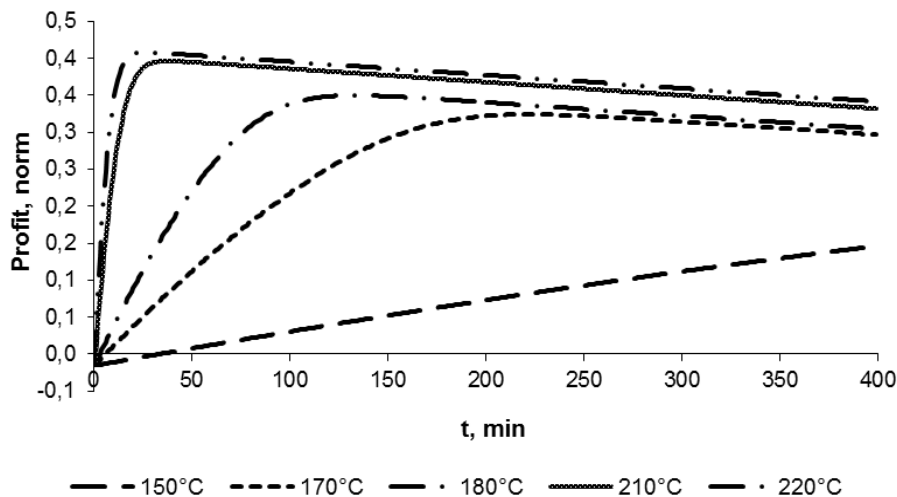
a) There is a temperature range with positive profits, and if the temperature is higher, maximum profit is attained earlier (170-220°C). It is worth noting that the decrease in the value of profit occurs more sharply, if the temperature is higher. This is probably because at a high temperature the reaction conversion occurs faster. When one of the reagents (alcohol) is completely consumed, value of the second reagent conversion cannot be increased and profit decreases due to variable costs (for example, to maintain a temperature).

b) At a temperature of 150°C, the profit value goes to negative area, but you can also see maximum profit (or minimum loss).

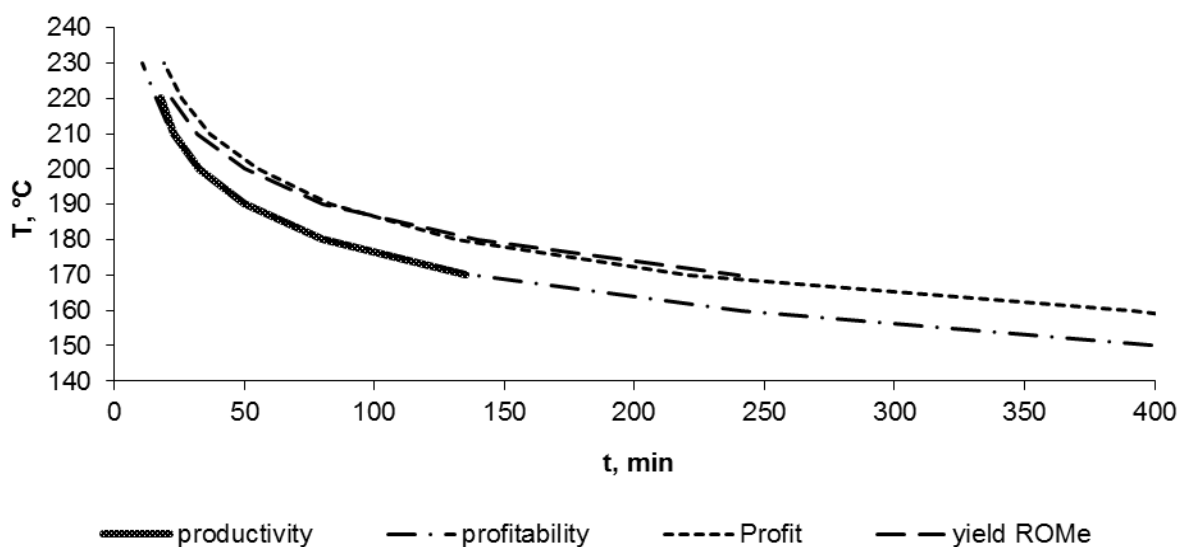
Profitability changing in the process from time is shown in Figure 4.

Figure 4 shows that the profitability reaches a maximum and decreases with time, which is explained by the costs of maintaining the set temperature. With increasing temperature, a maximum value of profitability comes earlier.





Each curve for the profit margin, as well as the productivity and yield of the target product passes through a maximum. From the points of maximums, we construct the optimum temperature profile, putting in correspondence to each instant of time the temperature at which the maximum of the target criterion is reached. Then the optimum temperature profile for all the indicators will take the form (Fig. 5).



The above reaction temperature profiles characterize the optimal reaction conditions (temperature and reaction time). Based on the results of the work, the overall temperature profile is given for all targets.

The overall temperature profile (Figure 5) for tungsten catalysts clearly demonstrates high correlation between productivity and profitability, as well as between yield and profit.

#### 4. Conclusion

The catalytic reaction of alcohols with DMC in the presence of tungsten hexacarbonyl is considered in this work. The formulation and solution of the inverse kinetic problem, the method of optimization of chemical reactions by economic criteria based on the kinetic model of the process are given. As indicators, the yield of the target product, productivity, profit and profitability are considered. For the reaction of alcohols with DMC in the presence of  $W(CO)_6$ , correlations are observed between productivity and profitability, as well as between yield and profit.

#### Acknowledgements

The work was supported by Russian Foundation for Basic Research N 15-07-01764 A.

#### References

- [1] Anastas PT, Warner J C. Green Chemistry: Theory and Practice. New York : Oxford University Press, 1998; 30 p.
- [2] Tundo P. New developments in dimethylcarbonate chemistry. Pure and Applied Chemistry 2001; 73: 1117–1120.
- [3] Khusnutdinov RI, Schadneva NA, Maykova YuYu. Synthesis of alkyl methyl ethers and alkyl methyl carbonates by reaction of alcohols with dimethyl carbonate in the presence of tungsten and cobalt complexes. Russian Journal of Organic Chemistry 2014; 50(6): 790–795.
- [4] Koledina KF, Koledin SN, Schadneva NA, Gubaidullin IM. Kinetics and mechanism of the catalytic reaction between alcohols and dimethyl carbonate. Russian Journal of Physical Chemistry A 2017; 91(3): 442–447.
- [5] Koledin SN, Koledina KF, Gubaidullin IM. Kinetic model of catalytic interaction alcohols with dimethyl carbonate in the presence of various catalysts. Materials V all-russian scientific-practical conference timed to the 110th anniversary of the birth of Academician A.N. Tikhonov, 17-19 november 2016. Sterlitamak. Part I. Sterlitamak: Sterlitamak branch of BashGU, 2016: 159–165. [in Russian]
- [6] Koledin SN, Koledina KF, Gubaidullin IM. Kinetics of reaction in the study economic efficiency in chemical production. Mathematical modeling of processes and systems. Materials III all-russian scientific-practical conference with international participation Sterlitamak, 4-6 december, 2014: 30–32. [in Russian]
- [7] Koledina KF, Koledin SN, Gubaidullin IM, Safin RR, Ahmetov IV. Information system for constructing a kinetic model of a catalytic reaction, planning an economically optimal chemical experiment. Control Systems and Information Technology 2015; 3(61): 79–84. [in Russian]
- [8] Koledina KF, Gubaidullin IM. Kinetics and mechanism of olefin catalytic hydroalumination by organoaluminum compounds. Russian Journal of Physical Chemistry A 2016; 90(5): 914–921.
- [9] Hook P, Jeeves TA. Direct solution search for numeric and static problems. M.: Mir, 1961; 219 p. [in Russian]
- [10] Koledin SN, Koledina KF, Gubaidullin IM, Spivak SI. Determination of optimal conditions for catalytic processes based on economic criteria. Chemical industry today 2016; 10: 24–35. [in Russian]
- [11] Koledin SN, Koledina KF. Optimal control and sensitivity of the optimum in problems of chemical kinetics. Journal of Middle Volga Math. Soc. 2016; 18(3): 137–144.

# Ab initio modeling of optical properties of the new $sp^3$ silicon and germanium allotropes

V.A. Saleev<sup>1</sup>, A.V. Shipilova<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The application of the hybrid topological-quantum-mechanical method to the search of new allotropes of 14th group elements is demonstrated for silicon and germanium. Starting from the databases of hypothetical and real zeolite nets and subsequently applying the geometrical and energetic selection criteria, we extract the most energetically favourable structures for the allotropic modifications of silicon and germanium, and study their optical properties. In the framework of density functional theory we calculate the frequency-dependent complex dielectric tensors, refraction and absorption coefficients of the selected allotropes and their electronic band gaps.

*Keywords:* crystal structure design; photonics; density functional theory; silicon and germanium optic properties.

---

## 1. Introduction

The elements of XIVth group of the periodic table (carbon, silicon and germanium) are widely used in the modern electronics, photonics and photovoltaics, as functional elements of the different electronic and optical devices. Of particular importance are their optical properties in the sense of absorption and refraction of the incident electromagnetic radiation of different wavelength ranges: infrared (IR), visible and ultraviolet (UV). The properties of carbon, silicon and germanium crystals with the diamond crystal lattice (the space group  $Fd\bar{3}m$ ) have been studied quite well, both experimentally [1] and theoretically [2]. One of the tasks of theoretical material science based on quantum ab initio calculations is the search and prediction of new allotropic modifications of carbon, silicon and germanium, which would have a set of properties exceeding diamond structures. These new structures are of interest in the practical applications to solve various problems of micro- and nanoelectronics, photonics and photovoltaics.

## 2. Computer design of new allotropes

There are four main steps in this theoretical study: the search for new crystalline structures, the verification of their stability under normal or specified conditions, the calculation of the basic physical properties, and the search for the ways of their synthesis. In our work we use a new method of search for new allotropic modifications, based on topological analysis of the modern bases of hypothetical zeolite nets [3,4]. For the first time this method was used in our work [5] to search for the carbon allotropes with  $sp^3$ -hybridization of chemical bonds, close in energy and other physical properties to diamond. We analyzed more than 600 thousand zeolite nets, applying subsequently topological and geometric selection criteria. As an example of such a criterion, at one stage we excluded from consideration all structures containing 3 and 4-membered rings, since the presence of such chains of bonds in carbon allotropes leads to a significant internal stress in the structure. Then we relaxed the positions of atoms in the structures using quantum-mechanical modeling packages and selected those structures in which the  $sp^3$ -hybridization of chemical bonds was preserved. For them, we calculated the binding energies of the ground state and compared them with the energy of the diamond configuration, selecting 6 most advantageous structures, with binding energy per atom not more than 0.12 eV compared to diamond. Taking into account the chemical relation of carbon, silicon and germanium, it is obvious to assume that the same allotropic modifications (from the point of view of crystallography, that is, with the same symmetry group and topology) exist for silicon and germanium. In Table 1 we give the chemical bond lengths, the angles between them, the distances to the nearest neighbors for the predicted allotropes of silicon and germanium.

Table 1. Chemical bond lengths, bond angles, the distances to the nearest neighbors for the predicted allotropes of silicon and germanium. Definitions are the same as in the work [5].

### Silicon

Structure	Bond lengths, Å	Bond angles, degrees	Nearest-neighbor distances, Å
#26	2.308-2.382	95.98-126.34	3.818
#27	2.308-2.415	95.31-125.84	3.820
#28	2.308-2.406	93.14-124.12	3.818
#50	2.301-2.386	96.95-127.46	3.826
#55	2.322-2.422	97.91-120.04	3.828
#88	2.322-2.380	98.16-119.32	3.837

---

**Germanium**

Structure	Bond lengths, Å	Bond angles, degrees	Nearest-neighbor distances, Å
#26	2.436-2.562	94.76-126.16	4.055
#27	2.436-2.533	94.76-126.16	4.055
#28	2.447-2.539	92.88-120.79	4.057
#50	2.439-2.530	96.29-122.11	4.067
#55	2.451-2.547	98.65-119.12	4.058
#88	2.459-2.514	97.51-120.61	4.068

**3. Calculation methods**

The recent development of methods for quantum mechanical calculations of energy and the electron density distribution of many-electron systems (atoms, molecules, crystals) is associated with the success of density functional theory (DFT) [6,7,8] and combined or hybrid methods based both on DFT and the Hartree-Fock (HF) approach [9], which take into account the exchange interaction more accurately than the DFT. Quantum mechanical calculations of the physical properties of crystals require a significant computational resources, in comparison with atomic or molecular calculations, and can be performed only on supercomputers or multiprocessor cluster systems. In our work, we use the most common and widely used licensed software packages CRYSTAL [10] and VASP [11], installed on the supercomputer "Sergey Korolev" of the Samara University. The CRYSTAL software package uses a basis of atomic orbitals and an all-electron approximation, while the VASP package uses a basic set of plane waves and a pseudopotential approximation.

To check the stability of the predicted silicon and germanium allotropes, the matrices of elastic constants and various elastic coefficients were calculated, and we showed the structures to be mechanically (energetically) stable at zero external pressure (see Fig. Tables 2 and 3). We also demonstrated the absence of imaginary frequencies in the phonon spectra of studied silicon and germanium allotropes. As an example, Fig. 1 shows the calculated phonon spectrum for the allotrope Si#50. All calculations were performed in the generalized gradient approximation (GGA) of the density functional theory with the exchange-correlation functional PBE [12].

Table 2. Density, binding energy difference, band gap, band gap in Gamma point, bulk modulus, bulk modulus derivative, shear modulus and static dielectric constants for silicon allotropes.

GGA PBE	88,Pnma	50,Pnma	55,Pmma	26,P2/m	27,C2/m	28,Pbam
$\rho$ g/cm <sup>3</sup>	2.293	2.291	2.305	2.298	2.291	2.271
$\Delta E$ , eV	0.037	0.036	0.035	0.039	0.047	0.036
$E_{\text{gap}}$ , eV	1.40	1.42	1.26	0.97	1.02	1.31
$\Delta E_{\text{gap}}(\Gamma-\Gamma)$ , eV	1.56	1.51	1.58	1.10	1.12	1.57
B, GPa	78	84	83	84	84	86
B'	2.74	4.31	3.75	4.14	4.35	4.64
G, GPa	48	48	48	50	51	51
$\epsilon_{xx}$ ,	11.88,	11.92,	11.74,	11.85,	12.74,	12.55,
$\epsilon_{yy}$ ,	12.05,	12.05,	11.40,	11.78,	11.94,	11.51,
$\epsilon_{zz}$	12.09	12.52	11.87	12.18	12.07	11.93

Table 3. Density, binding energy difference, band gap, band gap in the Gamma point, bulk modulus, bulk modulus derivative, shear modulus and static dielectric constants for germanium allotropes.

GGA PBE	88,Pnma	50,Pnma	55,Pmma	26,P2/m	27,C2/m	28,Pbam
$\rho$ , g/cm <sup>3</sup>	5.251	5.086	5.121	5.102	5.087	5.054
$\Delta E$ , eV	0.037	0.038	0.036	0.044	0.052	0.033
$E_{\text{gap}}$ , eV	0.87	0.86	0.61	0.42	0.23	0.65
$\Delta E_{\text{gap}}(\Gamma-\Gamma)$ , eV	0.95	0.96	1.08	0.42	0.60	0.65
B, GPa	57	59	59	59	59	60
B'	4.24	4.54	4.38	4.44	4.53	4.61
G, GPa	43	43	44	45	45	46
$\epsilon_{xx}$ ,	14.87,	15.69,	14.77,	15.57,	16.74,	16.07,
$\epsilon_{yy}$ ,	16.47,	16.38,	15.31,	15.92,	17.77,	14.82,
$\epsilon_{zz}$	15.74	16.18	15.49	15.51	16.76	15.50

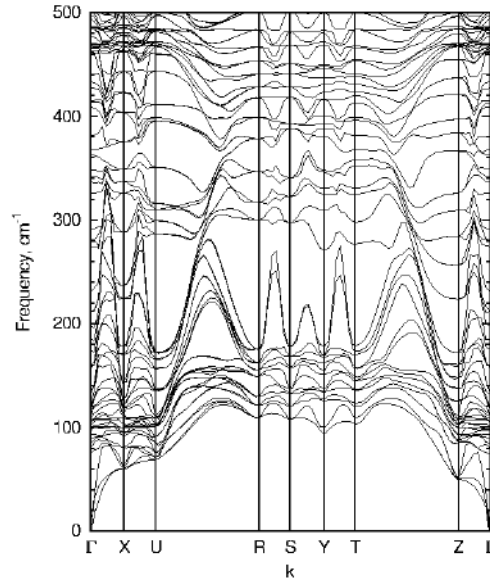


Fig.1. The calculated phonon spectrum of allotrope Si#50 along the high-symmetry pathway in the reciprocal space.

#### 4. The optical properties of allotropes

Methods for calculating the optical properties of crystals depend on the chosen electromagnetic wavelength range. The properties of the complex dielectric tensor in the infrared region are determined in the semi-classical Drude-Lorentz theory by singularities of the crystal lattice vibration spectrum, which can be calculated in the quasi-harmonic approximation [13] with the transverse and longitudinal optical vibration modes. The spectrum of eigenfrequencies of allotropes allows one to calculate their Raman scattering spectra and absorption spectra in the IR range. Investigation of Raman spectra can be used for experimental search for new allotropic modifications, since the position of the Raman peaks is uniquely related to the structure of the crystal lattice, the forces and lengths of the chemical bonds. Unlike the diamond modification of silicon, the Raman spectra of its allotropic modifications have a more complex structure, which makes it difficult to identify them experimentally. At the same time, the Raman spectra of germanium allotropes, like the Raman spectrum of the diamond modification of germanium, have only one strong peak, which is significantly shifted relative to the peak at a frequency of about  $300 \text{ sec}^{-1}$ , which is observed for the ground state of germanium, Fig. 2. It is well known that diamond and diamond-like crystals of silicon and germanium practically do not absorb in the IR range. The silicon and germanium allotropes predicted by us have narrow absorption bands in the IR range, which can also be used for their experimental search and identification, see Fig. 3.

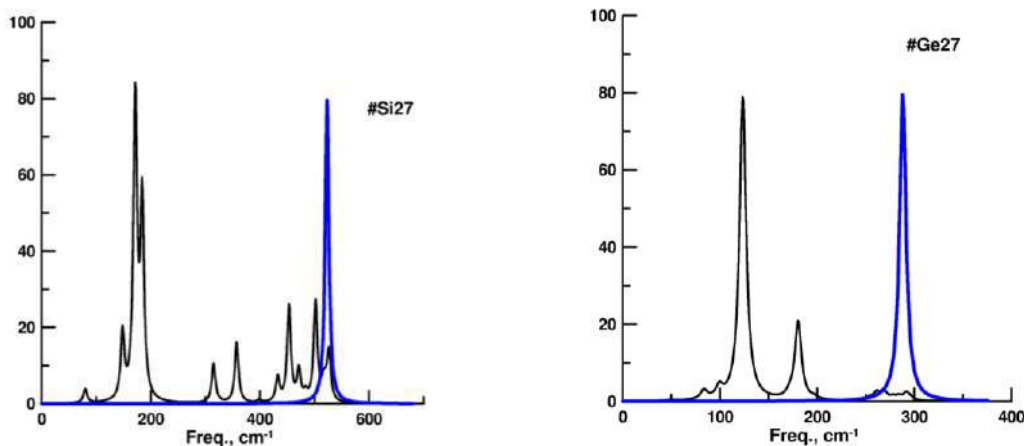


Fig. 2. Raman shift spectra for allotropes of silicon (Si#27) and germanium (Ge#27) – black curves, diamond-like silicon and germanium – gray curves.

