Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Tools and examples of intelligent processing, visualization and interpretation of GEODATA

To cite this article: V Taranchuk 2019 *J. Phys.: Conf. Ser.* **1425** 012160

IOP Conf. Series: Journal of Physics: Conf. Series **1425** (2020) 012160 doi:10.1088/1742-6596/1425/1/012160

Tools and examples of intelligent processing, visualization and interpretation of GEODATA

V Taranchuk

Belarusian State University, Nezavisimosti avenue, 4, Minsk, 220030, Belarus

e-mail: Taranchuk@bsu.by

Abstract. The article discusses the results of instrumental filling and usages of the integrated software complex designed for the preparation and testing of digital geological and geoecological models. Proposed methods of calculating the accuracy and confirming the reliability of the interpretation of the initial data are described. Examples of intelligent analysis of the results are given, the possibilities of obtaining and visualizing deviations are described.

1. Introduction

Digital geological, geoecological models are now a mandatory component of expertise in many areas, they occur in oil and gas production, in chemical industries, in the treatment of municipal and industrial liquid waste, in the construction industry, in biotechnology and many other industries [1]. Geological modeling includes the improvement of mathematical methods and algorithms; development of computer programs that provide a cycle of creating models; database design, their filling and maintenance. The data used in geological and geoecological models are a representative part of the geodata, which classify, summarize information about processes and phenomena on the earth's surface. This information becomes really useful when integrated into a single system [2]. Geodata, as a generalization of accumulated information, include information not only from the field of Earth sciences, but also from others, such as transport, economics, ecology, management, education, analysis, artificial intelligence. The technological feature of geodata ([2]) is that they are not obtained on the basis of direct measurements, but they are formed as a result of post-processing of the measured information, can have different accuracy, stored using different units of measurement. The system peculiarity is that after their formation geodata represent a set of parameters and descriptions of different types and structures integrated into a single complex, reflect different characteristics and properties, describe the existing spatial relations taking into account time and thematic factors. Interesting fact is that after geodata are grouped into three characteristics: place, time, topic, they behave like a new information resource. Another feature is the presence (implemented automatically) of mutual influence of graphic and attribute data – changing attribute data involves the replacement of graphic information, the refinement of the spatial position requires changes in coordinates, spatial relationships. This interaction provides a reliable foundation for spatial visual analysis and management.

The volume of geodata is growing at a very high rate. Accordingly, the application of Big Data Technologies (the specifics for geodata are noted in [3]), including Intelligent Data Analysis (IDA), is relevant. Article [3] emphasizes one of the main objectives of the IDA – detect “raw” (primary) amounts of data previously unknown, nontrivial, practically useful and intuitive interpretation of knowledge; the specificity of the observed data. In the author's statement [3] “data mining does not

IOP Conf. Series: Journal of Physics: Conf. Series **1425** (2020) 012160 doi:10.1088/1742-6596/1425/1/012160

exclude human participation in processing and analysis, but greatly simplifies the process of finding the necessary data from raw data, making it available to a wide range of analysts who are not experts in statistics, mathematics or programming. Human participation is expressed in the cognitive aspects of participation and application of information cognitive models". Geodata mining tools are the same as for ordinary data; the basis is the theory, methods, algorithms of applied statistics, databases, artificial intelligence, pattern recognition. There are many different existing and applied software tools for data mining. For example, in [4] the following classes of IDA systems are identified: "Industrial systems, Domain-specific analytical systems, Statistical packages, Artificial neural networks, Packages based on decision trees, reasoning systems based on similar cases, Genetic algorithms, Evolutionary programming".

The variety of proposed methods and software makes it necessary to assess the quality of geodata, determine their main characteristics. The quality assessment criteria of the geodata discussed in the article [5]. A number of the listed problems on the analysis, evaluation of the quality of spatial data can be solved using a computer complex "The generator of the geological model of deposit (GGMD)" – ([6-8]). In this paper, we discuss additional possible options for obtaining estimates, examples illustrate software tools which allow confirming the validity of interpretations, to visualize and obtain numerical values of errors calculated by different methods of intellectual data processing of results included and used in computer geological models.

2. Software platform. Components of GGMD

Technical solutions for the development of complex GGMD are detailed in [6-8]. The complex integrates the functions of the computer algebra system Wolfram Mathematica (The world's definitive system for modern technical computing) and of the geographic information system Golden Software Surfer (Surfer. Explore the depths of your data). The software implementation provides opportunities when the complex in a specific configuration can be operated after the assembly and saving in the computable document format CDF. Calculations, user work with a CDF version of the application is possible on any personal computer. When viewing the CDF version hosted on the web server, the viewer is automatically loaded as a browser plug-in. Offline work is possible after the installation of free distributed CDF Player.

