Using Some Features of Wavelets in the Search for Regularities in Applied Data

E.A. Nelyubina, V. V. Ryazanov, A. P. Vinogradov*
*vngrccas@mail.ru

Introduction

Previously, a number of studies were carried out where, when constructing a hypothesis about a regularity, suitable parametric space *Y* of the required form is chosen in the same way as it takes place in the Hough transformation scheme [1], [2], [3]. The secondary cluster structure $c^t \epsilon C^T$ arising in the space *Y* in this case contains information about the frequency of appearances of the regularity in the sample and about typical values of its parameters. Each realization of a regularity (as well as the vector representing it in the parametric space) is called 'generalized precedent' (GP), i.e., a precedent of the regularity itself. We show that certain local relationships of objects in the original sample may be considered as some multidimensional analogue of a small wave.

Regularity in data as a small wave

Here we focus on the local relationships of objects in the original sample, and form the space ***Y*** to represent these relationships in the form of some multidimensional analogue of a small wave.

Let *X* be a sample of digitized data of a large volume and dimensionality in the feature space R^N. The main object is a tuple of points $\boldsymbol{x}=\{x^1,\ldots,x^M\}$, $x^m \epsilon X$, $m=1,\ldots,M$, for which the conditions formulated by the expert are met:

$P_1(x^1,\ldots,x^M) = P_1(x^1_1,\ldots,x^1_N,\ldots,x^M_1,\ldots, x^M_N)$,
$P_2(x^1,\ldots,x^M) = P_2(x^1_1,\ldots,x^1_N,\ldots,x^M_1,\ldots, x^M_N)$,
…
$P_L(x^1,\ldots,x^M) = P_L(x^1_1,\ldots,x^1_N,\ldots,x^M_1,\ldots, x^M_N)$.

The set of conditions $\boldsymbol{P}=\{P_1,\ldots,P_L\}$ is a formulation of a hypothesis about the presence of a regularity. If the conditions ***P*** determine some specific form of location points of the tuple $\boldsymbol{x}=\{x^1,\ldots,x^M\}$ in the feature space, then this form is called 'basic cluster'. For example, if $\boldsymbol{P}=\{P_1,\ldots,P_L\}$, $L=N-2$, is a system of polynomial equations written by an expert, then in the space R^N an algebraic surface ***R*** is thus defined. Manifestation of the regularity at a point $\boldsymbol{x}\epsilon X$ corresponds to the fulfillment of the condition $\boldsymbol{x}\epsilon \boldsymbol{R}$. If this system ***P*** can be reduced to the form

$x_3=F_3(x^1_1, x^1_2)$,
$x_4=F_4(x^1_1, x^1_2)$,
…
$x_N=F_N(x^1_1, x^1_2)$,

and the functions $F_3,\ldots,F_N$ have no singularities, then the projection of the sample *X* onto the surface ***R*** can be seen on the plane (x^1_1,x^1_2), in particular, on the monitor screen. If the images of outlayers are weaken on the screen, the behavior of the regularity will then appear in certain "distilled" form. The example is very conditional, but the user can consider it as an ideal to which one should strive. In what follows, two illustrations are presented.

Now let *X*, $X \subset R^N$ be a sample of parameters of individuals in a population. If the expert is interested in the parent-child relationship for different pairs x^1, x^2 in *X*, $x^1 \neq x^2$, then he may define the measure of proximity of the tuple to the carrier of the similarity law: $\boldsymbol{\rho(x,R)} = \sum_1^N \boldsymbol{\sigma_n}, \boldsymbol{\sigma_n} = \begin{cases} \mathbf{1}, |\boldsymbol{x_n^1} - \boldsymbol{x_n^2}| \leq \boldsymbol{\varepsilon_n} \\ \mathbf{0}, |\boldsymbol{x_n^1} - \boldsymbol{x_n^2}| > \boldsymbol{\varepsilon_n} \end{cases}$, where ε_n, $n=1,2,\ldots,N$, is a set of boundaries on the scales of anthropometric indicators. Let's include the condition $x_1 \neq x_2$ in the list ***P***, and display marks about the presence of the regularity by points on the plane $(\rho,\varkappa)$, serving here as the GP space ***Y***, where $\varkappa = |x^1_{n'} - x^2_{n'}|$, and n' is the 'age' parameter. Then, in the vicinity of a certain point $(\rho=b,\ \varkappa\approx 25)$, $1<b<N$, a pronounced cluster will be observed on this plane, since the anthropometric indicators in the 'ancestor-descendant' pair are often close. But there are age-related changes during each individual life, and the definition of $\boldsymbol{\rho(x,R)}$ should be improved correspondingly.