The electronic band structure of a crystal determines the properties of its complex dielectric function, hence, the dependence of the absorption and refraction coefficients on the frequency of electromagnetic radiation in the visible and UV ranges. DFT allows to obtain a microscopic dielectric function in the random phase approximation within the theory of linear response, in the visible and ultraviolet frequency ranges. Neglecting the local field effects, we can derive from the microscopic dielectric function a macroscopic dielectric function. The imaginary part of the latter is a tensor and can be written in the form of a weighted sum over transitions between levels, and the real part is obtained from the imaginary one using the Kramers-Kronig relations. We can extract both the optical constants and the complex dielectric function averaged over the directions directly from the components of the macroscopic dielectric tensor, which allow us to determine the optical absorption and refraction spectra for the studied structures.

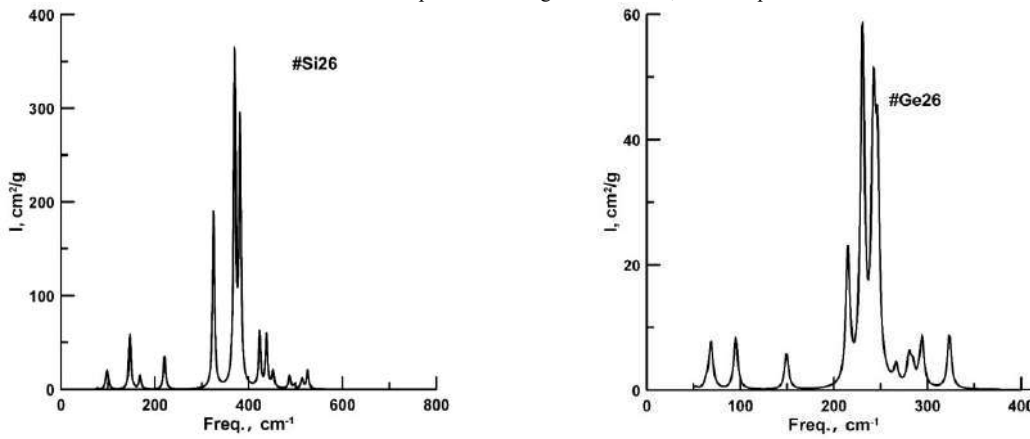


Fig. 3. Infrared absorption spectra for allotropes of silicon (Si#26) and germanium (Ge#26).

The electronic band gap of a semiconductor determines the energy boundary of absorption of optical photons. It is known that standard DFT methods do not allow to reproduce this gap correctly, in particular for materials with a narrow optical gap, such as germanium. Therefore, in our calculations we used the hybrid functional HSE06 [14], which allows to obtain results comparable to the experimental data. However, the standard functional PBE is more adequate in describing the position of the peaks of the complex dielectric function. Tables 2 and 3 show the calculated values of the band gaps (indirect and in the Gamma point of the reciprocal space) and the permittivity coefficients of silicon and germanium allotropes, respectively. We note that the small value of the band gap of the allotrope Ge#27, 0.23 eV, may indicate metallization of this structure at high temperatures and loss of semiconductor properties. In Fig. 4 we present an electronic band structure for the allotrope Si#27 and Ge#27.

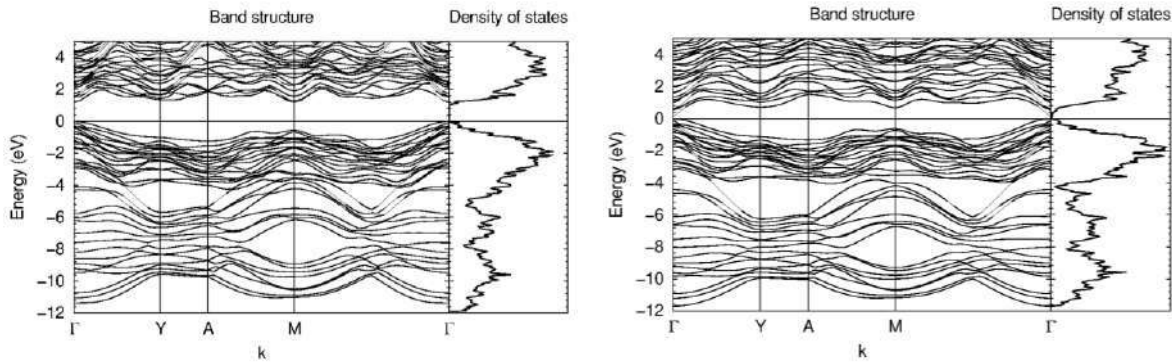


Fig.4. The electronic band structure of the allotropes Si#27 (left) and Ge#27 (right), calculated along the high-symmetry pathway in the reciprocal space, and the corresponding electron density of states.

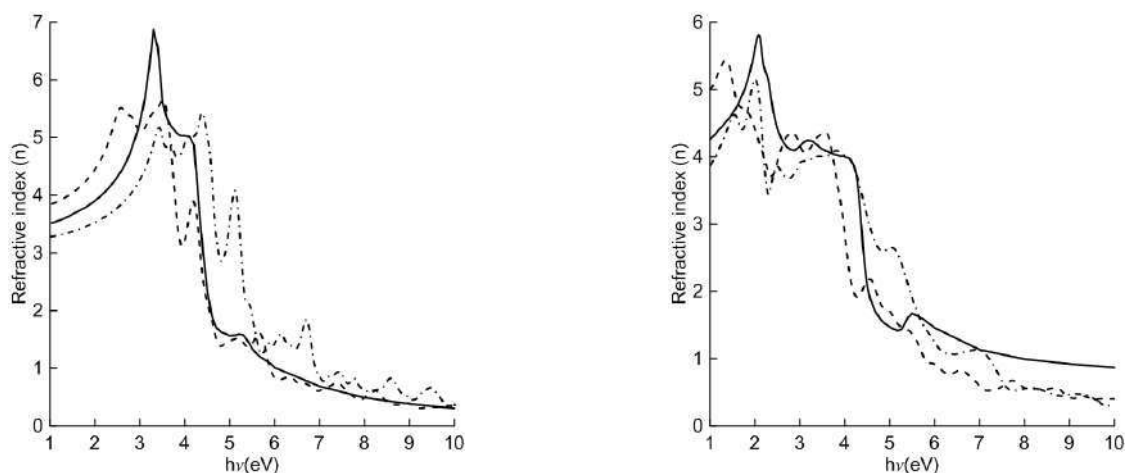


Fig. 5. The refractive index for diamond configurations of silicon (left) and germanium (right). Dash-dotted line – the results obtained for the functional HSE06, the dashed line – for the functional PBE, solid line – experimental data [1].

In Fig. 5 we show the calculated (dash-dotted line – HSE06, dashed – PBE) curves in comparison with the experimental data (black line) of the frequency-dependent refractive indices  $n$  for diamond configurations Si (left) and Ge (right). The absorption spectra ( $k$ ), together with the relative spectrum of solar irradiation at the reference air mass of 1.5, are shown in Fig. 6, respectively. Amorphous forms of Si and Ge, the so-called a-Si and a-Ge, as well as their various hydrogenated forms and some Si-Ge compounds [1] demonstrate promising properties for use in electronics and photovoltaics, especially in solar cells.

Comparing the results for our allotropes for the refraction and absorption coefficients with the corresponding spectra of amorphous forms, we observe quantitative and qualitative agreement both for the position of the refraction/absorption peak and for its absolute value (see Figures 7 and 8). This can be an evidence that the predicted by us allotropes can be a counterpart of the corresponding amorphous forms.

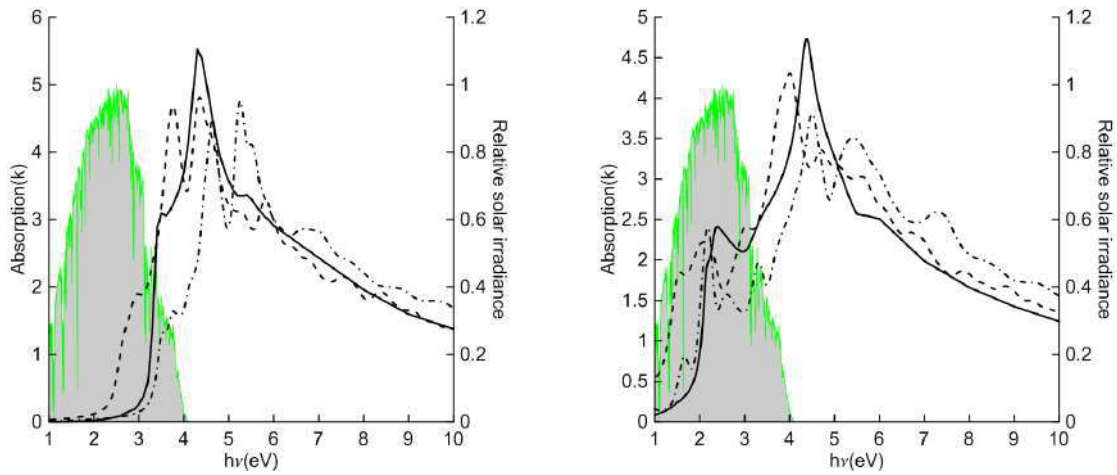


Fig. 6. The absorption index for diamond configurations of silicon (left) and germanium (right). Dash-dotted line – the results obtained for the functional HSE06, the dashed line – for the functional PBE, solid line – experimental data [1]. The shaded area is the relative spectrum of solar irradiance at the reference air mass of 1.5.

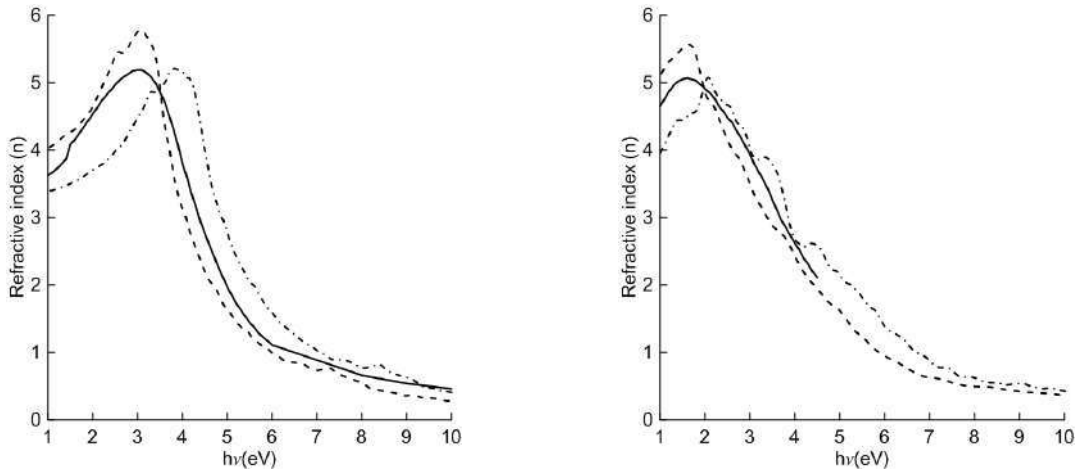


Fig. 7. The refractive index for allotropes #28 of silicon (left) and germanium (right). Dash-dotted line – the results obtained for the functional HSE06, the dashed line – for the functional PBE, solid line – experimental data for the amorphous forms [1].

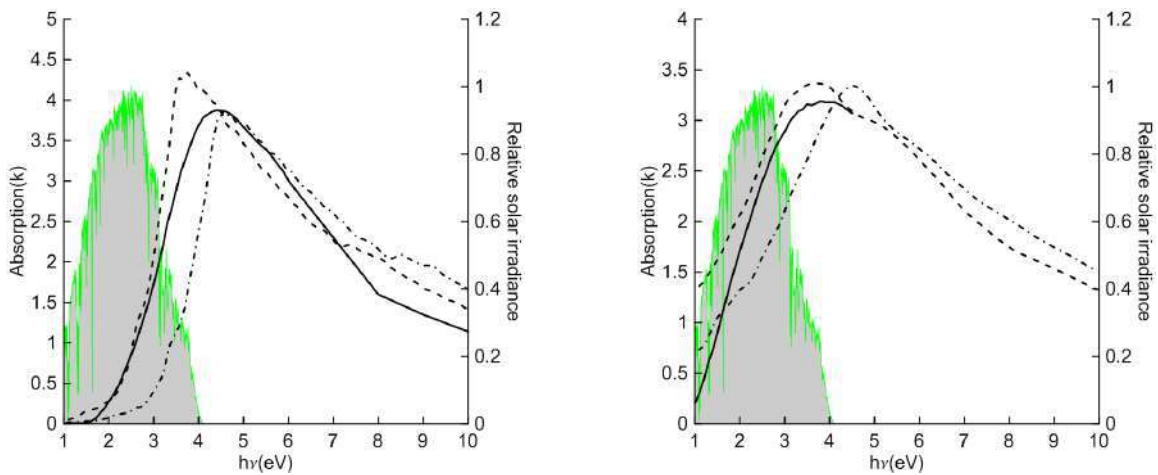


Fig. 8. The absorption index for allotropes of silicon (left) and germanium (right). Dash-dotted line – the results obtained for the functional HSE06, the dashed line – for the functional PBE, solid line – experimental data [1] for the amorphous forms. The shaded area is the relative spectrum of solar irradiance at the reference air mass of 1.5.

## 5. Conclusions

We investigated the six new low-energy allotropes of silicon and germanium [15], isostructural to the previously proposed carbon allotropes [5], demonstrating an application of the hybrid topology-quantum-mechanical approach [5] to prediction of new structures. Using the *ab initio* methods implemented in the CRYSTAL [10] and VASP [11] software packages, we calculated their mechanical, electronic and optical properties, which mostly resemble the properties of diamond configurations, but the observed differences in Raman shift spectra and infrared absorption spectra can allow to identify these allotropes if they are present in mixed phases. We have shown that the optical properties of allotropes under study are quantitatively and qualitatively close to the properties of amorphous modifications, a-Si and a-Ge. This leads to the conclusion that the considered allotropes are present in the experimentally observed amorphous phases, which are promising materials for electronics and photovoltaics.

## Acknowledgements

The authors thank the Ministry of Education and Science of the Russian Federation for financial support in the framework of the Samara University Competitiveness Improvement Program among the world's leading research and educational centers for 2013-2020, the task number 3.5093.2017/8.9.

## References

- [1] Adachi S. Optical constants of crystalline and amorphous semiconductors. New York: Springer Science + Business Media, 1999; 714 p.
- [2] Vivien L, Pavesi L. Handbook of Silicon Photonics. New York, London : CRC Press, 2013; 851 p.
- [3] Deem MW, Pophale R, Cheeseman PA, Earl DJ. Computational discovery of new zeolite-like materials. *J. Phys. Chem. C*. 2009; 113: 21353–21360. DOI: 10.1021/jp906984z.
- [4] Treacy MMJ, Randall KH, Rao S, Perry JA, Chadi DJ, Kristallogr Z. Enumeration of periodic tetrahedral frameworks. 1997; 212: 768–791. DOI: 10.1524/zkri.1997.212.11.768.
- [5] Baburin IA, Proserpio DM, Saleev VA, Shipilova AV. From zeolite nets to sp<sup>3</sup> carbon allotropes: a topology-based multiscale theoretical study. *Phys. Chem. Chem. Phys.* 2015; 17: 1332–1338. DOI: 10.1039/c4cp04569f.
- [6] Hohenberg P, Kohn W. Inhomogeneous Electron Gas. *Phys. Rev.* 1964; 136: B864. DOI: 10.1103/PhysRev.136.B864.
- [7] Kohn W, Sham LJ. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* 1965; 140: A1133. DOI: 10.1103/PhysRev.140.A1133.
- [8] Sholl D, Steckel JA. Density Functional Theory: A Practical Introduction. New York: Wiley, 2009; 252 p.
- [9] Fock VA. Fundamentals of quantum mechanics. Moscow: Nauka, 1976; 376 p.
- [10] Dovesi R. A program for the ab initio investigation of crystalline solids. *Int. J. Quantum Chem.* 2014; 114: 1287-1317. DOI: 10.1002/qua.24658.
- [11] Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*. 1996; 54: 11169–11186. DOI: 10.1103/PhysRevB.54.11169.
- [12] Perdew JP, Burke K, Ernzerhof M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* 1996; 77: 3865–3868. DOI: 10.1103/PhysRevLett.78.1396.
- [13] Pascale F, Zicovich-Wilson CM, Lopez F, Civalleri B, Orlando R, Dovesi R. The calculation of the vibration frequencies of crystalline compounds and its implementation in the CRYSTAL code. *J. Comput. Chem.* 2004; 25: 888. DOI: 10.1002/jcc.20019.
- [14] Heyd J, Scuseria GE, Ernzerhof M. Hybrid functionals on a screened Coulomb potential. *J. Chem. Phys.* 2006; 77: 219906. DOI: 10.1063/1.1564060.
- [15] Saleev VA, Shipilova AV, Fadda G, Proserpio DM. Prediction of the new sp<sup>3</sup> silicon and germanium allotropes from the topology-based multiscale method. Cornell University Library: <https://arxiv.org/abs/1701.04667>.



# Mathematical modeling of separation of watered oil-containing mixture

V.A. Zelenskiy<sup>1</sup>, A.A. Sushin<sup>1</sup>, A.I. Shchodro<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

A mathematical description of the separation process of the watered oil-containing mixture has been proposed. The mathematical model is based on the determination of the speed of movement of oil globules in a constrained flow. The formula for determining the velocity of the globules is derived from theoretical provisions and experimental data. The key issue of simulation is determining the separation time. This characteristic is calculated as a time period from the moment of arrival of the portion of oil-containing mixture to the first chamber of the separator to the moment of formation of the continuous phase layer of the given thickness on the surface. The mathematical model enables us to determine the separation time taking into account geometric characteristics of the separation device at the high level of water content. Thus, a topical problem of improving the performance of the separation device without sacrificing the quality of the commercial oil is solved.