In the computer system GGMD the following tools are implemented:

- tools and patterns for a preparation of a reference (calibration) model of digital field, which corresponds to the specified properties – "Digital field constructor (DFC)";
- tools and several options of "distortion" of a reference model;
- tools for data capture simulation, which are used in simulation practice – "Generator of profile observer (GPO)";
- modules for calculation, visualization, comparison of digital field approximation by several different methods – "Approximation component (ApC)";
- tools and adaptation modules for a digital model being formed, "Adaptation component (AdC)".

The main idea and purpose of the development of the GGMD complex is to provide the expert with tools for selecting and justifying the method of processing spatial data by comparing the reference digital field and the reconstructed from "observations". The reference digital field is considered as an explicit analytical function $z(x,y)$ defined in the domain with two independent variables. In the case of discrete representation, it is a grid function – set, an array of $z_{i,j}$ values at grid nodes (x_i,y_j).

In this statement, instead of the term "digital field" is often used "surface" as the most understandable for explanation. But it should be understood that from the standpoint of mathematical description we are talking about a method of explicitly specifying a function defined in the field with two independent variables. So you can model any distribution, for example: the concentration of contaminant or the density of the layer material, the saturation of the reservoir oil, porosity, hydroconductivity or capacity of the interlayer.

Software modules of the DFC group provide the construction of typical elements of the mathematical description (analytical functions) of the surface model. The resulting digital field (surface model) represents a visually synthesized set of fragments of different shapes. Construction in the DFC is performed in the interactive mode with the accompanying graphical visualization. At the same time, each expert advisor defines and includes, positions, scales, orients typical fragments into the model.

Then, the tools of the complex in the GPO component "are performed observations", simulated measurements of the reference distribution, and the geometry of the measurement points and their accuracy are also determined by the user of the complex and must correspond to the typical source data of the subject area. The user, using interactive tools of the GPO component, draws a profile diagram, setting the number of nodes on each of them. In a separate software module GGMD for each node of the profile scheme, data are recorded, from which five numbers are formed: the profile and node numbers, x and y coordinates, the calculated value of the surface function at the node point. The mentioned five numbers are exported to a file. The user can choose one of the varieties of offered file formats, including .XLS (.XLSX, .CSV). The scheme of placement of points with measurements does not have to be regular. As a result of this stage, the user receives a scattered (irregular) set of points, a set of "observations" data, the main of which are the coordinates of the node and the value of the parameter in it. This simulation of obtaining data on the shape of the surface (relief) gives the exact value of the reference surface level at the measuring point. Separately, in the GPO component there are tools to "distort the results of observations" (the simulation error of the measurements on each individual profile) by an additional overlay of noise with the use of different sets of functions with oscillation and random number generators [9].

The next step is to select the algorithm for processing the resulting data set ("reconstruction" of the digital field) by interpolating and extrapolating to a regular grid. Comparison of the results of "reconstruction" and the reference distribution ([9, 10]) will prompt the expert method of processing, the geometry of observation points.

The articles [6-8] describe and illustrate different examples of methodological and technical solutions of constructing, parsing, reference, and approximating digital field distributions of the parameters characterizing geological objects. There are more than ten variants of 1D, 2D and 3D graphics, including interactive modules for obtaining and designing maps and inserts for them, graphs on profiles, 3D visualizations of possible incisions (vertical, horizontal), clipping (simple and complex). In the mentioned articles the results are presented, examples of obtaining and visual comparison of different solutions are considered. Additional features and software tools that provide a more complete analysis of the quality of the generated models, allow to visualize and obtain numerical values of errors calculated by different methods, from the results of intelligent processing of spatial data are discussed below.