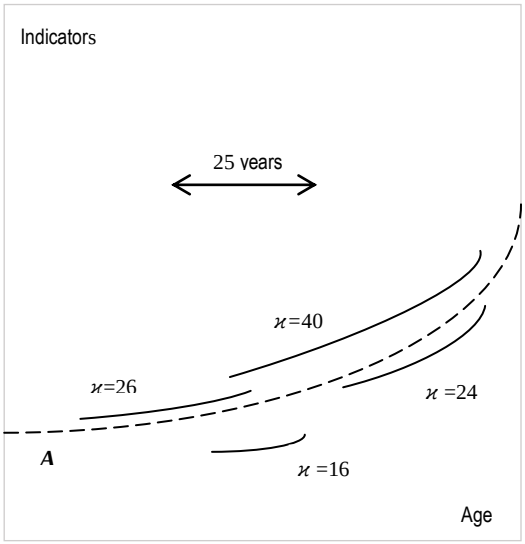

Fig.1. Modeled mean sequence ***A*** of changes in anthropometric indicators (dashed line). Small wave is a fragment of the vector function $\boldsymbol{\alpha(x)}$ cut from this sequence by two age points of two components of the tuple ***x***.

Thus, we change the definition of the proximity measure as a measure for comparing two tuples, as follows:

$$\rho(x,x^*) = \sum_1^N \sigma_n,$$

$$\boldsymbol{\sigma_n} = \begin{cases} \mathbf{1}, |\boldsymbol{x_n^1} + \boldsymbol{\alpha_n(x)} - \boldsymbol{x_n^2}| \leq \boldsymbol{\varepsilon_n} \\ \mathbf{0}, |\boldsymbol{x_n^1} + \boldsymbol{\alpha_n(x)} - \boldsymbol{x_n^2}| > \boldsymbol{\varepsilon_n} \end{cases}$$

where the vector function $\boldsymbol{\alpha(x)}$ acts as a small wave representing in this formula the correction of the reference tuple x^* in accordance with the age of the individuals. The previous statement without the function $\boldsymbol{\alpha(x)}$ corresponded to a situation of ideal similarity $x_n{}^1=x_n{}^2$, $n \neq n'$, between an ancestor and a descendant.

Testing the hypothesis may include testing of various admissible forms of the small wave of this type, and, in particular, assigning the detected typical forms to its various localizations in the parametric space ***Y***. So, the 'family' lines on Fig.1 show that the form of the small wave itself should be improved, too.

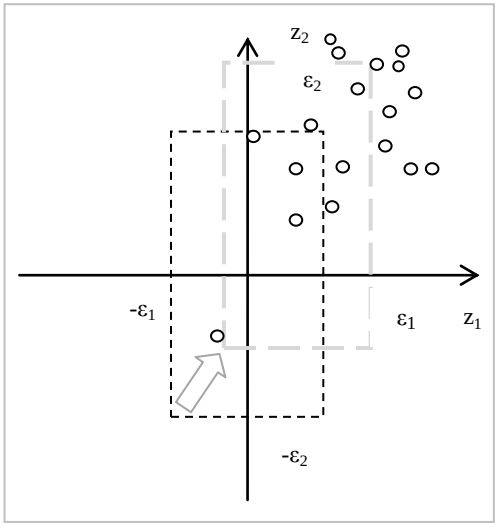

Fig.2. Modeled empirical distribution of GPs on the plane (z_1, z_2). Essential share of the main cluster lays out of the thresholds $\pm\varepsilon_n, n=1,2$. The rest of realizations is at the disposal, and one can form the direction (bold arrow) of improvement of $\boldsymbol{\alpha(x)}$ and thus of the thresholds (gray rectangle)

Conclusion

There are weighty considerations that the described above way to improve the shape of the wave has some methodological advantages. The poster presents an approach to the problem of finding regularities in applied data, which is based on the concept of GP as a numerical manifestation of some parameterized regularity in data. The paper shows that, in this case, it is appropriate to include some useful properties of the wavelet transform in the GP scheme and construct multidimensional analogues also of a small wave for sets of different and not necessarily spatiotemporal variables. Furthermore, in turn, it is shown how the GP scheme can be used again to correct hypotheses about the shape of a multidimensional small wave itself that describes complicated vicinity of an object by a small number of combined parameters. In general, when working with complex applied data, there is a huge spectrum of the new options open, and so, the central point again takes the adequate and operable integration of the modern applied and IT competencies into the structure and use of formal digital models.

References

1. Vladimir Ryazanov, Alexander Vinogradov. "Dealing with Realizations of Hidden Regularities in Data as Independent Generalized Precedents". IEEExplore, Proceedings of 2021 International Conference on Information Technology and Nanotechnology (ITNT), 2021, pp. 1-3.
2. Naumov V.A., Nelyubina E.A., Ryazanov V.V., Vinogradov A.P. "Analysis and prediction of hydrological series based on generalized precedents". Book of abstracts of the 12-th Int. Conf. Intelligent Data Processing (IDP-12), Gaeta, Italy, 2018, pp.178-179.
3. Vladimir Ryazanov, Alexander Vinogradov. "Analogues of Image Analysis Tools in the Problems of Finding Latent Regularities in Big Applied Data". Pattern Recognition and Image Analysis, v.32 No 3, 2022, pp. 639-644.