**Keywords:** separation device; watered oil-containing mixture; oil globules; constrained flow; separation time

## 1. Introduction

The separation of the oil-containing mixture is one of the main technological processes in the oil treatment control system (OTCS). In order for the OTCS to function effectively, it is necessary to determine the optimal separation time. The duration of the separation process is directly related to the performance improvement of the separation device and the OTCS as a whole without sacrificing the quality of the commercial oil [1]. Furthermore, the separation time is the most important parameter for managing the process in automated mode [2]. As a rule, the requirements for performance and oil quality are mutually contradictory. Curing mixture in separators (clarifying tanks) reduces OTCS efficiency. On the other hand, the quality of the commercial oil should be monitored when reducing the time of separation. One of the most important indicators of oil quality is its residual water content. The measurement of water content in oil is performed according to GOST P 51858-2002. The third-group oil standard is under 1%. More strict requirements are applied to oil of the 1st and 2nd group. The residual water content of these groups shall not exceed 0.5%. The percentage of water content in the oil-containing mixture will be called watercut. The oil and gas industry in many regions (Tatarstan, Bashkortostan, Samara, Orenburg regions) operates wells with watercut varying from 70% to 90%. The profitability of such wells can be achieved only through equipment upgrade, employment of up-to-date technologies and IT. This should be preceded by the analysis of processes taking place in OTCS based on their mathematical models. As a result, the simulation of a separation process is a relevant scientific task.

## 2. Physical description of a separation process

The separation process of the oil-containing mixture takes place in a separation device, also called an "oil and gas separator" or simply a "separator" [2]. The fragment of separator (first chamber) is shown in Figure 1. The main elements are: gas pressure sensor (1), fluid-level sensor (2), partition wall between separator's chambers of  $h_p$  height (3), control device (4), temperature sensor (5), fluid pressure hydrostatic sensor (6), water-discharge valve (7), inlet valve for oil-containing mixture (8).

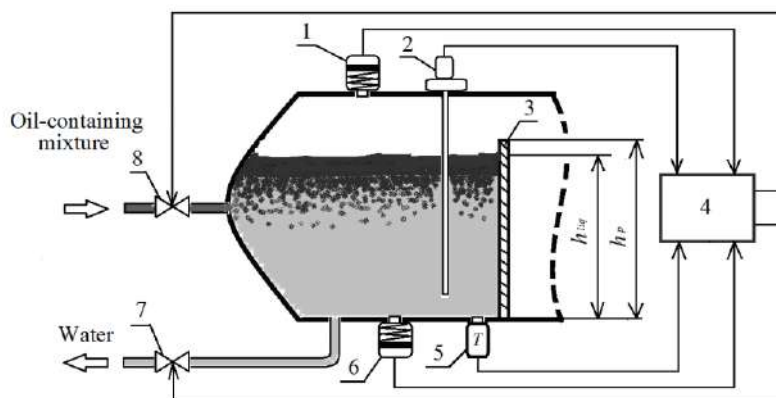


Fig. 1. The first chamber of separation device.

A complex mixture containing oil-associated gas, water, oil, metal salts and other impurities comes into the separator's chamber from an oil well through the inlet valve 8. The velocity of separation of the gas phase is much higher than fluid demulsification, so the influence of the gas factor can be neglected. After the chamber is filled up to the height of  $h_l$ , the gravity clarifying of oil-containing mixture takes place. As a result, a lighter oil globule comes to the surface. Relatively heavy globules of water are depositing to the bottom of the chamber and through the valve 7 get into the water treatment device and further back to the well. The separation process begins with the destruction of the globule membranes, which leads to their adhesion. Then the globules become larger as a result of coagulation. Finally, coalescence of globules results in the formation of continuous phases of water and oil. Sensors 1, 2, 5, 6 readings are used to generate control signals in control device 4. Thus, by

indirect measuring of density the mixture watercut value is obtained [3]. This information is used to determine the separation time [4]. Three approaches to obtain this parameter can be highlighted.

- 1) Separation time is a period of time from the moment the first chamber of the separator is filled (up to a given level, not exceeding the overflow level) until full decomposition of the oil-containing mixture. This approach did not find practical use as it does not take into account the real velocity of demulsification process.
- 2) Separation time is a period of time from the moment of arrival of the portion of oil-containing mixture to the first chamber of oil and gas separator to the moment of its decomposition to the state determined by GOST P 51858-2002. This approach is more progressive than the first one, but does not take into account the continuous nature of OTCS operation.
- 3) Separation time is a period of time from the moment of arrival of the portion of oil-containing mixture to the first chamber of the separator to the moment of formation of the continuous phase layer of the given thickness on the surface. This is the preferred method to be used in the framework of this research.

### 3. Mathematical description of separation process

There is no strict mathematical description of the stratification of the oil-containing mixture as Navier-Stokes equation is valid just for laminar flow and some special cases [5]. However, there are a large number of empirical and semi-empirical relationships that determine the nature of the processes in the stratification of emulsions that can be used as a basis [6, 7]. Under the conditions of strong watercut of oil mixture (70% ... 90%), oil is a dispersed phase and water is a dispersed medium. The oil particle rising to the surface experiences the difference between the gravity force and the lifting Archimedes' force [7].

$$\Delta F = \frac{\pi g}{6} d^3 \Delta \rho,$$

where  $\Delta \rho$  is the difference between dispersed phase and dispersed medium particles' density,  $g$  is the gravitational acceleration,  $d$  is the particle's diameter. Resistance force of the continuous medium:

$$F_c = \xi_o \frac{\pi d^2}{4} \frac{\omega_o}{2} \rho_c,$$

where  $\xi_o$  is the coefficient of hydraulic resistance of the continuous medium to the movement of a single particle in it,  $\omega_o$  is the velocity of a single particle relative to the medium,  $\rho_c$  is the density of the continuous medium. Let us assume that the temperature at all points of a separation device chamber is the same. Then there is no thermal convection. If the particle's velocity in the medium is constant:  $\Delta F = F_c$  the Reynolds criterion is determined by the following ratio:

$$Re_o = \frac{\omega_o d \rho_c}{\mu_c},$$

where  $\mu_c$  is the dynamic viscosity of the continuous medium. Archimedes' criterion is as follows:

$$Ar = \frac{d^3 g \rho_d - \rho_c}{\nu_c^2 \rho_c},$$

where  $\nu_c$  is the kinematic viscosity of continuous medium,  $\rho_d$  is the density of the dispersed phase. Taking these criteria into account, we can derive the following equality:

$$\xi_o Re_o^2 = \frac{4}{3} Ar.$$

Under the conditions of constrained surfacing, which is characterized by the interaction between particles, the following equality is valid:

$$\xi_d Re_d^2 = \frac{4}{3} Ar,$$

where  $\xi_d$  is the coefficient of hydraulic resistance for the dispersed phase in the emulsion,  $Re_d$  is the Reynolds criterion under the conditions of constrained flow. Hence the following formula is obtained:

$$\xi_d Re_d^2 = \xi_o Re_o^2.$$

Let us assume that

$$\xi_d = f(\varphi) \xi_{o\varphi},$$

where  $\xi_{o\varphi}$  is the coefficient of hydraulic resistance for a continuous medium for a single particle under the constrained flow conditions,  $\varphi$  is the volume fraction of the dispersed phase within the system. It would be useful to define the type of function  $f(\varphi)$  for the small and large Reynolds criterion values. Experimental studies showed that the velocities of particles depositing are connected by the following relation:

$$\omega_{o\varphi} = \omega_o (1 - \varphi)^n,$$

where  $\omega_{o\varphi}$  is the depositing rate of the particle relative to the continuous medium in constrained flow conditions,  $\omega_o$  is the rate of free depositing of the particle, and  $n$  index is to be determined. Using the obtained parameter called the volume fraction of the dispersed phase, we have:

$$Re_d = (1 - \varphi) Re_o.$$

It is experimentally shown [7] that with small values of Reynolds criterion ( $Re < 500$ ) the hydraulic resistance coefficient of the medium equals:

$$\zeta_d = \frac{24(1 + 0,15 Re_0^{0.687})}{0.843 \lg(\theta / 0.065) Re}$$

where  $\theta$  is the coefficient of the particle shape, equal to the ratio of the surface area of the spherical particle to the surface area of the real particle of the same volume. With small values of Reynolds coefficient, we can assume that:

$$f(\varphi) \approx (1 - \varphi)^{-n}$$

For large values of Reynolds coefficient  $Re$  the following expression is valid:

$$f(\varphi) \approx (1 - \varphi)^{-2n}$$

It is experimentally shown [7] that the function  $f(\varphi)$ , both in case of large and small values of  $Re$  varies from  $(1 - \varphi)^{-4.65}$  to  $(1 - \varphi)^{-4.78}$ . Then we can take an average index and write:

$$f(\varphi) = (1 - \varphi)^{-4.72}$$

Taking into account the expressions derived, it is obtained that the ratio of the particle depositing rate relative to the continuous medium under the constrained flow conditions to the particle's free depositing rate is equal to:

$$\omega_{od} / \omega_0 = (1 - \varphi)^{-4.72}$$

There exist empirical formulas which enable us to account for the influence of constraining [7]. For example, when  $\varphi < 0.3$ , the following formula is applied:

$$\omega_{od} / \omega_0 = (1 - \varphi)^2 10^{-1.82\varphi}$$

For  $\varphi > 0.3$  the following formula is applied:

$$\omega_{od} / \omega_0 = \frac{0.123}{\varphi} (1 - \varphi)^3$$

Table 1 shows comparative data of the calculation of the velocity of the oil globules rising for the known relations and calculated according to the obtained formula. Numerical values are reduced to the velocity of oil globules freely rising to the surface, i.e. relative ones.

Table 1. Calculation results for the velocity of oil globules rising according to known formulas and formulas obtained.

Watercut, %	Results of calculations according to formulas, relative units		
	$\omega_{od} / \omega_0 = (1 - \varphi)^{-4.72}$	$\omega_{od} / \omega_0 = (1 - \varphi)^2 10^{-1.82\varphi}$	$\omega_{od} / \omega_0 = \frac{0.123}{\varphi} (1 - \varphi)^3$
5	0.7558	0.7319	
10	0.6095	0.5327	
20	0.3504	0.2768	
30	0.1871		0.1406
40	0.0906		0.0664
50	0.0385		0.0308
60	0.0135		0.0131
70	0.0035		0.0047

Data for the watercut degree from 70% to 90% have not been found. The velocities of oil globules rising distributed by fractions are of practical interest. The known relations are obtained for a case when the dispersed phase is water. In our case, the dispersed phase is oil. In paper [7] it is assumed that the distribution of water drops in oil after filling the chamber of separator is uniform. Therefore, the watercut  $B$  in any vertical section is the same. The relative velocity of the constrained surfacing of oil globules of  $d_i$  diameter in this case equals:

$$(w_{od} / w_0)_i = \left[ \frac{1 - B}{1 - B \sqrt{1 - (d_i / d_{\max})^2}} \right]^{4.72},$$

where  $d_{\max}$  is the maximum size of the globule. Generally, it is suggested to define the separation time through the velocity of the oil globules rising and the geometric parameters of the oil and gas separator. The velocity of constrained rising of oil globules is as follows:

$$w_0 = \frac{(p_d - p_c) d^2 g}{18 \mu_c},$$

where  $p_d, p_c$  is the density of dispersed and continuous medium, kg/m<sup>3</sup>;  $\mu_c$  is the viscosity of the continuous medium, Pa s;  $d$  is the diameter of the globule;  $g$  is the gravitational acceleration, m/s<sup>2</sup>. While rising, globules of different sizes are moving at different speeds. It is proposed to use the following expression to describe the calculation of the constrained rising of globules:

$$(w_{0d} / w_0)_i = \frac{(p_d - p_c)d^2 g}{18\mu_c} \left[ \frac{1 - B}{1 - B\sqrt{1 - (d_i / d_{\max})^2}} \right]^{4.72},$$

The equation obtained enables us to calculate the spectrum of velocities of constrained rising of the oil globules, taking into account the change of watercut of emulsion according to the height of the partition wall between the chambers of the separation device. In Table 2 there are given the data for the watercut values of the studied range from 70% to 90%. In accordance with table 2 and geometric characteristics of the separation device, it is possible to determine the separation time for oil-containing mixture, taking into account the composition of the mixture and watercut value.

Table 2. The velocity of oil globules rising in the chamber of separation device.

	Watercut, %	90	85	80	75	70
Globules diameter, $\mu\text{m}$		The velocity of the oil globules rising, cm/s				
50	78.2609	128.35	184.70	240.26	329.47	417.91
60	54.5455	89.45	128.76	167.50	229.70	291.35
80	30.7692	50.46	72.62	94.46	129.54	164.31
100	19.6721	32.26	46.42	60.39	82.81	105.04
150	9.6257	15.79	22.73	29.56	40.54	51.42
200	4.9113	8.05	11.59	15.07	20.67	26.22

#### 4. Conclusion

The obtained relations enable us to mathematically describe the separation process through the velocity of the globules in the case of a high watercut value for oil-containing mixture, which demonstrates the scientific novelty. Using the mathematical model data and knowing the geometric characteristics of the separation device, separation time can be calculated. The precise definition of separation time improves the performance of the separation device without sacrificing the quality of the commercial oil, which is of great practical importance.

#### Acknowledgements

The authors express their gratitude for LLC Coordination (Ufa), LLC "GIRS", "Neftestroy" Educational Centre (Samara) for the fruitful cooperation and experimental data provided.

#### References

- [1] Zelenskiy VA, Shchodro AI. Increasing the efficiency of separation by controlling the differential density of the oil and gas mixture. Bulletin of Samara State Technical University. Series "Technical Sciences" 2015; 1(45): 178–183.
- [2] Zelenskiy VA, Shchodro AI. The automated control system of the oil and gas separator providing control over the density of the oil-containing mixture. Bulletin of Samara State Technical University. Series "Technical Sciences" 2016; 1(49): 15–23.
- [3] Zelenskiy VA, Shchodro AI. Analysis of errors of measuring the density of the oil-containing mixture and their impact on the determination of the separation time. Bulletin of Samara Scientific Centre of the Russian Academy Sciences 2016; 18(3): 896–901.
- [4] Zelenskiy VA, Shchodro AI. Method, mathematical model, and the algorithm of control over the oil separation process. Bulletin of Samara State Technical University. Series "Technical Sciences" 2016; 3(51): 21–28.
- [5] Cochin NE, Kibel IA, Rose NV. Theoretical hydromechanics, part 2. Moscow: Fizmatlit, 1963; 727 p.
- [6] Ponomarev SV, Mishchenko SV. Methods and devices for measuring the effective thermal performance of the flows of technological liquids. Tambov: Publ. Tambov State Technical University, 1997; 249 p.
- [7] Astarita J, Maruchchi J. The basics of hydromechanics of non-Newtonian fluids. Moscow: "Mir", 1978; 312 p.

# Isopropylbenzene oxidation reaction computer simulation

M.K. Vovdenko<sup>1</sup>, I.M. Gubaidulin<sup>1,2</sup>, K.F. Koledina<sup>1,2</sup>, S.N. Koledin<sup>1,2</sup>

<sup>1</sup>Institute of Petrochemistry and Catalysis Russian Science Academy, Prospect Oktyabrya 141, 450075, Ufa, Russia

<sup>2</sup>Ufa State Technological Petroleum University, Kosmonavtov street 1, 450062, Ufa, Russia

## Abstract

Isopropylbenzene (cumene) oxidation by air oxygen is intermediate stage of phenol and acetone production in cumene method. Cumene method of synthesis is currently the most wide-spread in the world to produce these two chemicals. Reaction is radical-chained and there also unwanted byproducts received along with desired ones.

This reaction has been subject of study since 30th of XX century. Different authors propose different ways of chain reactions, initiation mechanisms and influence degree of different factors and parameters during process.

In this paper different kinetic models of this reaction are compared and results of computations are shown.

*Keywords:* isopropylbenzene; oxidation; phenol; acetone; radical-chain mechanism; mathematic modeling.

## 1. Introduction

Phenol and acetone are very important substances in today petrochemistry industry. Acetone is used in paints, varnishes and solvents production. Also it is used as intermediate substance in petrochemistry industry. Phenol is used as intermediate substance in big number of petrochemistry units, such as bisphenol-A (which is lately used for polycarbonate and epoxies production), phenolic resins production, etc. Currently world industry produces more than 7 million tons of phenol per year [1] and this number growth annually. More than 97 % of phenol in the world is produced from cumene method and isopropylbenzene oxidation [1].

## 2. Isopropylbenzene oxidation reaction computer simulation

First cumene type production units were built at 40<sup>th</sup> years of XX century in USSR (Russia) and Canada [2], but more detailed studies of process kinetics and mathematical model appeared in 60-70<sup>th</sup> of XX century [3,4,5,6]. Because of low calculation machines capacities of that time researchers tried to simplify kinetics and mathematical models making different assumptions.

The general scheme of substances conversion is shown on Fig. 1 [7]:

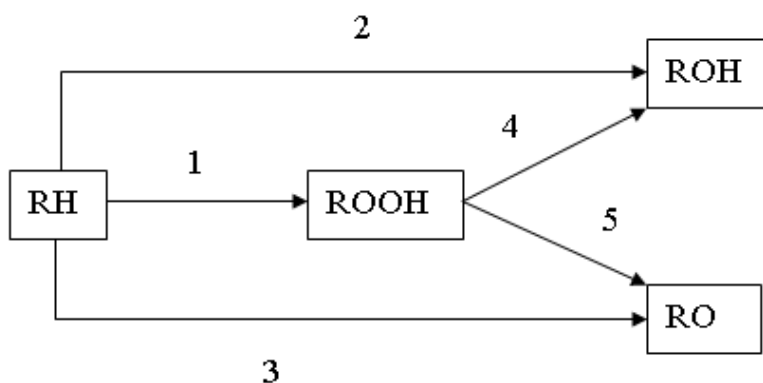


Fig. 1. General reactions scheme.

R• is cumyl radical ( $C_6H_5(CH_3)_2C\bullet$ ), RH is cumene ( $C_6H_5(CH_3)_2CH$ ), ROOH is cumene hydroperoxide ( $C_6H_5(CH_3)_2COOH$ ), ROH – dymethylphenylcarbinol (DMPC is  $C_6H_5(CH_3)_2COH$ ), RO is acetophenone ( $C_6H_5CH_3CO$ ).

But in reality this reactions follow radical-chain mechanism. There are number of elementary steps on which decomposition of particles and radicals recombination happens. Let's look at Hattori radical-chain reaction mechanism and on reaction scheme from [6] as examples [3].

Where In is initiator, which decomposes on radicals with an easy, so overall reaction process is enhanced. In some experiments author used benzene hydroperoxide as initiator [3] but we studied experiment without any additional initiator. Cumene hydroperoxide itself is used as initiator in most production units and laboratory researches. So reactions of stage 1 (1a and 1b) are replaced by 3a, 3b, 3c reactions for this case.

B1 and B2 are some substances which can be detected exactly (different organic acids, aldehydes, alcohols).