In the preparation of this work, analytical expressions (1-7) are used in the computing, new results for the distributions zSurfT(x,y) and zSurfB(x,y) are discussed, the basic illustrations of which are given in Figures 1-2. Functions defined in the rectangle [xMin,xMax]×[yMin,yMax], xMin = 0, xMax = 4000, yMin = 0, yMax = 2500, zMin=0, zMax=10, xOtk1 = 1600, xOtk2 = 2200, xSkl2 = 2600, zPlt = 10, ugOtk = 0.08, ugSkl = 0.01).

$$\begin{gathered}
\mathrm{fPlt}(x,\mathrm{y}) = \mathrm{zPlt};\ \mathrm{fOtk}(x,z) = \mathrm{zPlt} + \mathrm{Tan}(\mathrm{ugOtk})(x - \mathrm{xOtk1});\ \mathrm{zOtk2} = \mathrm{fOtk}[\mathrm{xOtk2},0];\\
\mathrm{fPrh}(x,z) = \mathrm{fOtk}(x,z) - \mathrm{perkoef}(x - \mathrm{xOtk2})\hat{}2;\ \mathrm{zSkl2} = \mathrm{fPrh}(\mathrm{xSkl2},0);\\
\mathrm{fSkl}(x,z) = \mathrm{zSkl2} + \mathrm{Tan}(\mathrm{ugSkl})(x - \mathrm{xSkl2});\\
\mathrm{zBasicB}(x) = \mathrm{If}[\mathrm{xMin} \le x \le \mathrm{xOtk1},\mathrm{fPlt}(x,0),0] + \mathrm{If}[\mathrm{xOtk1} < x \le \mathrm{xOtk2},\mathrm{fOtk}(x,0),0] +\\
+ \mathrm{If}[\mathrm{xOtk2} \le x \le \mathrm{xSkl2},\mathrm{fPrh}(x,0),0] + \mathrm{If}[\mathrm{xSkl2} \le x \le \mathrm{xMax},\mathrm{fSkl}(x,0),0];
\end{gathered} \tag{1}$$

$$\begin{gathered}
\mathrm{fPltT}(x,\mathrm{y}) = \mathrm{zPltT};\ \mathrm{fOtkT}(x,z) = \mathrm{zPltT} + \mathrm{Tan}(\mathrm{ugOtkT})(x - \mathrm{xOtkT1});\mathrm{zOtkT2} = \mathrm{fOtk}[\mathrm{xOtkT2},0];\\
\mathrm{fPrhT}(x,z) = \mathrm{fOtkT}(x,z) - \mathrm{perkoefT}(x - \mathrm{xOtkT2})\hat{}2;\mathrm{zSklT2} = \mathrm{fPrhT}(\mathrm{xSklT2},0);\\
\mathrm{fSklT}(x,z) = \mathrm{zSklT2} + \mathrm{Tan}(\mathrm{ugSkT1})(x - \mathrm{xSklT2});\\
\mathrm{zBasicT}(x) = \mathrm{If}[\mathrm{xMin} \le x \le \mathrm{xOtkT1},\mathrm{fPltT}(x,0),0] + \mathrm{If}[\mathrm{xOtkT1} < x \le \mathrm{xOtkT2},\mathrm{fOtkT}(x,0),0] +\\
+ \mathrm{If}[\mathrm{xOtkT2} \le x \le \mathrm{xSklT2},\mathrm{fPrhT}(x,0),0] + \mathrm{If}[\mathrm{xSklT2} \le x \le \mathrm{xMax},\mathrm{fSklT}(x,0),0];
\end{gathered} \tag{2}$$

IOP Conf. Series: Journal of Physics: Conf. Series **1425** (2020) 012160 doi:10.1088/1742-6596/1425/1/012160

$$
\begin{gathered}
\mathrm{fH}(x,y)=\cos(\pi x/2)\cdot\cos(\pi y/2),\\
\mathrm{fHill}(x,y)=\begin{cases}\mathrm{fH}(x,y), & -1\le x\le 1\wedge -1\le y\le 1,\\ 0, & \end{cases}\\
\mathrm{fHillXY}(x,y)=\begin{cases}\mathrm{fH}(x,y), & -1\le x\le 1\wedge -1/4\le y\le 1,\\ 0, & -1\le x\le 1\wedge -1\le y\le -1/4,\\ 0, & \end{cases}\\
\mathrm{fHillZz}(x,y)=\begin{cases}\mathrm{fH}(x,y), & -1\le x\le 1\wedge -1\le y\le 1\wedge \mathrm{fH}(x,y)<1/4,\\ 1/4, & -1\le x\le 1\wedge -1\le y\le 1\wedge \mathrm{fH}(x,y)\ge 1/4,\\ 0, & \end{cases}
\end{gathered} \tag{3}
$$

$$
\begin{gathered}
\mathrm{fTrenchA}(x,y)=\begin{cases}-1, & -1/2\le x\le 1/2\wedge -1\le y\le 1/2,\\ 0, & \end{cases}\\
\mathrm{fTrenchB}(x,y)=\begin{cases}-1, & -1/2\le x+k\cdot y\le 1/2\wedge -1\le y\le 1/2\\ 0. & \end{cases}
\end{gathered} \tag{4}
$$

$$
\begin{gathered}
\mathrm{fPyramidA}(x,y)=\begin{cases}1-\max(|x|,|y|), & |x|<1\wedge|y|<1,\\ 0, & \end{cases}\\
\mathrm{fPyramidB}(x,y)=\begin{cases}\max(0,\ 1-|x|-|y|), & |x|<1\wedge|y|<1,\\ 0. & \end{cases}
\end{gathered} \tag{5}
$$

$$
\begin{gathered}
\mathrm{zSurfB}(x,y)=\mathrm{fOriginB}(x,y)+45\mathrm{fHillZz}[0.004(x-400),0.002(y-1900)]+\\
+37\mathrm{fPyramidA}[0.004(x-350),0.002(y-600)]+15\mathrm{fTrenchB}[0.003(x-1100),0.0015(y-400)]+\\
+25\mathrm{fHillXY}[0.003(x-1100),0.0014(y-1700)]+15\mathrm{fHill}[0.005(x-2350),0.003(y-950)-\\
-15\mathrm{fHill}[0.006(x-1900),0.003(y-1300)]-32\mathrm{fHill}[0.0012(x-3000),0.0025(y-2000)]+\\
+35\mathrm{fPyramidB}[0.002(x-3300),0.002(y-650)],\\
\mathrm{fOriginB}(x,y)=\mathrm{zBasicB}(x).
\end{gathered} \tag{6}
$$

$$
\begin{gathered}
\mathrm{zSurfT}(x,y)=\mathrm{fOriginT}(x,y)+42\mathrm{fHillZz}[0.004(x-400),0.002(y-1900)]+\\
+37\mathrm{fPyramidA}[0.004(x-350),0.002(y-600)]+10\mathrm{fTrenchB}[0.004(x-900),0.0015(y-500)]+\\
+15\mathrm{fHill}[0.003(x-1100),0.0014(y-1700)]+15\mathrm{fHill}[0.005(x-2150),0.003(y-1050)-\\
-15\mathrm{fHill}[0.006(x-1900),0.003(y-1300)]-13\mathrm{fHill}[0.0012(x-3000),0.0025(y-2000)]+\\
+25\mathrm{fPyramidB}[0.002(x-3300),0.002(y-650)],\\
\mathrm{fOriginT}(x,y)=\mathrm{zBasicT}(x).
\end{gathered} \tag{7}
$$

In addition, we note that the given analytical expressions (1, 2, 6, 7) are issued from a software module. They are obtained as a result of visual design of surfaces. Perturbations (3 – 5) placed on the base surfaces (1, 2) are automatically "tied" to the base profile. Note that such surfaces (distributions) can be interpreted as the top and bottom of the layer (in particular, aquifer, oil reservoir, sedimentary rock layer).