Reaction number	Reaction	Constant
1a	$\text{In}_2 \rightarrow 2 \text{In}\cdot$	-
1b	$\text{In}\cdot + \text{RH} \rightarrow \text{R}\cdot + \text{InH}$	-
2a	$\text{RH} \rightarrow \text{R}\cdot + \text{H}\cdot$	k2a
2b	$\text{RH} + \text{O}_2 \rightarrow \text{R}\cdot + \text{HO}_2\cdot$	k2b
3a	$\text{ROOH} \rightarrow \text{RO}\cdot + \cdot\text{OH}$	k3a
3b	$\text{RO}\cdot + \text{RH} \rightarrow \text{R}\cdot + \text{ROH}$	k3b
3c	$\cdot\text{OH} + \text{RH} \rightarrow \text{R}\cdot + \text{H}_2\text{O}$	k3c
4a	$\text{R}\cdot + \text{O}_2 \rightarrow \text{RO}_2\cdot$	k4a
4b	$\text{RO}_2\cdot + \text{RH} \rightarrow \text{R}\cdot + \text{ROOH}$	k4b
5a	$2 \text{R}\cdot \rightarrow \text{R-R}$	k5a
5b	$\text{R}\cdot + \text{RO}_2\cdot \rightarrow \text{ROOR}$	k5b
5c	$2 \text{RO}_2\cdot \rightarrow \text{ROOR} + \text{O}_2$	k5c
6a	$\text{RO}\cdot \rightarrow (\text{C}_6\text{H}_5)\text{CH}_2\text{CO} + \text{CH}_3\cdot$	k6a
6b	$\text{CH}_3\cdot + \text{O}_2 \rightarrow \text{CH}_3\text{O}_2\cdot$	k6b
6c	$\text{CH}_3\text{O}_2\cdot + \text{RH} \rightarrow \text{HCHO} + \text{H}_2\text{O} + \text{R}\cdot$	k6c
6d	$\text{RO}\cdot + \text{RH} \rightarrow (\text{C}_6\text{H}_5)\text{CCH}_2\text{CH}_2 + \text{H}_2\text{O} + \text{R}\cdot$	k6d

Table 2. Cumene oxidation kinetics scheme [6].

Reaction number	Reaction	Constant
0	$\text{RH} + \text{O}_2 \rightarrow \text{R}\cdot + \text{HO}_2\cdot$	k0
1	$\text{R}\cdot + \text{O}_2 \rightarrow \text{RO}_2\cdot$	k1
2	$\text{RO}_2\cdot + \text{RH} \rightarrow \text{ROOH} + \text{R}\cdot$	k2
3	$\text{ROOH} + \text{RH} \rightarrow \text{RO}\cdot + \text{R}\cdot + \text{H}_2\text{O}$	k3
4	$2 \text{ROOH} \leftrightarrow [\text{ROOH}]_2$	k4
5	$[\text{ROOH}]_2 \rightarrow \text{RO}_2\cdot + \text{RO}\cdot + \text{H}_2\text{O}$	k5
6	$\text{RO}\cdot + \text{RH} \rightarrow \text{ROH} + \text{R}\cdot$	k6
7	$\cdot\text{OH} + \text{RH} \rightarrow \text{H}_2\text{O} + \text{H}_2\text{O}$	k7
8	$\text{RO}\cdot \rightarrow \text{R}'\text{O} + \text{CH}_3\cdot$	k8
9	$\text{CH}_3\cdot + \text{RH} \rightarrow \text{CH}_4 + \text{R}\cdot$	k9
10	$\text{CH}_3\cdot + \text{O}_2 \rightarrow \text{CH}_3\text{OO}\cdot$	k10
12	$\text{CH}_3\text{OO}\cdot + \text{RH} \rightarrow \text{CH}_3\text{OOH} + \text{R}\cdot$	k11
13	$\text{H}_2\text{CO} + \text{R}\cdot \rightarrow \text{HC}\cdot\text{O} + \text{R}\cdot$	k12
14	$\text{HC}\cdot\text{O} + \text{O}_2 \rightarrow \text{HCOOO}\cdot$	k13
15	$\text{HCOOO}\cdot + \text{RH} \rightarrow \text{HCOOOH} + \text{R}\cdot$	k14
16	$\text{HCOOOH} \rightarrow \text{HCOO}\cdot + \cdot\text{OH}$	k15
17	$\text{HCOO}\cdot + \text{RH} \rightarrow \text{HCOOH} + \text{R}\cdot$	k16
18	$\text{HCOO}\cdot \rightarrow \text{CO}_2 + \cdot\text{OH}$	k17
19	$2 \text{RO}_2\cdot \rightarrow 2 \text{R}'\text{O} + 2 \cdot\text{CH}_3 + \text{O}_2$	k18
20	$2 \text{RO}_2\cdot \rightarrow \text{B1}$	k19
21	$\text{RO}_2\cdot + \text{R}\cdot \rightarrow \text{B2}$	k20

For Hattori reaction scheme simulation we made equations system based on Table 1 [8] where rate reactions are unknown:

$$W_{2a} = k2a \cdot [\text{RH}] \quad (1)$$

$$W_{2b} = k2b \cdot [\text{RH}] \cdot [\text{O}_2] \quad (2)$$

$$W_{3a} = k3a \cdot [\text{ROOH}] \quad (3)$$

$$W_{3b} = k3b \cdot [\text{RO}\cdot][\text{RH}] \quad (4)$$

$$W_{3c} = k3c \cdot [\cdot\text{OH}][\text{RH}] \quad (5)$$

$$W_{4a} = k4a \cdot [\text{R}\cdot] \cdot [\text{O}_2] \quad (6)$$

$$W_{4b} = k4b \cdot [RO_2 \cdot][RH] \quad (7)$$

$$W_{5a} = k5a \cdot [R \cdot]^2 \quad (8)$$

$$W_{5b} = k5b \cdot [R \cdot][RO_2 \cdot] \quad (9)$$

$$W_{5c} = k5c \cdot [RO_2 \cdot]^2 \quad (10)$$

$$W_{6a} = k6a \cdot [RO \cdot] \quad (11)$$

$$W_{6b} = k6b \cdot [CH_3 \cdot] \cdot [O_2] \quad (12)$$

$$W_{6c} = k6c \cdot [CH_3O_2 \cdot][RH] \quad (13)$$

$$W_{6d} = k6d \cdot [RO \cdot][RH] \quad (14)$$

Then we created system of differential equations, in which reagents concentration changing through time is shown [3,8]:

$$\frac{d[RH]}{dt} = -k2a \cdot [RH] - k2b \cdot [RH] - k3b \cdot [RO \cdot][RH] - k3c \cdot [\cdot OH][RH] - k4b \cdot [RO_2][RH] - k6c \cdot [CH_3O_2 \cdot][RH] - k6d \cdot [RO \cdot][RH] \quad (15)$$

$$\frac{d[ROOH]}{dt} = -k3a \cdot [ROOH] + k4b \cdot [RO_2][RH] \quad (16)$$

$$\frac{d[R \cdot]}{dt} = k2a \cdot [RH] + k2b \cdot [RH] + k3b \cdot [RO \cdot][RH] - k4a \cdot [R \cdot] + k4b \cdot [RO_2][RH] - k5a \cdot [R \cdot]^2 - k5b[R \cdot][RO_2] + k6c \cdot [CH_3O_2 \cdot][RH] + k6d \cdot [RO \cdot][RH] \quad (17)$$

$$\frac{d[RO_2 \cdot]}{dt} = -k4b \cdot [RO_2 \cdot][RH] + k4a \cdot [R \cdot] - k5b \cdot [R \cdot][RO_2 \cdot] - k5b \cdot [R \cdot][RO_2 \cdot] \quad (18)$$

$$\frac{d[RO \cdot]}{dt} = -k3b \cdot [RO \cdot][RH] + k3a \cdot [ROOH] - k6a \cdot [RO \cdot] - k6d \cdot [RO \cdot][RH] \quad (19)$$

$$\frac{d[\cdot OH]}{dt} = k3a \cdot [ROOH] - k3c \cdot [\cdot OH][RH] \quad (20)$$

$$\frac{d[R-R]}{dt} = k5b \cdot [R \cdot][RO_2 \cdot] + k5c \cdot [RO_2 \cdot]^2 \quad (21)$$

$$\frac{d[ROOR]}{dt} = -k2a \cdot [RH] - k2b \cdot [RH] - k3b \cdot [RO \cdot][RH] - k3c \cdot [\cdot OH][RH] - k4b \cdot [RO_2] \quad (22)$$

$$\frac{d[H_2O]}{dt} = k3c \cdot [\cdot OH][RH] + k6c \cdot [CH_3O_2 \cdot][RH] + k6d \cdot [RO \cdot][RH] \quad (23)$$

$$\frac{d[ROH]}{dt} = k3b \cdot [RO \cdot][RH] \quad (24)$$

$$\frac{d[H \cdot]}{dt} = k2a \cdot [RH] \quad (25)$$

$$\frac{d[HO_2 \cdot]}{dt} = k2b \cdot [RH] \quad (26)$$

$$\frac{d[CH_3 \cdot]}{dt} = k6a \cdot [RO \cdot] - k6b \cdot [CH_3 \cdot] \quad (27)$$

$$\frac{d[CH_3O_2 \cdot]}{dt} = k6b \cdot [CH_3 \cdot] - k6c \cdot [CH_3O_2 \cdot][RH] \quad (28)$$

$$\frac{d[HCHO]}{dt} = k6c \cdot [CH_3O_2 \cdot][RH] \quad (29)$$

$$\frac{d[(C_6H_5)CCH_3CH_2]}{dt} = k6d \cdot [RO \cdot][RH] \quad (30)$$

In order to solve this equation system we used MATLAB program. In fact we need to solve reverse kinetic problem [9] which means that we need to find chemical rates reactions constants for given experimental data. We used data from [3,6,7]. On Fig.2 graphs of concentration are shown, both for experimental and calculated data for experiment in [3].

### 3. Conclusion

In this paper different cumene oxidation mechanism were shown and calculation experiment was conducted.

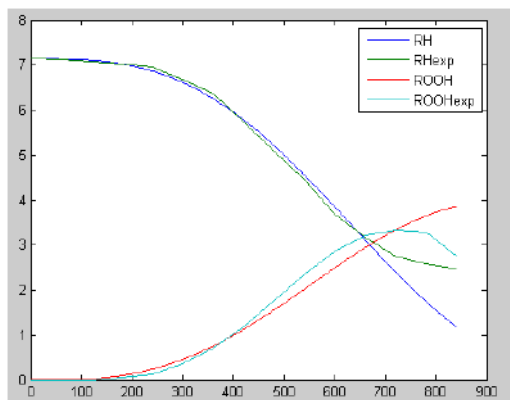


Fig. 2. MATLAB calculations result and experimental data [3].

## References

- [1] Zakoshansky VM. Alternative technologies of phenol and acetone production. Russian Journal of Chemistry 2008; LII(4): 53–71.
- [2] Zakoshansky VM. Phenol and acetone: Technologies, kinetics and main reactions mechanisms analyziz. SPb.: Khimizdat, 2009.
- [3] Hattori K, Tanaka Y, Suzuki H. Kinetics of liquid phase oxidation of cumene in bubble column. Journal of chemical Engineering of Japan 1970; 3(1): 72–78.
- [4] Andriago P, Caimi A, Cavalieri d'Oro P, Fiat A, Roberti L, Tampieri M, Tartari V. Phenol-acetone process: cumene oxidation kinetics and industrial plant simulation. Chemical Engineering Science 1992; 47(9-11): 2511–2516.
- [5] Dale GH. Rate Constants of Osidation of Cumene. Journal of the American Chemical Society 1967; 89(21): 5433–5438.
- [6] Makalec BI, Kirichenko GS, Stryigin EI. Liquid-phase model of cumene oxidation to hydroperoxide. Petrochemistry 1978; 18(2): 250–255.
- [7] Dahnavi EM, Ryazanov IG, Hardampidi HE. Temperature influence on catalytic cumene oxidation. Journal of Kazan Technological University 2009; 6: 263–266.
- [8] Gubaidullin IM. Informational and analytic system of inverse kinetic tasks: tutorial . Ufa: Published in BSU, 2003; 89 p.
- [9] Koledian KF, Gubaidullin IM. Program complex for inverse kinetic tasks solution and its realisation as virtual test stand. Science and education: science edition MSTU N.E. Bauman 2013; 7: 385–398.



# Optical digital system for DOE computation

S.K. Misievich<sup>1,2</sup>, R.V. Skidanov<sup>1,2</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

<sup>2</sup>Image Processing Systems Institute – Branch of the Federal Scientific Research Centre “Crystallography and Photonics” of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

## Abstract

This article considers algorithm implemented in an optoelectronic circuit. The main particularity of this algorithm is implementation in the working circuit of a part consisting of laser on the level of hardware, grey-level optical modulator of light and a diffraction pattern camera in the Fourier plane. This design allows using a laser with any initial distribution of intensity and adjusting a phase function of calculated DOE exactly for this distribution in order to decrease an error in forming a diffraction pattern of the output distribution.

*Keywords:* laser; light optical modulator; DOE; coding method; Fourier plane; Fourier transform

## 1. Introduction

For problems of synthesizing diffraction optical component (DOE) iterational (iterative) methods were developed and are used widely [1-11]. Their main advantage is that iterational algorithms prove to be more precise in comparison with other algorithms for DOE phase computation[6-10]. On the other hand, focusators computed with their help have irregular microrelief, which raises requirements for production technology of the components computed. Besides, the DOE computation using iterative algorithms requires significant expenses.

Disadvantage of algorithms executed with a computer is an implementation of intensity distribution approximation for a laser beam used as an illuminating beam for the DOE. This paper discusses an algorithm implemented in an optoelectronic circuit. The main particularity of this algorithm is implementation in the working circuit of a part consisting of laser on the level of hardware, grey-level optical modulator of light and a diffraction pattern camera in the Fourier plane. Such design allows using a laser with any initial distribution of intensity and adjusting a phase function of calculated DOE exactly for this distribution in order to decrease an error in forming a diffraction pattern of the output distribution.

## 2. Iterational algorithm

The problem of image recovery in a lens focussing plane, set by its amplitude-phase distribution is reduced to problem of minimizing functionality of amplitude deviation in a recovered image from a set value [1-3]:

$$\Phi = (|G(u, v)|^2 - |F(u, v)|^2), \quad (1)$$

where  $|F(u, v)|$  and  $|G(u, v)|$  - is a set and calculated wave amplitude in the plane of spatial spectrum.

Let us take a coordinate descent algorithm as a basis for functionality minimizing algorithm (1). For this, solving one-dimension problems of optimization shall be carried out with dichotomy method.

Let us use coefficients obtained with two-dimension unary re-expression [4] as coordinates for the coordinate descent:

$$F(u, v) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} F(x, y)A(u, v, x, y), \quad (2)$$

where  $A(u, v, x, y)$  - is the null-space for forward transformation;

$F(x, y)$  - size imaging matrix  $N \times M$ ;

$F(u, v)$ -transformed size imaging matrix  $N_1 \times M_1$ .

In the process of transforming (2) an initial image is described by a set of coefficients, quantity of which is significantly lower than dimensions of the initial data, which in its turn facilitates computational speed.

Recover  $y$  of an initial image is carried out by means of inverse transformation:

$$F(x, y) = \sum_{u=0}^{N_1-1} \sum_{v=0}^{M_1-1} F(u, v)B(u, v, x, y), \quad (3)$$

where  $B(u, v, x, y)$  - is the null-space for the forward transformation;

One of possible forward and inverse transformations representations for images by size  $N \times N$  may have form of forward and inverse Fourier transforms:

$$F(u, v) = \frac{1}{N} \sum_{j=0}^{N-1} \sum_{k=0}^{N-1} F(j, k) \exp \left\{ \frac{-2\pi i}{N} (uj + vk) \right\},$$
$$F(j, k) = \frac{1}{N} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} F(u, v) \exp \left\{ \frac{2\pi i}{N} (uj + vk) \right\}.$$

Let us compile algorithm scheme:

- 1) computation of expansion factors for a given initial approximation of a recovered image with expansion formula (3);
- 2) finding interval of expansion factor modification  $F(u_0, v_0)$ ;
- 3) carrying out of functionality minimizing with dichotomy method computed at previous stage of interval algorithm:
  1. image recovery using known expansion coordinates with an inverse transform (4);
  2. performing Fourier transform for obtaining distribution in lens focussing plane;

3. functionality recalculating (1) in relation to the distribution deduced;
  4. calculation of functionality Euclidean norm;
  5. finding interval of expansion factor modification;
  6. checking exit condition from dichotomy method;
- 4) checking exit condition from algorithm. Exit is effected in case we achieved functionality minimum with a set definiteness or we reset all the expansion factors.

At realization of this algorithm in an optoelectronic circuit, Fourier transform performance, which make a significant part of computations is transferred to hardware component [5]. Realization of these calculations is carried out by means of spatial light modulator, which modifies amplitude of illuminating beam. After this, the light passes through a collecting lens, forming a distribution in focussing plane, which is recorded by the camera and serves as a basis for functionality recalculating (1).

### 3. Experimental research

Let us conduct a simulation experiment for a DOE phase function optimization using a worked out iterative algorithm.

Let us use a result received with the iterative algorithm as initial approximation of a phase. We conduct results of numerical experiments for images that describe amplitude and phase distributions with dimensions of 256x256 pixels. For acceleration of the algorithm work we employ radially-symmetrical phase and amplitude distributions as investigated.

A ring given at figure 11 is used as a reference distribution.

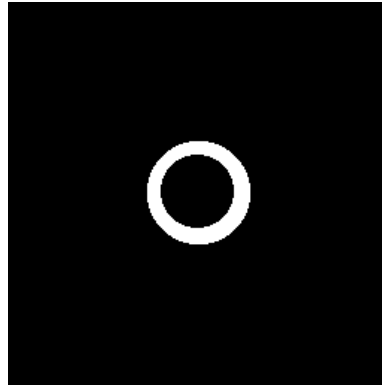


Fig. 1. Reference distribution of intensity.

As a method for inaccuracy estimation, a mean square deviation was used:

$$\varepsilon = \frac{\sqrt{\frac{1}{S} \sum_{(x,y) \in R} [I(x,y) - \hat{I}(x,y)]^2}}{\frac{1}{S} \sum_{(x,y) \in R} I(x,y)}, \quad (4)$$

where  $I(x, y)$  - is the distribution of the intensity formed;

$\hat{I}(x, y)$  – reference distribution of intensity;

$R$  – area of inaccuracy estimation;

$S$  - space of the area  $R$ .

For the calculated initial approximation, inaccuracy was 0.81. After implementation of iterative algorithm, the inaccuracy reduced to 0.48. Results of the algorithm performance are given in figure 2.