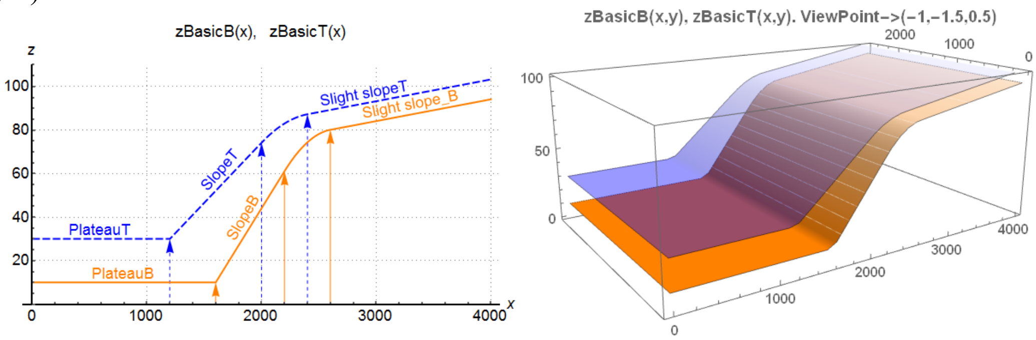

Figure 1. Plot of the basic surface, profile and 3D view

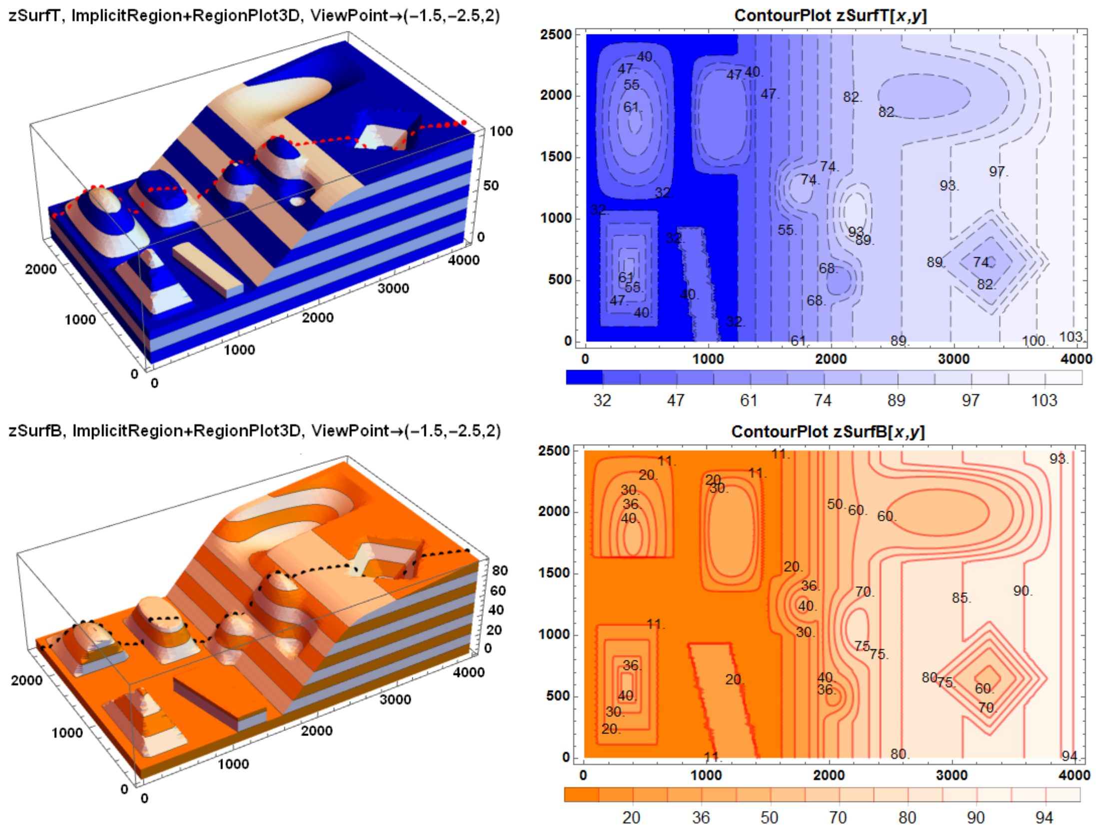

Figure 2. Visualization of the considered reference surfaces

3. Examples of data analysis using artificial neural networks

Below we consider several options for the formation of data sets, which can be interpreted as typical results of observations destined for use in computer geological, geoecological models. Instead of extracting data from the appropriate storages, we will use the approach when we simulate observations and measurements. In this approach, we have a standard, we know how the data are distorted, and the choice of processing method is focused on assessing the accuracy of the original. For the constructed and the above described models, we will generate data for interpretation and analysis by "distorting" the results of level marks on the profiles of reference surfaces by adding "noise", using different random number generators.

Wolfram *Mathematica*, as a system of intelligent computing, provides the user not only the means of mathematical transformations, accurate and approximate calculations, but also machine learning tools. They can be used in the interpretation and processing of input data and computational experiments. The following results and examples of data analysis with artificial neural networks are obtained in the GGMD complex using the corresponding functions that are available in versions 11 and 12 of *Mathematica*.

Let us consider the tools and several applications of the modules “distortion” of the reference model, examples of pre-processing of respectable data, visualization of profiles using the tools of the component GPO. The following describes several options for working with the program module “distortion” of the reference model. For the above examples, using the coordinates and values of the functions in the profile nodes for reference surfaces, we show several methods for simulating observations and obtaining representative data. For this purpose, we use the values calculated from the analytical expressions on the selected profile, to which we add “noise”, using different random number generators ([9, 10]). The following results of computational experiments were performed in one of the typical profile sections, the visualization of which is given in Figure 2. "Measurements are made" at points on surfaces in the direction of (0,2300) – (4000,200). The level values are calculated by formulas (6), (7). One-dimensional graphs illustrating levels of surfaces are shown in Figure 3. Points connected by solid lines indicate the calculated values of the levels on the reference surfaces in the "observation nodes" (for measurements without errors). Circles and diamonds mark data interpreted by observations with errors – they are obtained with additives formed by the corresponding generators of noise. For the bottom surface *zSurfB* (6), the base values at the nodes are distorted using a sample generated by the *Mathematica* RandomVariate (random variable implementation) and Cauchy distribution with location parameter -0.3 and scale parameter 0.8; the marks are shown as circles, the data set is dataProfB. For the top surface *zSurfT* (7), the initial values are distorted using a sample generated by the RandomVariate and UniformDistribution[{-5,10}] functions (a uniform statistical distribution that gives values from -5 to 10); the marks are shown as diamonds, the data set is dataProfT.

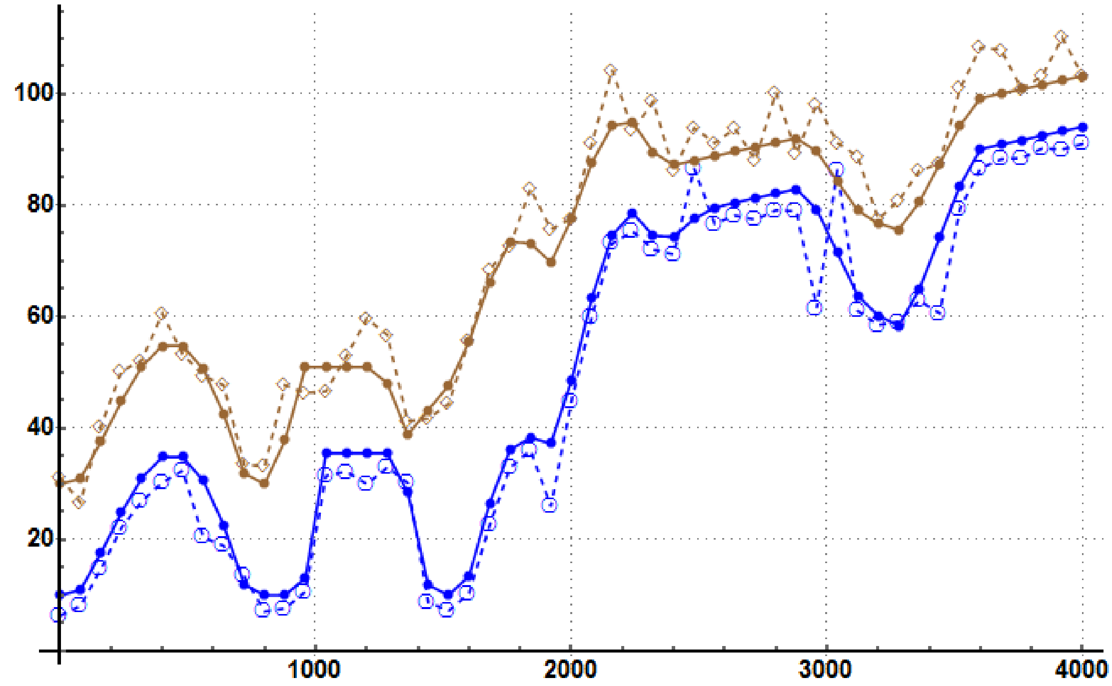

Figure 3. Basic and “observed” data on profiles dataProfT, dataProfB

Data with “distortions” can be used for different purposes, respectively, different methods of pre-processing will be used. The following is important – the mining methods which are discussed below “do not know” the source, the reference functions. The task of the expert is to reproduce (guess) the original according to the available data “measurements”. By the way, the distributions mentioned in the text may not be the original. As noted above, at present many methods and software for data mining are based on the achievements of artificial neural networks.

When solving problems of mathematical modeling the original equations are written in differential form, so the original data (facilities of model) must be continuous, moreover, as a rule, the distribution should be smooth functions. Such requirements are typical not only for models in geology similar examples occur in models of microbiology [11], industrial facilities [12]. In particular, the distribution of the every observed parameter along the profile should be transferred to the initial data of the computer theoretical model as a smooth function. Note that the considered reference data do not satisfy the requirement of smoothness (fHillXY, fHillZz, fTrenchA, fTrenchB), and the data with "distortions" are not suitable for numerical models at all – such raw data will immediately cause a "bumpiness" in the solutions (computational volatility) and the results will be unusable.