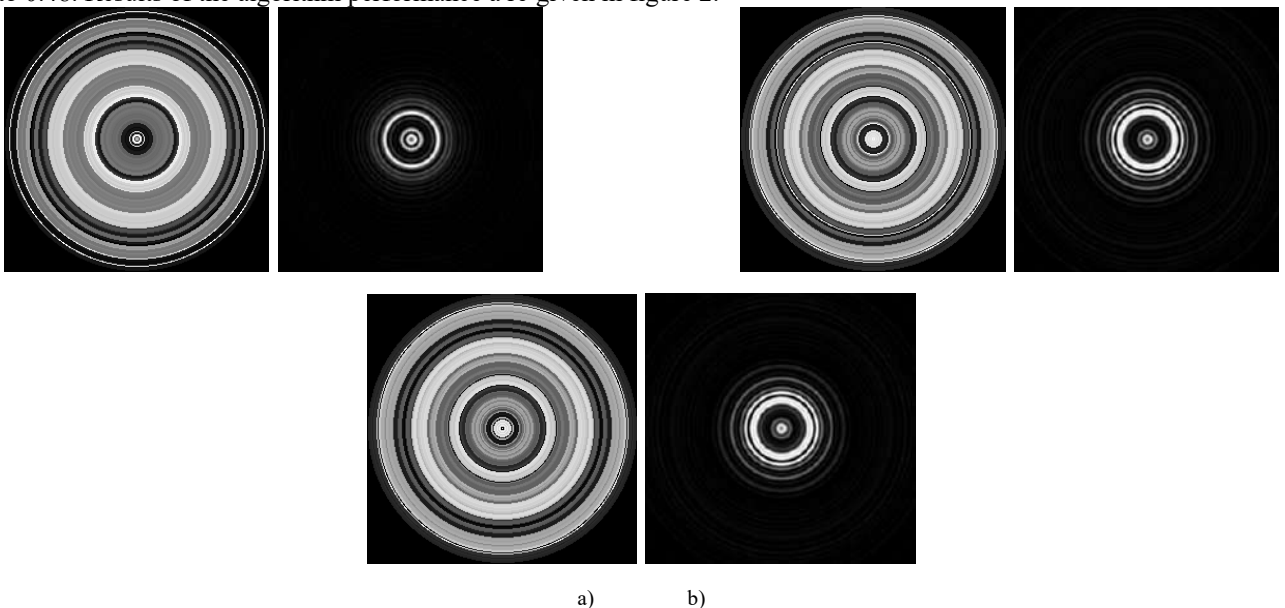
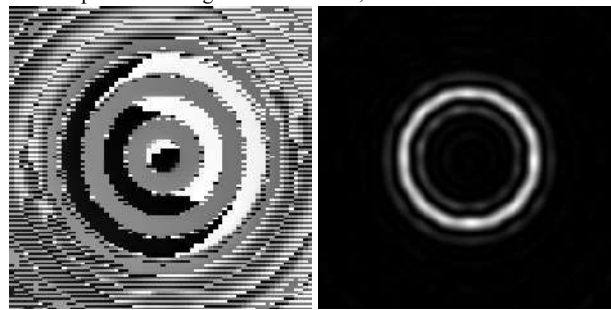
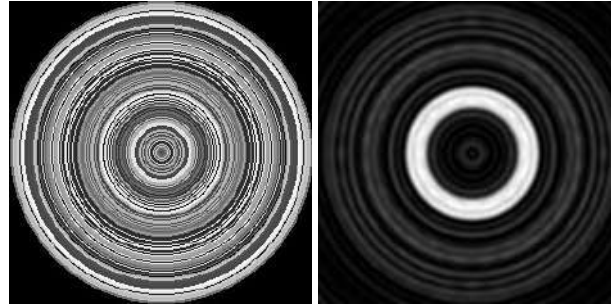


Fig. 2. Phase (a) and its corresponding intensity (b) for consecutive iterations of the algorithm with inaccuracy level 0,67; 0,54; 0,48 correspondingly.

Carrying out simulation implementing various initial approximations, we obtained results given in figures 3 and 4.

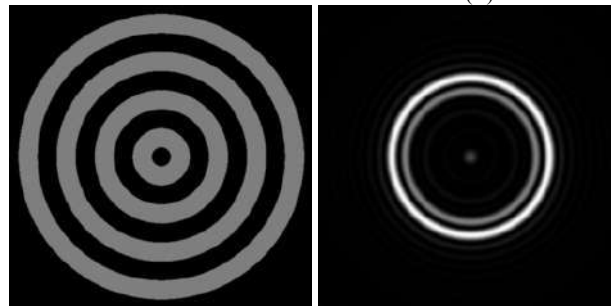


a)

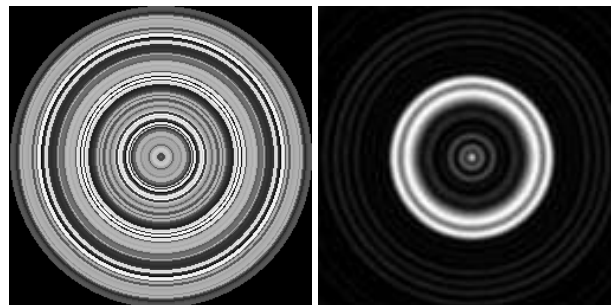


b)

Fig. 3. DOE phase and intensity obtained from Fourier plane for the initial approximation, which is computed using method of local phase jump (a) and at exit of the ALGORITHM PERFORMANCE (B).



a)



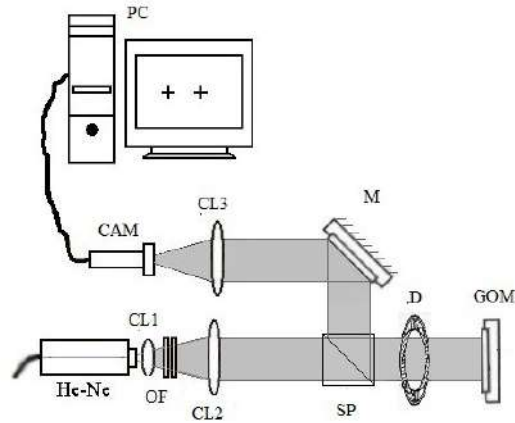
b)

Fig. 4. DOE phase and intensity obtained from Fourier plane for the initial approximation, which is an axicon (a) and at the exit of the algorithm's performance (b).

Fig. 3 and 4 show relative recovery inaccuracies, 25% and 32% correspondingly. Collation of the results obtained proves decrease in the algorithm's convergence speed when initial approximation increases.

To conduct an experiment, an optical design was built (see figure 15), with a grey-level optical modulator of light, model SLM PLUTO Phase Only [10,11,14].

At the exit of the laser a Fourier correlator is installed consisting of two lens (CL1, CL2) with different focal distance for beam blooming to a size that is capable of covering work panel of the modulator. Immediately before the modulator, the diaphragm is installed, which is necessary for varying diameter of the illuminating beam and for illuminating a certain area of the modulator. The laser beam falling onto the work panel of the modulator attached to the PC, changes its intensity and returns on the same trajectory. Reaching the splitter (SP), the beam divides in two parts, one of which as reflected by the mirror (M) and passing through a convex lens (CL3) forms some distribution in the Fourier plane. The distribution obtained is recorded by the camera attached to the PC displaying a formed diffraction pattern. The optical scheme was also equipped with different darkening optical filters (OF).



He-Ne – helium-neon solid-state laser, OF – optical filters, CL1, CL2, CL3 – convex lens, SP – splitter, D – diaphragm, GOM – grey-level spatial optical modulator CRL OPTO, CAM – camera VSTT-252, M – rotating mirror, PC – personal computer  
Fig. 5. Optical scheme used in the experiment.

Figure 6 shows a photography of the optoelectronic system in action.

Implementation of such scheme in an algorithm for computing DOE phase function is that the phase function corrected at each algorithm's iteration is brought off to the modulator, which functions as an actual DOE, then the camera records the diffraction pattern in the lens focussing plane. Further, the distribution obtained is processed by the computer. Discrepancy error between the reference and the obtained distribution is calculated and an operation of the phase correction occurs.

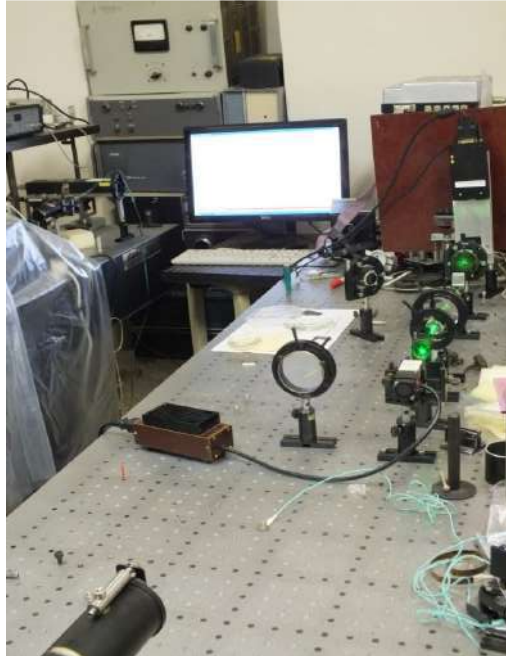


Fig. 6. The optoelectronic system.

As an initial approximation of the intensity formed we shall employ DOE shown on figure 3a. After its forming on the optical scheme, the distribution shown on fig. 7 was received.

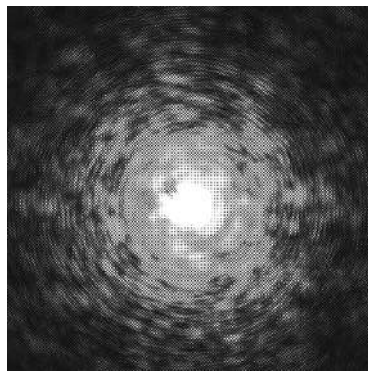


Fig. 7. The reference distribution of intensity.

Such distribution possesses a significant peak related to re-reflection effect, which occurs in the modulator. Further, fig.8 shows the result of a programme deduction from the distribution, which describes this peak.



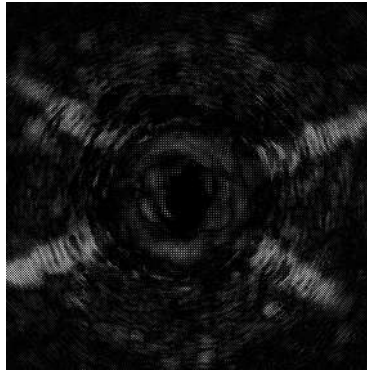


Fig. 8. The difference of intensity.

The calculated deviation of this intensity distribution from the reference one makes approx. 95%.

The result of the algorithm's work, which is implemented in the optoelectronic circuit is given in fig. 9. The distribution corresponding it, as obtained by means of diminution from the central peak is given in fig. 10. The DOE phase function computed has the form represented in fig. 11.

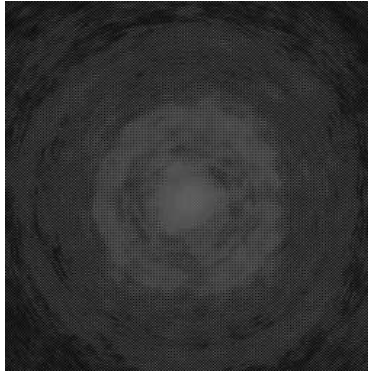


Fig. 9. The resultant distribution of intensity.

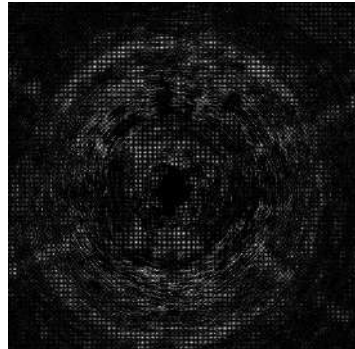


Fig. 10. The result of the programme deduction in the central peak.

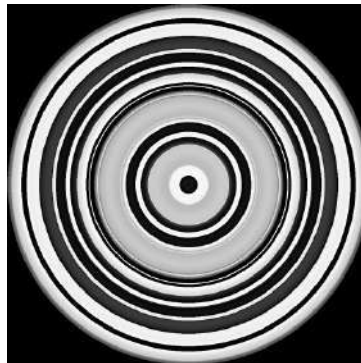


Fig. 11. The calculated DOE phase.

The distribution obtained as a result of the algorithm performance has a more distinct ringed structure, which corresponds more to the set reference distribution. The calculated deviation of this image from the reference one makes 83%.

An essential disadvantage of such computation method for the DOE phase function is a low speed of the algorithm's work, which is correlated to the output of a frequently alternating phase distribution onto the modulator.

#### 4. Conclusion

In work process, an algorithm for calculating the DOE phase function was elaborated, which is used for forming a radially-symmetrical distribution in the Fourier plane of the convex lens. Such algorithm was implemented on the computer for conducting a work simulation and convergence research, then transposed to the optoelectronic circuit.

The results of conducting a live experiment showed possibility of implementing this algorithm for calculating the diffraction optical components. The main advantage of the algorithm is its universality in relation to the intensity distribution of the illuminating beam. When calculating the DOE phase function, there is no need for taking into account laser intensity distribution, as it is counted on the hardware level, as a result of the algorithm's work the phase function is corrected exactly for the given distribution. Although, a low speed of the algorithm's work in the optoelectronic circuit hinders its implementation.

Such algorithm may be used in cases, when it is necessary to calculate a component that gives a more distinct image at the output, using a laser with a specific intensity distribution.

#### Acknowledgements

The work was funded by the Russian Federation Ministry of Education and Science of state-assigned task No. 3.3025.2017/8.9

#### References

- [1]. Lanina EP. Organization of ECM and systems. Site of Irkutsk State Technical University, 2004. URL: [http://paralichka85.px6.ru/11future/glava11\\_1.htm](http://paralichka85.px6.ru/11future/glava11_1.htm) (date of reference: 09.01.2013).
- [2]. Kotlyar VV, Khonina SN, Melekhin AS, Soifer VA. Coding of diffraction optical components using method of phase jump. *Computer Optics* 1999; 19(9): 54–64.
- [3]. Soifer VA, Doskolovich LL, Golovashkin DL, Kazanskiy NL, Kharitonov SI, Khonina SN, Kotlyar VV, Pavelyev VS, Skidanov RV, Solovyev VS, Uspleniev GV, Volkov AV. *Methods for computer design of diffractive optical elements*. New York: John Wiley & Sons, Inc., 2002.
- [4]. Pratt W. *Digital processing of images*. Moscow: Mir Publishing house, 1982; 1: 312 p.
- [5]. Fast spatial light modulators speed optical-computing applications. *Vision Systems Design*, 1997. URL: <http://www.vision-systems.com/articles/print/volume-2/issue-8/applications/spotlight/fast-spatial-light-modulators-speed-optical-computing-applications.html> (date of reference: 23.12.2012).
- [6]. Kazanskiy NL, Kotlyar VV, Soifer VA. Computer-aided design of diffractive optical elements. *Optical Engineering* 1994; 33: 3156–3166. DOI: 10.1117/12.178898.
- [7]. Doskolovich LL, Golub MA, Kazanskiy NL, Khamov AG, Pavelyev VS, Seraphimovich PG, Soifer VA, Volotovskiy SG. Software on diffractive optics and computer generated holograms. *Proceedings of SPIE* 1995; 2363: 278–284.
- [8]. Golovashkin DL, Kazanskiy NL. Solving diffractive optics problem using graphics processing units. *Optical Memory and Neural Networks (Information Optics)* 2011; 20: 85–89. DOI: 10.1134/S1063776110120095.
- [9]. Kharitonov SI, Doskolovich LL, Kazanskiy NL. Solving the inverse problem of focusing laser radiation in a plane region using geometrical optics. *Computer Optics* 2016; 40(4): 439–450. DOI: 10.18287/2412-6179-2016-40-4-439-450.
- [10]. Kazanskiy NL. Research and education center of diffractive optics. *Proceedings of SPIE* 2012; 8410: 84100R. DOI: 10.1117/12.923233.
- [11]. Kovalev AA, Kotlyar VV, Porfirev AP. Generation of half-pearcey laser beams by a spatial light modulator. *Computer Optics* 2014; 38 (4) 658–662.
- [12]. Method of descent by coordinates. Multi-dimensional methods of optimization. URL: <http://school-sector.relarn.ru/dckt/projects/optim/pocspusc.htm> (date of reference: 15.12.2012).
- [13]. Dichotomy method. Encyclopedias and dictionaries. URL: <http://dic.academic.ru/dic.nsf/ruwiki/1034684> (date of reference: 17.12.2012).
- [14]. PLUTO: High-Resolution LCOS Phase Only Spatial Light Modulators. *HOLOEYE Pioneers in Photonic Tecnology*, 1997. URL: [http://www.holoeye.com/spatial\\_light\\_modulators\\_pluto.html](http://www.holoeye.com/spatial_light_modulators_pluto.html) (date of reference: 15.01.2012).

# Modeling of geometrical stability of the diffraction lens mount for a promising project of the outer space observation satellite

G.P. Anshakov<sup>1</sup>, V.V. Salmin<sup>1</sup>, K.V. Peresykin<sup>1</sup>, A.S. Chetverikov<sup>1</sup>, I.S. Tkachenko<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

---

## Abstract

The article gives an overview of a promising project of an observation spacecraft fitted with a diffractive optical system. The problem of ensuring a stable position of the elements of the optical system is considered. For this purpose, the load-bearing scheme of the mount of a diffraction lens previously proposed by the authors is used. In this work, a study is made of the influence of the geometric parameters of the structure on its stiffness characteristics. With the help of numerical optimization, the optimal values of the design parameters for minimizing structural mass are calculated.

*Keywords:* diffraction optics, space membrane optical system, finite element simulation, natural oscillations, numerical optimization

---

## 1. Introduction

In recent years, the project of an observation satellite that uses a diffractive lens to focus the light flux instead of a conventional mirror [1] is actively discussed in the scientific community. Due to the fact that the Fresnel lens is a thin perforated membrane, its weight is much less than the weight of a conventional mirror. This creates a possibility for creating an observation satellite with optical payload with a large aperture.

The MOIRE project [2] is of particular interest. It implies creation of an observation satellite fitted with a Fresnel lens 10 meters in diameter. This lens is to be mounted 60 meters from the body of the satellite. The overview of the system is resented in figure 1.

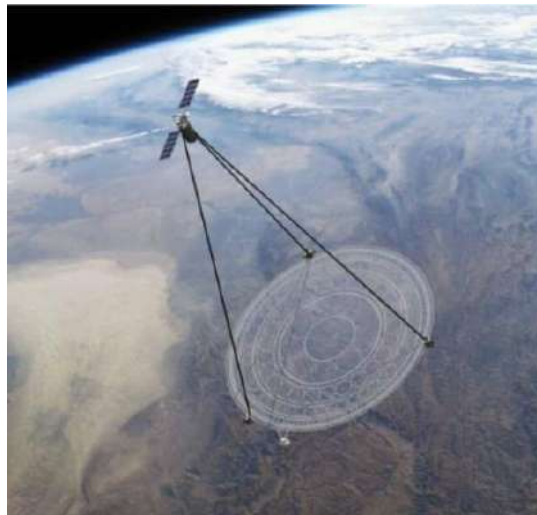


Fig. 1. MOIRE - Membrane Optical Imager for Real-Time Exploitation [1-3].