Let us consider the options for data preprocessing, implying the obtaining of approximating a smooth function – the smoothing problem ([9, 10]). Several examples are illustrated in Figures 4, 5. Figure 4 shows the results of calculations after training an artificial neural network using the *Mathematica* NetTrain function and specifying the ADAM method. The Adam method is programmed with an optimization algorithm that combines the idea of motion accumulation and the idea of a weaker update of the weights for typical features; the stochastic gradient descent method is implemented using an adaptive learning rate invariant to the diagonal scaling of gradients. Figure 4 presents the results obtained with a minimum number of settings. This is done purposefully, since in this presentation the task is to demonstrate the capabilities as such, rather than detailing the best (adapted) solutions with minimal time for calculations, which for the considered options require only a few tens of seconds. Codes of software module in the implementations of the examples include the assignment of the network netB design NetChain with arguments {vectLength, Tanh, vectLength, Tanh,1}, "Input"–"Scalar", "Output"–"Scalar" and getting dataProfB design NetTrain with arguments netB, dataProfB, Method–"ADAM", MaxTrainingRounds (MaxTrainingRounds is an option for NetTrain and related functions that specifies the maximum number of rounds of training to do).

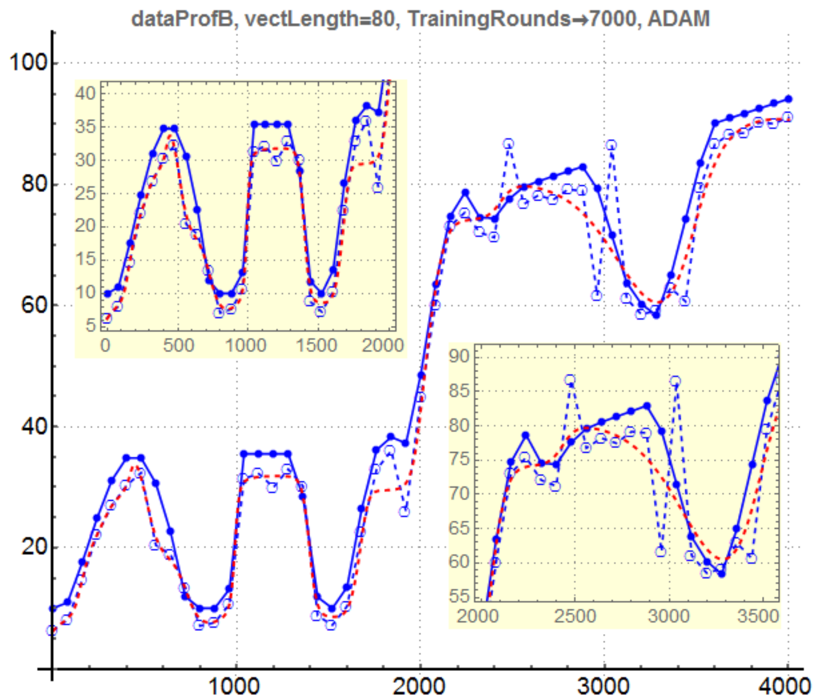

Figure 4. The result of smoothing the dataProfB dataset using a neural network method "ADAM"

The obtained smooth approximation dataProfB shown in Figure 4 the dashed red line. Additionally, in the graphics field, two inserts of the situation plan (copies) are given, which detail the graphs. Calculations were performed with different values of vectLength from the intervals 20-100. For the dataProfB data set, the best results are obtained at vectLength ~ 80. Calculations were also performed using the RMSProp – stochastic gradient descent method using an adaptive learning rate derived from an exponentially smoothed mean gradient. For the profile under consideration, the optimal variant of this method was vectLength ~ 100, the graph is similar to the one shown in Figure 4.

Figure 5 shows the results of calculations after learning artificial neural network using NetTrain method RMSProp for dataProfT data. The smoothed approximation is illustrated by a dashed brown line. Calculations were performed with different values of vectLength. The best results are obtained at vectLength~60. Also the calculations were made with the option method ADAM. For the profile under consideration, the optimal variant in this method was vectLength=110. In method ADAM for considered data, we need significantly less MaxTrainingRounds.