## 2. Structural requirements of the diffraction lens mount

Elements of the optical system are to be precisely positioned against each other coaxially and exactly on the right distance. This means that the dimensional stability is critical for the lens mount. Obviously, this design should be folded in the process of orbital injection, and be unfurled in orbit into the operating state. Ensuring that a structure this big remains dimensionally stable is a complex engineering task. Dimensional stability of the structure might be compromised by numerous factors: plastic deformations induced during the launch, temperature deformations, structural oscillations. The first two factors could be avoided by correct selection of the mount's material. Structural oscillations, however, will be inevitably induced during the process of positioning the telescope for image capturing. These oscillations in such a large-sized system can change the position of the lens relative to other elements of the optical system. If the amplitude of these oscillations is sufficiently large to distort the image and the decay time of these oscillations is high, it will severely hinder the image-capturing capabilities of the system. To prevent the occurrence of long-term oscillations of a lens with a large amplitude, the structure must have a sufficiently high rigidity. We propose to formulate rigidity requirements in the form of a restriction on the values of the natural vibration frequencies of the structure. Traditional observation satellites usually are equipped with solar arrays that have natural vibration frequencies in the range from 1 Hz to 2.5 Hz. We have to match these values for our proposed lens mount.

### 3. Load-bearing structure of the diffractive lens mount

In the previous paper [4] we have proposed a load-bearing scheme that could meet the aforementioned structural requirements for the mount. High rigidity in this scheme is achieved by combining three trusses into one structure by means of cables stretched between them. Strained cables load the trusses with transverse forces, which can lead to large deformations. To avoid this, in the proposed load-bearing scheme the trusses are arc-shaped and their ends are connected by a longitudinal cable as shown in figures 2-4. Such trusses function as arches and are capable of accommodating transverse loads. An overview of an observation satellite with a diffractive optic payload utilizing the proposed mount structure is shown in figure 5.

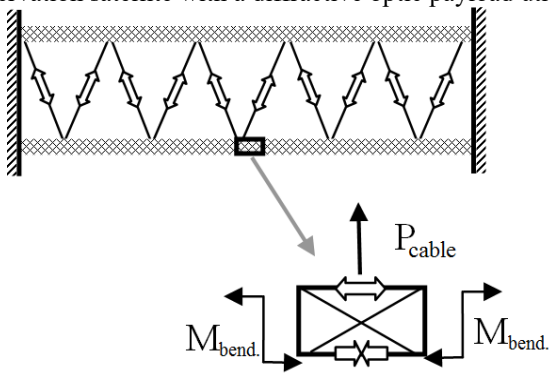


Fig. 2. Loading of straight trusses tightened with cables. Thick arrows represent internal forces induced in the structural elements and thin arrows represent external forces applied to the structure from other elements.

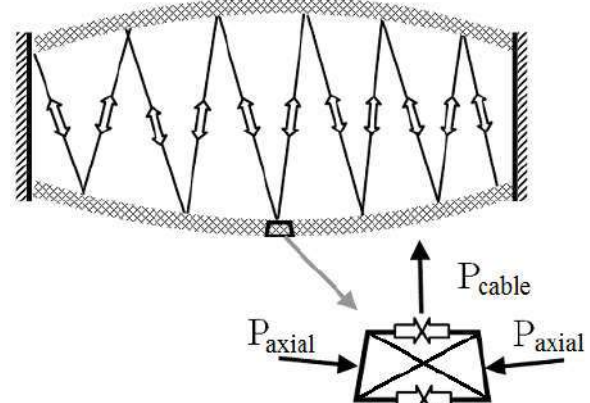


Fig. 3. Loading of arch-shaped straight trusses tightened with cables. Thick arrows represent internal forces induced in the structural elements and thin arrows represent external forces applied to the structure from other elements.

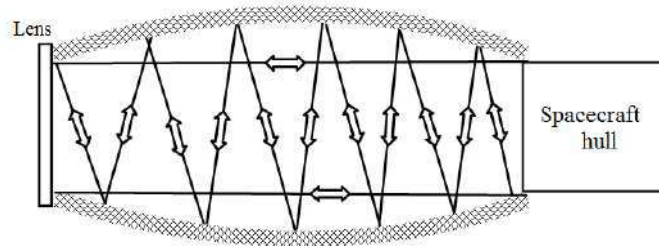


Fig. 4. Maintaining the distance between the ends of the arch-shaped trusses using longitudinal cables.

In the paper [4] we have presented a model of the loaded state of the proposed mount structure for a set of predetermined design variables and have shown the feasibility of the structural layout regarding the stiffness constraints. However, that paper did not contain a method for parametric optimization of the structure’s mass. The proposed design variables are: radius of arched trusses,  $R$ ; the cross-sectional area of the truss bars; the cross-sectional area of the cables. In this paper we study the influence of these parameters on the behavior of the lens mount structure.

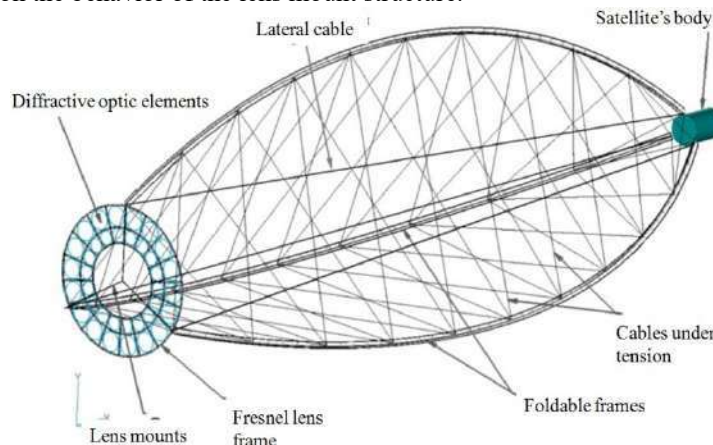


Fig. 5. An overview of an observation satellite with a diffractive optic payload utilizing the proposed mount structure.

### 4. Modeling method

Simulation of the behavior of the structure is performed in the finite element system MSC.Nastran. To estimate the rigidity of the mounting structure, the natural oscillations of a spacecraft with a diffraction lens were sought.

The search for natural oscillations in the method of finite elements consists in solving the following eigenvalue problem [2, 3]:

$$(-\omega_i^2 \cdot [M] + [K]) \cdot \{U_i\} = 0,$$



where  $\omega_i$  -  $i$ -th own circular frequency;  $[M]$  - mass matrix;  $[K]$  - stiffness matrix;  $\{U_i\}$  -  $i$ -th eigen form. The solution is performed for several lower tones of natural oscillations by the Lanczos method. The matrix of the rigidity of an elastic system within the framework of the finite element method has the following form:

$$[K] = \sum_{k=1}^{Ne} \int_{Vek} [B]_k^T \cdot [D]_k \cdot [B]_k dv,$$

where  $Ne$  - number of finite elements;  $Vek$  - volume of the  $k$ -th finite element;  $[D]_k$  - Hooke matrix for the material of the  $k$ -th finite element;  $[B]_k$  - matrix of connection between nodal displacements and deformations:  $\{\varepsilon\}_k = [B]_k \cdot \{u\}_k$ ;  $\{u\}_k$  - nodal displacements of the  $k$ -th finite element;  $\{\varepsilon\}_k$  - deformations of the  $k$ -th finite element. Coefficients of the matrix  $[B]_k$  could be obtained from differentiating the form function  $[\Phi]_k$  of the finite element by the corresponding coordinates. The matrix of the masses of the elastic system in the framework of the finite element method has the following form:

$$[M] = \sum_{k=1}^{Ne} \rho_k \cdot \int_{Vek} [\Phi]_k^T \cdot [\Phi]_k dv,$$

where  $[\Phi]_k$  - from function of the  $k$ -th finite element:  $\{u(x)\}_k = [\Phi]_k \cdot \{u\}_k$ ;  $\{x\}$  - coordinates of a point inside of a finite element;  $\rho_k$  - material density of the  $k$ -th finite element.

**5. Influence of the radius of truss arches on the behavior of the structure**

Radii of arched trusses determine the general geometry of the structure. To determine the influence of this parameter on the natural oscillations, a number of finite element models with different values of the radius of arched trusses were built in the MSC.Nastran system (Fig. 6). Some forms of oscillations for one of the considered radii are shown in Fig. 7-10. Dependences of natural frequencies on the radius of arched trusses are shown in Fig. 11.

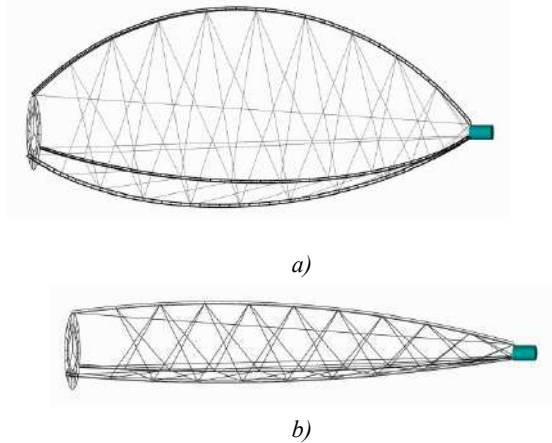


Fig. 6. Finite element models of the lens mount design with different values of the radii of the arched trusses: a) 40 m; b) 150 m.

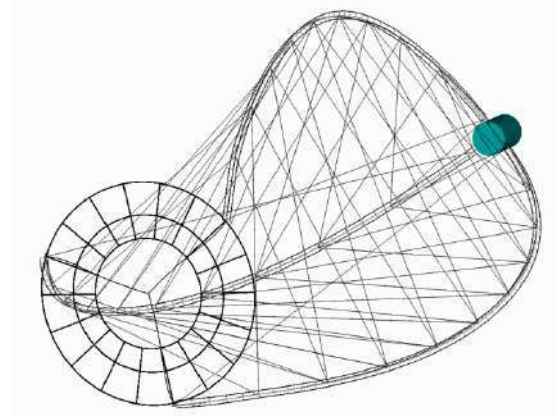


Fig. 7. Form of the first elastic tone of natural oscillations. Frequency - 0,687 Hz. Torsion of the lens around its axis.

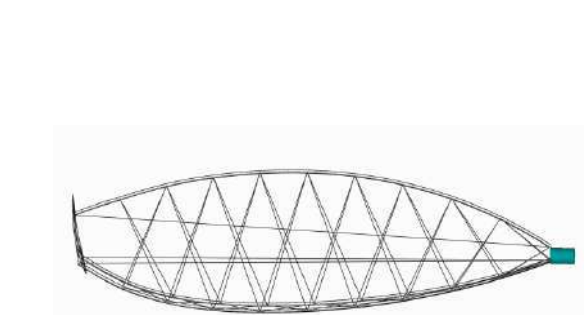


Fig. 8. Form of the second and fourth elastic tone of natural oscillations. Frequency - 0,948 Hz. The first flexural shape.

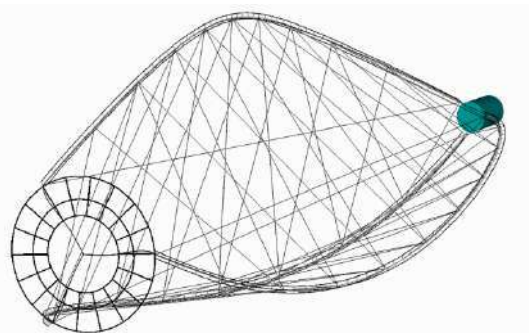


Fig. 9. Form of the fourth elastic tone of natural oscillations. Frequency - 1,14 Hz. Torsion of the lens around its axis.

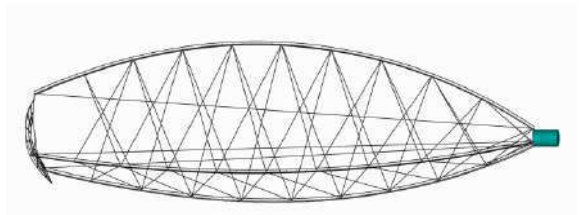


Fig. 10. Form of the fifth and sixth elastic tone of natural oscillations. Frequency - 1,48 Hz. The second flexural shape.

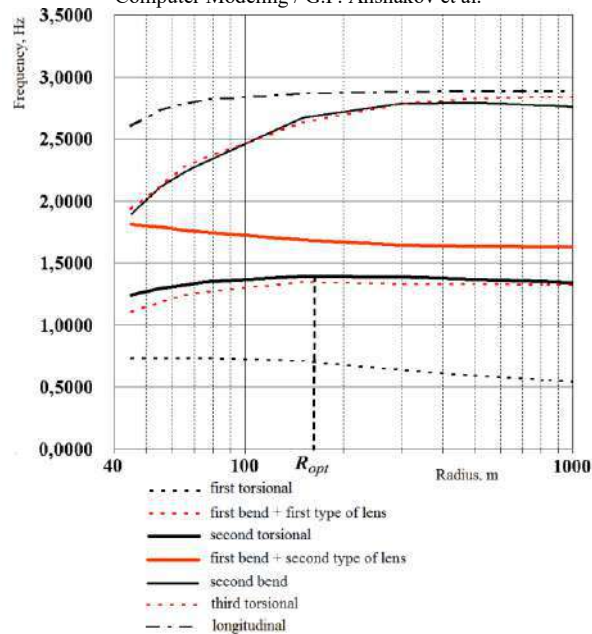


Fig. 11. Dependences of natural frequencies on the radius of arched trusses.

As far as the lower natural frequency is concerned, the smaller is the radius the higher is the stiffness. However, in the first tone of natural oscillation the lens is rotating around its optical axis. This kind of displacement does not affect the positioning of optical elements against each other and therefore does not affect the image quality of the system. The second and third forms of natural oscillations lead to tilting of the axis of the lens, which will lead to image distortion. If we would choose the radius of the arched trusses in order to maximize the frequencies of these tones, then an optimal radius value would be 150 m.

## 6. Parametric optimization of the parameters of the construction of the diffraction lens mount

The selection of the remaining parameters of the load-bearing structure was carried out using the procedure of parametric optimization of the MSC.Nastran system. The following formulation of the optimization problem was used [5,6]:

- Areas of the cross sections of the elements of the structure: the rods of the truss; beams of the lens mount, longitudinal cables and lateral cables were taken as design variables
- Design constraints: the frequencies of the five lowest tones of oscillation should not be less than 1 Hz; the tension of the cables should not lead to buckling of the structure; inertial loads from a typical orbital rotational maneuver should not cause destruction of the material of structural elements.
- The purpose of optimization is to minimize the weight of the structure.

For parametric optimization the MSC. Nastran system uses a gradient optimization method. Figures 12 and 13 show the changes of the structure's mass during optimization, the maximum value of the constraints and the values of the design variables. The results of optimization are shown using the example of a structure with a radius of arched trusses equal to 40 m.

As a result of optimization, the mass of the spacecraft decreased from 3620 kg to 3566 kg. The value of the natural tone frequency of the lower tone increased from 0.736 Hz to 1.0 Hz. Having carried out similar optimization calculations for models with different values of the radii of arched trusses, one can choose the design variant with the least mass. Thus, the optimal parameters of the load-bearing structure will be found. The authors did not perform a full series of calculations due to the preliminary and methodological nature of the study, but in the case of real design, there are no obstacles to finding optimal values of the parameters using the proposed method.

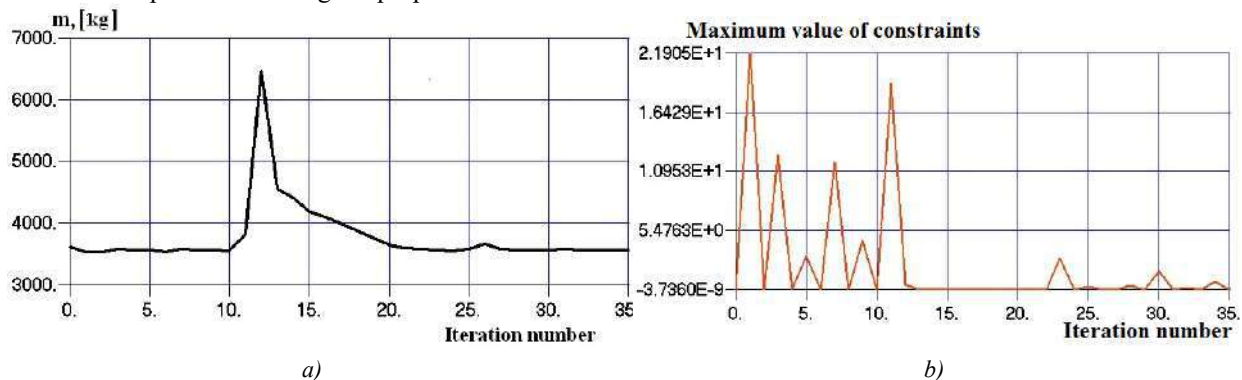


Fig. 12. Change in the process of optimization: a) mass of the spacecraft (target function) and b) the maximum value of the constraints (if the value is greater than zero - the constraint is not satisfied).

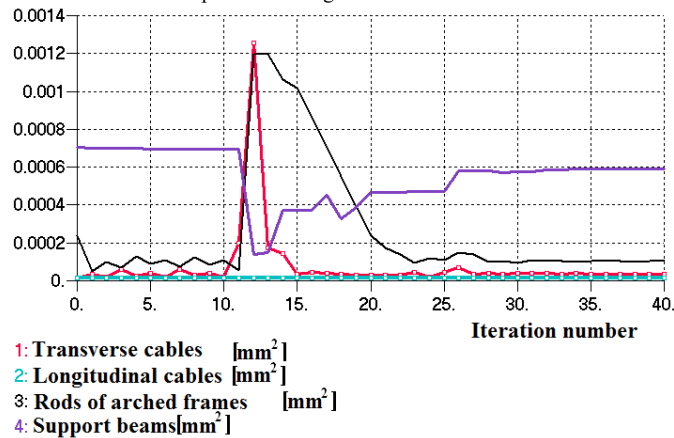


Fig. 13. Change of design variables during the process of optimization.

## Conclusion

The problem of stable arrangement of the optical elements for an observation satellite with a diffractive optic payload was considered. Particular attention was given to modeling and shaping the appearance of the diffraction lens mounting structure. To solve this problem, the authors proposed a large scale load-bearing structure of high rigidity. With the help of finite element modeling, the effect of geometric parameters of a structure on its stiffness characteristics is studied. The optimal values of the design parameters ensuring minimal mass of the structure were found using numerical optimization.

## Acknowledgements

This work is supported by the Ministry of Education and Science of the Russian Federation in the framework of The Federal purpose-oriented program "Research and development on priority directions of development of scientific-technological complex of Russia for 2014-2020" (agreement № 14.578.21.0229, unique identifier of the project RFMEFI57817X0229).

## References

- [1] Early J, Hyde R, Baron R. Twenty meter space telescope based on diffractive Fresnel lens. Proceedings of SPIE – The International Society for Optical Engineering 2004; 5166: 148–156.
- [2] Atcheson P, Stewart C, Domber J, Whiteaker K, Cole J, Spuhler P, Seltzer A, Smith L. MOIRE – Initial demonstration of a transmissive diffractive membrane optic for large lightweight optical telescopes. Proceedings of SPIE – The International Society for Optical Engineering 2012; 8442: 844221.
- [3] Atcheson P, Domber J, Whiteaker K, Britten JA, Dixit SN, Farmer B. MOIRE – Ground demonstration of a large aperture diffractive transmissive telescope. Proceedings of SPIE – The International Society for Optical Engineering 2014; 9143: 91431W.
- [4] Salmin VV, Chtverikov A, Peresypkin KV, Tkachenko IS. Modeling control of a large-size structure in the geostationary orbit. International Conference Information Technology and Nanotechnology. Session Mathematical Modeling. CEUR Workshop Proceedings 2017; 1904: 168–173. DOI: 10.18287/1613-0073-2017-1904-168-173.
- [5] Zienkiewicz O, Morgan K. Finite elements and approxiamtions. Moscow: Mir, 1986; 318 p. [in Russian]
- [6] Zienkiewicz OC, Taylor R. The finite element method. Fifth edition. Butterworth-Heinemann, 2000.

# Development of method for selecting motion control laws of space optical system on based diffractive membranes during transfer into geostationary orbit

G.P. Anshakov<sup>1</sup>, V.V. Salmin<sup>1</sup>, A.S. Chetverikov<sup>1</sup>, K.V. Peresyphkin<sup>1</sup>, I.S. Tkachenko<sup>1</sup>

<sup>1</sup>Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia

## Abstract

The article presents a formulation of a problem of trajectory optimization using low-thrust engines for an optical space system based on diffractive membranes. A methodology, where first stage nominal trajectories and control programs are selected and then corrected at the long-range guidance, has been developed for solving the problem of optimizing the trajectories of a flight to a geostationary orbit. At the final stage, algorithms for terminal control are formed, which allows to deliver a cosmic optical system based on diffraction membranes to a given point in the geostationary orbit. The end result is acquisition of Pareto-optimal solutions in the coordinates "characteristic speed-duration of the flight", where each point of the set of solutions has a corresponding a measure of accuracy of payload delivery to a geostationary orbit at a given set of coordinates.

*Keywords:* cosmic optical system; diffraction membrane; interorbital flight; geostationary orbit; multicriteria problem; terminal control algorithms; low-thrust engine

## 1. Introduction

Modern remote Earth sensing satellites function on low orbits. This imposes several constraints on their performance. For example, it is impossible to perform continuous monitoring of an object from a single satellite. In order to do so the satellite must be delivered to the geostationary orbit.

Since the geostationary orbit is quite high (about 36000 km), obtaining high-quality images requires complex optical systems. Remote geostationary Earth sensing satellites fitted with traditional refractive payloads capable of obtaining high-quality images are inevitably heavy which complicates their delivery to the orbit.

To solve this problem, the US Agency for Advanced Defense Research and Development of the US DARPA is developing a project of a modern space telescope with a membrane diffractive optical system. The project was named MOIRE or Membrane Optical Imager for Real-Time Exploitation [1-3].

The spacecraft with a membrane diffractive optical system will have a much smaller mass compared to a spacecraft with a refractive optical system. However, such a spacecraft would have large dimensions - the diameter of the lens is of the order of 10-20 m, the distance from the lens to the spacecraft body is of the order of 50 to 70 m.

The paper [4] presents a design of the load-bearing structure for an Earth remote sensing satellite with diffractive optics (Figure 1).

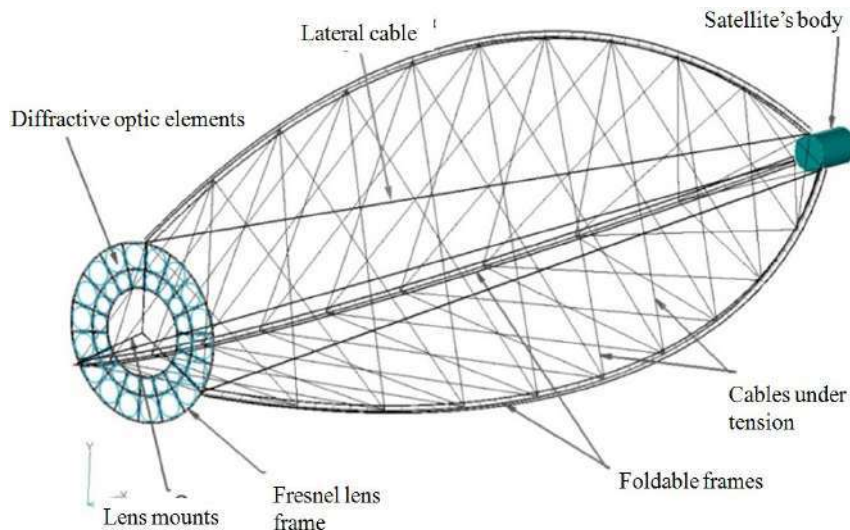


Fig. 1. Load-bearing structure for an Earth remote sensing satellite with diffractive optics [4].

When such a structure is being propelled by a chemical booster block, significant overloads can occur, which can lead to undesirable structural changes of the diffractive observation system.

In this case, it is preferable to use low-thrust electric propulsion engines, that are creating accelerations of the order of  $0.5-1.0 \text{ mm/s}^2$ , in order to bring the cosmic optical system on the basis of a large diffraction membrane to the geostationary orbit (GSO). The duration of transportation from low Earth orbit to GSO would be in the range from 100 to 200 days.

When optimizing ballistic schemes for such flights, it is necessary to seek a compromise between the mass of the payload and the duration of the flight - the main efficiency criteria [5]. The problem of ensuring the required accuracy of delivery is also very important.

Since spacecraft of this type have significant dimensions and, accordingly, mass-inertial characteristics, the process of motion control is significantly hampered. This calls for a solution of the problem of joint optimization of trajectory and angular movements.

## 2. Statement of the Problem

Let us consider a ballistic scheme for a transfer of a spacecraft from initial circular orbit to operational (geostationary) orbit, with insertion into a given station and return of the space tug to the initial orbit.

The optimization problem in a general statement is shaped as the problem of tied choice of design parameters,  $p \in P$ , ballistic parameters,  $b \in B$  and set of functions  $u(t, x)$ ,  $x(t)$  out of allowable multitude  $D$ , ensuring the transfer of a spacecraft from the initial state  $x(t_0) = x_0$  to a finite multitude  $x(t_f) \in X_f$  with the maximum value of the optimization criterion.

Relative payload mass (relation of the payload mass  $M_{pl}$  to launch mass of the spacecraft  $M_0$ ) is adopted as the main optimality criterion  $\mu$ :

$$\mu = \frac{M_{pl}}{M_0} \rightarrow \max.$$

Another criterion, no less important, is the overall transfer time  $T_\Sigma$ , which is a sum of the powered flight time  $T_P$  and the time of unpowered legs  $T_{UP}$ . The latter result from the need to make navigational observation and to "phase" precision targeting of an EP-powered spacecraft.

Let us represent the launch weight of a spacecraft as the sum of the masses of its functional elements: power plant (consisting of the reactor, energy transformer and radiator); thruster unit (consisting of cruising EP thrusters and controlling thrusters with their controls); propellant supply for the direct and return transfer, including additional expenses for control; propellant storage and supply system; payload; the body of the spacecraft:

$$M_0 = M_{PP} + M_{EP} + M_{P1} + M_{P2} + M_{PSS} + M_{PL} + M_B, \quad (1)$$

where  $M_0$  is the launch mass of the spacecraft;  $M_{P1}$ ,  $M_{P2}$  is the propellant mass for the direct and the return transfer respectively;  $M_{PL}$  is payload mass;  $M_{PP}$  is the power plant mass;  $M_{EP}$  is electric propulsion mass;  $M_{PSS}$  is the propellant storage and supply system;  $M_B$  is the spacecraft body mass.

The criterion  $\mu$  may be represented as [6]:

$$\mu = \max_{\substack{u(t) \in U \\ p \in P}} \mu(p, b, u(t), T, M_0) = \max_{\substack{u(t) \in U \\ p \in P}} \left( \frac{1 - A(z_1 + z_2 - z_1 z_2) - \mu_B}{1 - A z_2} \right), \quad (2)$$

where  $A = 1 + \gamma_{PSS} + \frac{c}{T} \left( \frac{\alpha_{PP} \cdot c}{2 \cdot \eta_P} + \gamma_{EP} \right)$ ;  $\alpha_{PP}$  - is the relative mass of the power plant;  $\gamma_{EP}$  - is the relative mass of electric propulsion;  $\mu_B$  is the relative mass of the body;  $\gamma_{PSS}$  is the relation of the propellant storage and supply system mass to propellant mass;  $P$  is the thrust of the cruising and controlling thrusters;  $c$  is the thruster exhaust velocity;  $z_1 = 1 - \exp\left(-\frac{W_{f1}}{c}\right)$ ;

$z_2 = 1 - \exp\left(-\frac{W_{f2}}{c}\right)$ ;  $W_{f1}$ ,  $W_{f2}$  - is the final characteristic velocity expense for direct and return transfer.

In this statement, the optimization problem is traditionally divided into two parts: dynamic (selection of optimal trajectories and control laws) and parametrical (selection of optimal design parameters).

Thus, two main optimality criteria are set: the relative payload mass  $\mu$  or, considering the expression (2), the final characteristic velocity expense  $W_f$ , as well as the total transfer time  $T_\Sigma$ . Then the optimization of the ballistic schemes, trajectories, and control modes as the ultimate goal is reduced to building a Pareto set in coordinates  $W_f$ -  $T_\Sigma$ . In addition, every point in the Pareto set must correspond to a measure of meeting the final boundary conditions  $\Phi$ , for example, in the following expression:

$$\Phi = \Delta x_f^T \Lambda \Delta x_f.$$

Here  $\Delta x_f$  is the vector of deviation of the final values of the state vector from required values;  $\Lambda$  is a positively determined weighted coefficient matrix.

The general mathematical model of motion includes the equation of the motion of the center of mass in equinoctial elements, dynamic equations of angular motion, kinematic equations in quaternion form.

## 3. The expansion principle in solution of a dynamic problem

The optimal control problem in this statement is extremely difficult, as the state of the controlled object is characterized by a set of variables, some of which are "slow-changing" (trajectorial motion) and some are relatively "fast-changing" (angular motion); besides, the mathematical model of motion includes a perturbation vector. It is apparently problematic to approach this problem with the classical methods of optimal control (the maximum principle and dynamic programming).

Therefore, we propose an approach to the solution of the dynamic problem, which is based on the principle of the extension – narrowing of the class of admissible states and controls [7].

The control vector  $u$  will include only the angles of the thrust orientation vector, and other components (controlling torques  $M_x, M_y, M_z$ ) will be subject to constraints. Let us impose constraints also on the state vector  $x$  at the final point of the trajectory. In this manner we obtain the following mathematical model:

$$x^* = f(x^*, u^*), \quad x^*(t_0) = x^*_{t_0} \quad (3)$$

where  $x^* = (h, e_x, e_y, i_x, i_y, F)^T$ ,  $u^* = (\vartheta, \psi)^T$ ,

with extended functional

$$L = V_f = \int_0^T a dt \rightarrow \min \quad (4)$$

and constraints  $x^*(t_f) = x^*_{t_f} = (h_f, e_{xf}, e_{yf}, i_{xf}, i_{yf}, F_f)^T \in X_f$ ,

$$\max_t |M_x(t)| \leq M_{\max\_x}; \quad \max_t |M_y(t)| \leq M_{\max\_y}; \quad \max_t |M_z(t)| \leq M_{\max\_z}. \quad (5)$$

Here  $h, e_x, e_y, i_x, i_y, F$  - are the equinoctial elements;  $\vartheta, \psi$  - are the thrust orientation vector angles;  $a$  - is reactive acceleration;  $M_{MAX\_X}, M_{MAX\_Y}, M_{MAX\_Z}$  are the maximum possible controlling torques that can be created by controlling thrusters of the spacecraft.

The conditions of (5) are checked as the result of a numerical modeling of the basic system of equations, with optimal control found by solving the problem for the simplified mathematical model.

Assume that the spacecraft is moving along a near circular orbit, so the eccentricity can be taken as zero. Assume also that the radial component of the acceleration is zero.

Then the thrust direction is determined by the angle  $\psi$  between the transversal and the thrust vector, and the projections of the reactive acceleration to the axis of the orbital system of coordinates are:

$$a_T = \frac{P}{M} \cos \psi, \quad a_S = 0, \quad a_W = \frac{P}{M} \sin \psi \quad (6)$$

Let us also exclude from the differential components the equations that describe the changes in the longitude of the ascending node and the perigee argument.

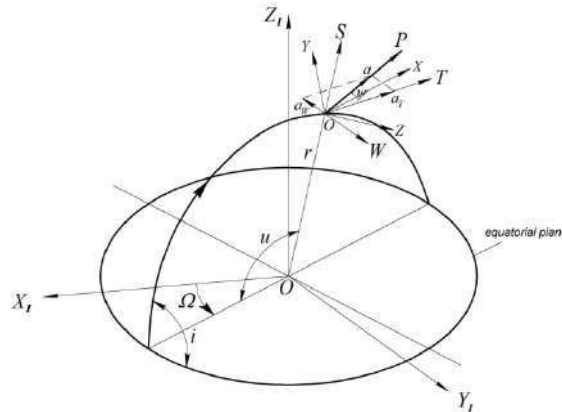


Fig. 2. Position of the spacecraft orbit and thrust vector control.

The system (3), taking into consideration (6), is transformed to the equations of a near-circular motion of a small-thrust spacecraft. By averaging the "slow-changing" variables  $r$  (average radius of a near-circular orbit, equivalent to semi-major axis) and  $i$  (orbit inclination) along the "fast-changing" variable  $u$  (polar angle of the argument of latitude) we can obtain the "asymptotic" model of motion [8].

Thrust vector control for transfers between non-coplanar orbits requires the sign of the  $a_W$  to change twice per revolution. The thrust angle orientation control is set in the following way:

$$\psi(W, u) = \psi_m(W) \text{sign}(\cos u), \quad (7)$$

where  $\psi_m$  is the amplitude of oscillations of the angle  $\psi$ ,  $W$  is current characteristic velocity and  $u$  is the argument of latitude.

Analytical solutions, obtained within the "asymptotic" model, describe a "universal" trajectory of a transfer between non-coplanar orbits, that does not depend on the spacecraft's design parameters (Figure 3).

#### 4. Method of selecting motion control laws of the space optical system on based diffractive membranes during transfer to given point on the geostationary orbit

The method includes an algorithm of developing nominal programs for thrust vector control, an algorithm of EP thrust magnitude adjustment, algorithm of terminal control development using motion models in discrete setting, numerical algorithm of building a set of Pareto-optimal solutions.

The goal of the control at the stage of lifting to GEO is narrowing the area  $G$  to an allowable area  $G_a$ .

The problem will be solved consequentially, in two stages:

- development of an algorithm for moving the final state deflection vector  $\Delta X_f$  into an area  $G'$ , where one or more of the components of the vector  $\Delta X_f$  satisfy the required precision (e.g. deflection by inclination  $\Delta i$ );



- developing the control laws and algorithms for narrowing the area  $G'$  to area  $G_a$ , where all components of the vector  $\Delta X_f$  satisfy the required precision conditions.

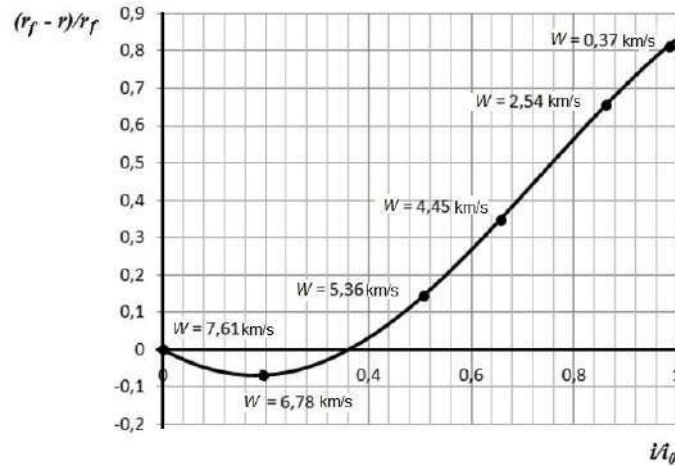


Fig. 3. Phase trajectory of a transfer to GEO.

#### 4.1 Goal of control at the long-range guidance stage

If the deflections are considerable, or it is impossible to build in advance a precise model of perturbing accelerations, which may change significantly during the transfer, it is advisable to use multistage control algorithms with end state prognosis and identification of perturbations.

Since sufficiently precise models of perturbations from the Earth's gravity field, solar and lunar perturbations are presently available, the parameter to be adjusted is the magnitude of reactive acceleration.

Thrust magnitude has been chosen as the adjusted parameter. Adjustment of actual thrust magnitude will be carried out according to the expression:

$$P^i = P_{nom} \left( 1 + \frac{\Delta T_{\Omega}^i}{T_{\Omega}^{calc}} \right), \quad (8)$$

where  $T_{\Omega}^{calc}$  is the oscillating orbit time at the measured unpowered leg;  $\Delta T_{\Omega}^i$  is the deviation of oscillating orbit time from calculated value.

An example of modeling the trajectorial motion of an EP-powered spacecraft during a transfer to GEO with set values of initial orbit ( $A_0, i_0, e_0$ ), exhaust speed  $c$  and initial acceleration  $a_0$  with adjustment of the thrust magnitude depending on the number of corrections  $N$  is represented in Table 1. Here  $\Delta P$  is the systematic error in thrust magnitude;  $t_{up}$  is the length of unpowered legs, where the low thrust propulsion unit is, as a rule, shut down to ensure more precise orbit parameter calculation and consequent adjustment of thrust magnitude.