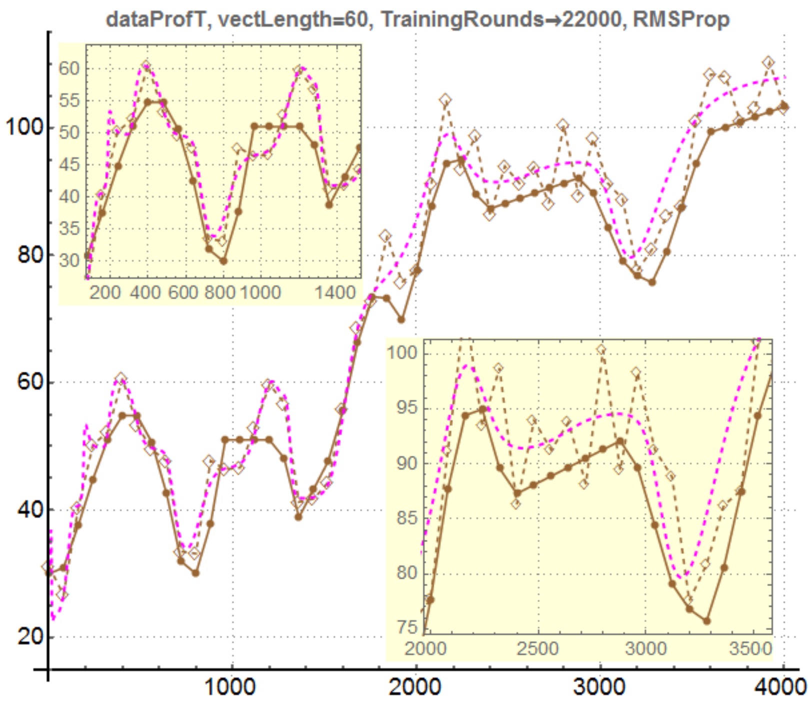

Figure 5. The result of smoothing the dataProfT dataset using a neural network

4. Conclusion

The components of the integrated software complex of the compiler of digital geological models are described. The developed complex provides the possibility of manipulating the source data, intelligent analysis, comparison of interpretations and variants of experts. The discussed results, examples of processing and visualization of spatial data, methods of tuning the tools of artificial neural networks are confirmation of the wide possibilities of the considered technology of intelligent data processing, which were distorted in different ways. However, it should be understood that the training network involves working with a black box, in each case, when obtaining a better reproduction of the model, there is subjectivity, requires a sufficiently large preliminary series of calculations, relatively long calculations. From this point of view, in particular, for data smoothing problems, the use of methods such as moving average, moving median provides repetitive results and is significantly less expensive in calculations.

References

[1] Bedrikovetsky P 2013 Mathematical theory of oil and gas recovery: with applications to ex USSR oil and gas fields (Springer Science & Business Media) p 576

[2] Savinyh V P, Tsvetkov V Ya 2014 Geodannye kak sistemnyi informacionnyi resurs *Vestnik Rossiiskoi akademii nauk* vol **84** no 9 pp 826–829

[3] Tsvetkov V Ya , Lobanov A A 2014 Big Data as Information Barrier *European Researcher* vol **78** no 7-1 pp 1237–1242

[4] Shaitura S V 2015 Intellektual'nyi analiz geodannyh *Perspektivy nauki i obrazovaniya*

[5] Dyshlenko S G 2016Analiz i razrabotka harakteristik kachestva geodannyh *Perspektivy nauki i obrazovaniya* vol **20** №2 pp 23–27

[6] Taranchuk V B 2017 Integrirovannyi programmnyi kompleks testirovschika geologicheskih modelei [Integrated programming complex of the tester of geological models] *Nauchnye*

vedomosti Belgorodskogo gosudarstvennogo universiteta. Seriya: Ekonomika. Informatika vol **43** no 16/265 pp 148–159

[7] Taranchuk V B 2017 The integrated computer complex of an estimation and adapting of digital geological models *Studia i Materiały* vol **2** no 14 pp 73–86

[8] Taranchuk V 2017 The integrated program complex of the composer of geological models *Computer Algebra Systems in Teaching and Research* vol **VI** (University of Natural Sciences and Humanities. Poland) pp 186–203

[9] Taranchuk V 2019 Methods and tools of intelligent data processing for geological models *Computer Algebra Systems in Teaching and Research* vol **VIII** (University of Natural Sciences and Humanities. Poland) pp 147–158

[10] Taranchuk V B 2019 Metody i primery intellektual'noi obrabotki dannyh dlya geologicheskih modelei [Methods and examples of intelligent data processing for geological models] *Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Seriya: Ekonomika. Informatika* vol **46** no 3 pp 511–522 DOI 10.18413/2411-3808-2019-46-3-511-522

[11] Chichurin A, Shvychkina H 2016 Computer simulation of two chemostat models for one nutrient resource *Mathematical Biosciences* vol **278** pp 30–36

[12] Orlov V and Kovalchuk O 2019 Mathematical problems of reliability assurance the building constructions *E3S Web of Conferences* **97** no 03031 p 7