Table 1. Results of modeling the motion of a EP OTV during transfer to GEO with adjustment of thrust magnitude ( $a_0 = 0,0006 \text{ m/s}^2$ ,  $c = 25 \text{ km/s}$ ,  $t_{up} = 0,5 \text{ day}$ ,  $i_0 = 51,6^\circ$ ,  $e_0 = 0$ ,  $A_0 = 7171 \text{ km}$ ,  $\Delta P = -1,5\%$  ( $P_r = 11,82 \text{ N}$ ))

N	Thrust and length of the $i$ -th powered flight leg			$\Delta A_f = A_f - A_f^{real}, \text{ km}$	$i_f, \text{ deg}$	$e_f$	$W_f, \text{ km/s}$
	$i$	$P^i, \text{ N}$	$T_a^i, \text{ days}$				
1	0	12	63,278	-464,1	-0,72	0,002	7,670
	1	11,665	66,079				
2	0	12	63,278	-12,8	-0,02	0,004	7,614
	1	11,665	33,039				
	2	11,789	32,222				

#### 4.2 Development of control laws and algorithms at the final stage of GEO transfer

**The goal of the terminal control.** The goal of the control is to move the end state deviation vector  $\Delta X_K$  from area  $G'$  to area  $G_a$ . Assume that the correction maneuver is performed with the help of transversal low thrust, which creates the transversal acceleration  $a_T$ .

This problem is set as a terminal control problem with functional

$$\Phi = \Delta x_f^T \Lambda \Delta x_f \rightarrow \min. \quad (9)$$

Here  $\Lambda$  is the fixed coefficient matrix;  $\Delta x_f = (\Delta T_f, \Delta \lambda_f, \Delta e_f)^T$ .

The control is structured as a sequence of powered and unpowered leg lengths  $u = \{\tau_1 \dots \tau_i, t_{up1} \dots t_{upi}\}^T$ .

Deviation of the semi-major axis  $\Delta A$  is equivalent to the deviation of the orbit time  $\Delta T = T - T_S$ . Here the orbit time on GEO equals star day  $T_S = 86164,09$  s.

In addition, the position of the OTV on GEO is described by longitude  $\lambda$ , which deviates from the required value of the station longitude  $\lambda_S$  by  $\Delta\lambda = \lambda - \lambda_S$ .

We shall solve this problem by increasing the sophistication of the motion model.

1. Neglecting the eccentricity value due to its insignificance, we obtain a near optimal analytical solution of the problem by Hamilton-Jacobi-Bellman's formalism.

The characteristic velocity budget for the correction maneuver with  $a_T = \text{const}$  is determined by the relation:

$$\Delta W_f = \frac{|\Delta T_0|}{3(T_S + \Delta T_0)^3 \sqrt{\frac{T_S + \Delta T_0}{2\pi\mu_E}}} . \quad (10)$$

The total time of the maneuver is determined by the equation

$$\Delta T_\Sigma = - \frac{\Delta\lambda_0}{\left( \frac{2\pi}{T_S + \Delta T_0} - \omega_E \right)} . \quad (11)$$

Here  $\tau$  and  $t_{up}$  are the lengths of the powered and unpowered flight legs respectively. Therefore, the characteristic velocity budget and time required for execution of the maneuver do not depend on design parameters (transversal acceleration  $a_T$ ), and are determined only by the initial deflections  $\Delta T_0$  and  $\Delta\lambda_0$ . Near optimal control by orbit time and longitude is performed in one step. Here by step we mean a sequence of a powered and an unpowered legs.

2. The problem of terminal control is solved with the help of multistage control algorithm with control parameters adjustment. Let the control law be set by a sequence of powered legs that is taken as decreasing and is defined by the expression [12]:

$$\tau_i = a \cdot \left[ 1 - \left( \frac{i-1}{n} \right)^b \right], \quad (12)$$

where  $i, n$  are the number of adjustment and total number of adjustments respectively;  $a$  and  $b$  are the parameters that describe the law of decreasing the lengths of the powered legs.

Then the problem of determining the optimal control law is reduced to a bi-parametric optimization problem, which is stated in the following way: for set initial values of orbital elements, acceleration  $a_T$ , number of adjustments  $n$ , lengths of unpowered flight legs  $t_{up}$  one should find such parameters  $a$  and  $b$  that ensure the minimum of the functional (9).

Control parameters  $a$  and  $b$  are found as the result of minimizing the functional of the view (9) and at that for better precision the dependency of the functional from parameter  $a$  is approximated by the least squares methods. When the control is adjusted (during motion modeling accounting for perturbations) at every unpowered leg the number of adjustment steps  $n$  is also corrected.

A series of calculations of control laws for a transfer of and EP-powered spacecraft to a given station by orbit time and longitude were carried out [9].

The characteristic velocity  $W$  budget, depending on the initial deviation by orbit time ( $\Delta T_0 = 300 \dots 1000$  s) is on the order of 4 to 14 m/s.

3. A discrete model for flat motion of a spacecraft under small transversal acceleration (with changes in orbit time, eccentricity and station longitude) was developed in [11], and an analytical solution for the problem of the search for the optimal control (lengths of powered and unpowered legs), minimizing final orbit time, eccentricity and station longitude errors, was obtained.

In [23] the authors propose an approximate method for solving the problem based on the three-step control algorithm for circulation period, eccentricity and longitude of the point of standing. Orbit correction is carried out using low-thrust electric rocket engine that produces acceleration in the transversal direction.

Discrete model of plane motion of geostationary satellites under the influence of small transversal acceleration is presented in as [11]:

$$\Delta T(k+1) = \Delta T(k) + 3a_T (T_3 + \Delta T(k)) \sqrt{\frac{T_3 + \Delta T(k)}{2\pi \cdot \mu}} \tau(k),$$

$$\Delta\lambda(k+1) = \Delta\lambda(k) + \left( \frac{2\pi}{T_3 + \Delta T(k)} - \omega_3 \right) \cdot (t_{\Pi}(k) + \tau(k)),$$

$$\Delta e(k+1) = \left| \Delta e(k) - 2 \cdot a_T \left( \frac{T_3 + \Delta T(k)}{2\pi \cdot \mu} \right)^{1/3} \tau(k) \right|,$$

where  $k = 0, \dots, N-1$ ,  $t_{\Pi}(k)$  is determined by formulas

$$t_{\Pi} = \frac{T_0}{2} (1 + 2m) - \frac{\tau}{2} - \frac{9_0 T_0}{2\pi}, \text{ for } a_T > 0,$$



$$t_{II} = T_0(1+m) - \frac{\tau}{2} - \frac{\vartheta_0 T_0}{2\pi}, \text{ for } a_T < 0.$$

Functional

$$I = \Delta X_K^T \Lambda \Delta X_K \rightarrow \min.$$

Where  $a_T$  – transversal acceleration,  $\vartheta_0$  – true anomaly angle before correction;  $\Delta X_K = \{\Delta T_K, \Delta \lambda_K, \Delta e_K\}^T$  – the final state vector, where  $\Delta T_K = T_K - T_3$ ,  $\Delta e_K = e_K - e_{GEO}$ ,  $\Delta \lambda_K = \lambda_K - \lambda_p$ ;  $T_K, e_K, \lambda_K$  – values of the orbital period, eccentricity and longitude of the satellite's standing point on the orbit at the end of the correction maneuver;  $T_3$  – circulation period of the spacecraft in geostationary orbit, equal to a star day  $T_3 = 86164,09$  c;  $e_{GEO}$  – eccentricity of the geostationary orbit;  $\lambda_p$  – longitude of the working point of standing of a satellite;  $\Delta T_0 = T_0 - T_3$ ,  $\Delta e_0 = e_0 - e_{GEO}$ ,  $\Delta \lambda_0 = \lambda_0 - \lambda_p$ , where  $T_0, e_0, \lambda_0$  – values of the orbital period, eccentricity and longitude of the point of standing on the orbit before the spacecraft correction maneuver.

Based on the proposed control structure an analytical solution is obtained for  $\tau_0, \tau_1, \tau_2, t_{II1}, t_{II2}$  [11]

$$\begin{aligned} \tau_0 &= \frac{\Delta e_0}{2 \cdot a_T \sqrt[3]{\frac{T_3 + \Delta T_0}{2\pi \cdot \mu}}}, \quad \tau_1 = \frac{1}{a_T \left(\frac{T_3 + \Delta T_C}{2\pi \cdot \mu}\right)^{1/3}} \left(1 - \sqrt{1 + \frac{\Delta T_C}{3 \cdot (T_3 + \Delta T_C)}}\right), \\ \tau_2 &= \frac{1}{a_T \left(\frac{T_3 + \Delta T_C}{2\pi \cdot \mu}\right)^{1/3}} \left(1 - \sqrt{1 + \frac{\Delta T_C}{3 \cdot (T_3 + \Delta T_C)}}\right) \times \sqrt{1 + \frac{\Delta T_C}{3 \cdot (T_3 + \Delta T_C)}}, \end{aligned} \quad (13)$$

where  $\Delta T_C = \Delta T_0 + 3a_T(T_3 + \Delta T_0) \sqrt[3]{\frac{T_3 + \Delta T_0}{2\pi \cdot \mu}} \tau_0$ ;

$$\begin{aligned} t_{n2} &= (1+3S) \left[ T_C \left(\frac{1}{2} + m\right) - \tau_1 \cdot \left(\frac{1-S}{2(1+3S)} + \frac{2}{2+3S}\right) \right], \text{ for } a_T > 0, \\ t_{n2} &= (1+3S) \left[ T_C(1+m) - \tau_1 \cdot \left(\frac{1-S}{2(1+3S)} + \frac{2}{2+3S}\right) \right], \text{ for } a_T < 0; \end{aligned} \quad (14)$$

where  $m \in Z$ ,  $S = 1 - \sqrt{1 + \frac{\Delta T_C}{3T_C}}$ ;

$$t_{n1} = (\Delta \lambda_C - \Delta \lambda_B) / \left( \frac{2\pi}{T_3 + \Delta T_C} - \omega_3 \right), \quad (15)$$

where  $\Delta \lambda_C = -2\pi - \left( \frac{2\pi}{T_3 + \Delta T_C} - \omega_3 \right) \cdot \tau_1 - \left( \frac{2\pi}{(T_3 + \Delta T_C) \cdot \left(1 + 3a_T \sqrt[3]{\frac{T_3 + \Delta T_C}{2\pi \cdot \mu}} \cdot \tau_1\right)} - \omega_3 \right) \times (t_{n2} + \tau_2)$ ,

$$\Delta \lambda_B = \Delta \lambda_0 + \left( \frac{2\pi}{T_3 + \Delta T_0} - \omega_3 \right) \cdot \tau_0.$$

The presented algorithm has shown high accuracy in modeling of the correction of orbit for a surveillance satellite, fitted with an electric propulsion engine. For example, for  $\Delta T_0 = 1000$  s,  $e_0 = 0,005$ ,  $\Delta \lambda_0 = 0,087$  rad the final orbit parameters deviations were: orbital period  $\Delta T_K = 1,3$  s, standing point longitude  $\Delta \lambda_K = 0,15^0$ , eccentricity  $\Delta e_K = 1 \times 10^{-4}$ . Durations of the active and passive sections were  $\tau_0 = 7758$  s,  $\tau_1 = 1997$  s,  $\tau_2 = 1998$  s,  $t_{n1} = 260200$  s  $\approx 3$  days,  $t_{n2} = 40170$  s  $\approx 0,46$  days.

Figure 4 and 5 shows an example simulation of the orbit control of geostationary spacecraft by using low thrust of EP.

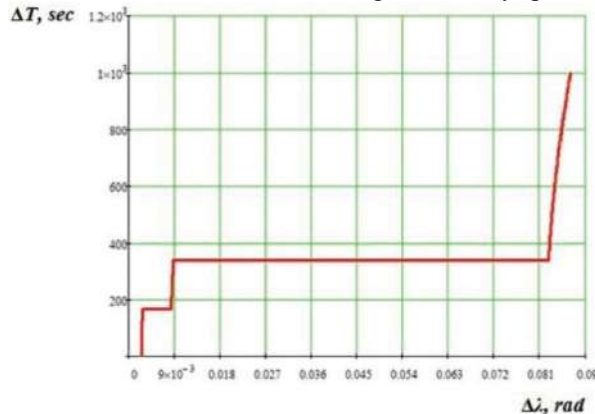


Fig. 4. Simulation the orbit correction for geostationary spacecraft using low thrust of EP ( $a_0 = 0,001$  m/s<sup>2</sup>,  $\Delta T_0 = 1000$  s,  $e_0 = 0,005$ ,  $\Delta \lambda_0 = 0,087$  rad).

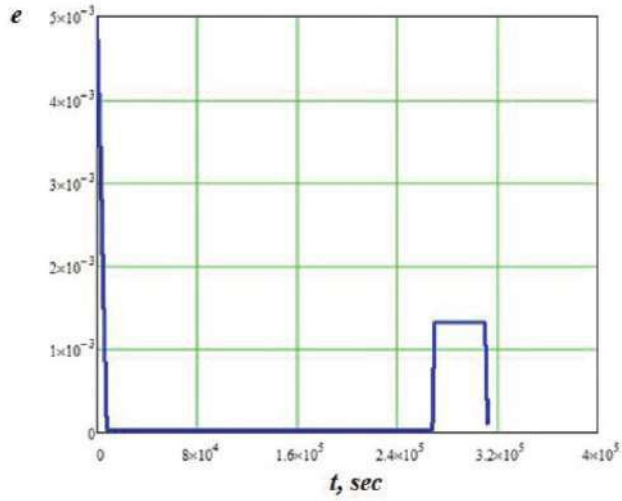


Fig. 5. The eccentricity change of geostationary spacecraft orbit in simulation the orbit correction using low-thrust of EP ( $a_0 = 0,001 \text{ m/s}^2$ ,  $\Delta T_0 = 1000 \text{ s}$ ,  $e_0 = 0,005$ ,  $\Delta\lambda_0 = 0,087 \text{ rad}$ ).

### 4.3 Building a set of Pareto-optimal solutions of a dynamic problem

Suggested control algorithms make it possible to build a set of Pareto-optimal solutions of the dynamic optimization problem of a space optical system on based diffractive membranes transfer to GEO using low thrust of EP with insertion into a given station.

The sequence of building the set of Pareto-optimal solutions is represented in Table 2.

Table 2. Building Pareto-optimal set

№	Control algorithms for insertion of powered EP spacecraft to GEO	Final error margin
1	1. Control program (7) with EP thrust adjustment algorithm (8) and correction of control program without precision GEO formation stage.	$\Phi_1$
2	1. Control program (7) with EP thrust adjustment algorithm (8) and correction of control program. 2. A near-optimal control by extended set used on the final stage.	$\Phi_2$
3	1. Optimal control program (7) with EP thrust adjustment algorithm (8) and correction of control program. 2. Control algorithm (12) used on the final stage.	$\Phi_3$
4	1. Optimal control program (7) with EP thrust adjustment algorithm (8) and correction of control program. 2. Three-stage control algorithm of terminal control (13) – (15) used on the final stage.	$\Phi_4$

Figure 6 shows a sample calculation of a set of Pareto-optimal solutions for a transfer to GEO with systematic thrust error  $\Delta P = -2,5\%$ .

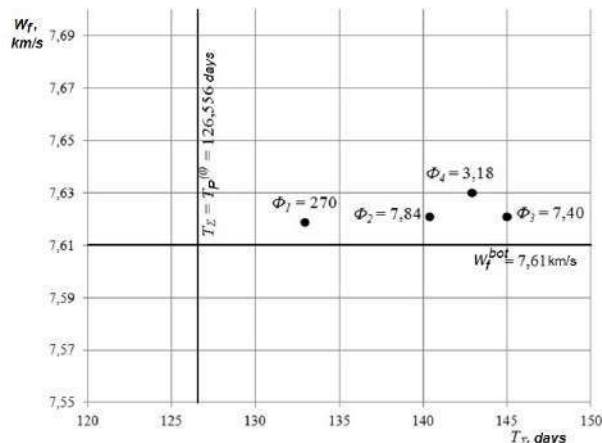


Fig. 6. Sample set of Pareto-optimal solutions ( $\Delta P = -2,5\%$ ,  $i_0 = 51,6^\circ$ ,  $r_0 = 7171 \text{ km}$ ,  $c = 25 \text{ km/s}$ ).

### Conclusion

A method of solving the dynamic optimization problem of a transfer to a given station on geostationary orbit was developed, including: an algorithm for obtaining nominal thrust control vector programs; algorithm for EP thrust magnitude adjustment on the basis of actual orbit time measurement at the long-range targeting stage; algorithm for obtaining terminal control, using discrete motion models; algorithm for obtaining a set of Pareto-optimal solutions.

## Acknowledgements

This work is supported by the Ministry of Education and Science of the Russian Federation in the framework of The Federal purpose-oriented program "Research and development on priority directions of development of scientific-technological complex of Russia for 2014-2020" (agreement № 14.578.21.0229, unique identificator of the project RFMEFI57817X0229).

## References

- [1] Early J, Hyde R, Baron R. Twenty meter space telescope based on diffractive Fresnel lens. Proceedings of SPIE - The International Society for Optical Engineering 2004; 5166: 148–156.
- [2] Atcheson P, Stewart C, Domber J, Whiteaker K, Cole J, Spuhler P, Seltzer A, Smith L. MOIRE - Initial demonstration of a transmissive diffractive membrane optic for large lightweight optical telescopes. Proceedings of SPIE - The International Society for Optical Engineering 2012; 8442: 844221.
- [3] Atcheson P, Domber J, Whiteaker K, Britten JA, Dixit SN, Farmer B. MOIRE - Ground demonstration of a large aperture diffractive transmissive telescope. Proceedings of SPIE – The International Society for Optical Engineering 2014; 9143: 91431W.
- [4] Salmin VV, Karpeev SV, Peresyarkin KV, Chetverikov AS, Tkachenko IS. Feasibility study and modeling of components for an informational space system based on a large diffractive membrane. CEUR Workshop Proceedings 2016; 1638: 132–148.
- [5] Grodzovsky GL, Ivanov YH, Tokarev VV. Space Flight Mechanics (Optimization Problems). Moscow: Nauka, 1975; 704 p. [in Russian]
- [6] Salmin VV, Ishkov SA. Trajectory and Parameter Optimization of Inter-Orbital Transfer Vehicles with Low-Thrust Propulsion Systems. Kosmicheskie Issledovaniya 1989; 27(1): 42–53. [in Russian]
- [7] Gurman VI. Extension Principle in Control Problems. Moscow: Nauka, 1985; 284 p. [in Russian]
- [8] Lebedev VN. Calculations of Low Thrust Spacecraft Motion. Moscow: CS AS USSR Press, 1968; 108 p. [in Russian]
- [9] Chernyavsky GM, Bartenev BA, Malyshev VA. Controlling the orbit of a geostationary satellite. Moscow: Mashinostroyenie, 1984; 144 p. [in Russian]
- [10] Salmin VV, Chetverikov AS. Selecting control laws for trajectorial and angular motion of a nuclear powered electric propulsion spacecraft during non-coplanar inter-orbital transfers. Bulletin of Samara Science Center of RAS 2013; 15(6). [in Russian]
- [11] Salmin VV, Chetverikov AS. Controlling the flat orbit parameters of a geostationary spacecraft with low thrust propulsion. Bulletin of Samara State Aerospace University 2015; 14(4): 92–101. [in Russian]