Examples of Integration of Intelligent Computing Modules and the System GeoBazaDannych

Valery B. Taranchuk
Department of Computer Applications and Systems
Belarusian State University
Minsk, Republic of Belarus
taranchuk@bsu.by

***Abstract*— Methodological and technical solutions for the integration of modules of intelligent computing of the Wolfram Mathematica system and tools of the GeoBazaDannych software complex in the tasks of creation, interpretation, processing, visualization of digital fields in computer modeling of objects of geology, geoecology are discussed.**

***Keywords*—system GeoBazaDannych, intelligent adaptation of digital fields, methods of computer model adaptation, clustering**

I. Introduction

Computer modeling includes the development of mathematical methods and algorithms; software development, and computational experiments. The appropriate software should always provide downloading from different sources and preprocessing of data, correlation, formation of digital cubes of object characteristics, interactive data analysis, visualization using graphics. A modification of the typical clustering method is proposed, and the advantages are confirmed by calculations based on representative data.

One of the most important components in the development and implementation, in particular, of constantly operating computer-based geological and geoecological models is the task of assessing the adequacy and accuracy of the proposed digital descriptions. The key issues are automation of the creation, adaptation of models taking into account the constantly incoming additional data, as well as revision of the results of processing the initial information using new interpretation methods [1].

One of the main goals of data mining is the detection of previously unknown, non-trivial, but understandable interpreted knowledge in "raw" (primary) arrays of information.

At the same time, following [2], “data mining does not exclude human participation in processing and analysis, but significantly simplifies the process of finding the necessary data from raw data, making it available to a wide range of analysts who are not specialists in statistics, mathematics or programming. Human participation is expressed in the cognitive aspects of participation and the application of informational cognitive models".

Geodata mining tools are the same as for usual data; the basis is the theory, methods, and algorithms of applied statistics, databases, artificial intelligence, and image recognition.

A number of issues related to the analysis and evaluation of spatial data quality can be solved using the computer system GeoBazaDannych [3].

Possible options, methodological solutions, and software tools that allow you to confirm the validity of interpretations, visualize and obtain numerical values of errors calculated by different methods of intellectual data processing results included and used in computer geological models are discussed below. For illustrations, the key task of forming and processing digital fields used in computer models is selected. In particular, the methods proposed and tested in solving various applied problems are discussed, as well as specialized algorithms for calculating approximating digital fields implemented in the interactive computer complex GeoBazaDannych.

The GeoBazaDannych is the interactive computer complex of intelligent computer subsystems, mathematical, algorithmic and software for filling, maintaining and visualizing databases, input data for simulation and mathematical models, tools for conducting computational experiments, algorithmic tools and software for creating continuously updated computer models. GeoBazaDannych subsystems allow you to calculate and perform expert assessments of local and integral characteristics of ecosystems in different approximations, calculate distributions of concentrations and mass balances of pollutants; create permanent models of oil production facilities; generate and display thematic maps on hard copies [3], [4].

The main components of the GeoBazaDannych: the data generator Gen_DATv; the generator and editor of thematic maps and digital fields Gen_MAPw; the software package Geo_mdl — mathematical, algorithmic

and software tools for building geological models of soil layers, multi-layer reservoirs; the Generator of the geological model of a deposit (GGMD) — the integrated software complex of the composer of digital geological and geoecological models.

We note the main additions to the GeoBazaDannych, including methodological aspects and software components implemented using artificial intelligence tools. At the same time, in order to understand the stages of development and the relationship of the components, we will give examples of the use of updated components of the complex and recall fragments of the results that were discussed and published in the proceedings of OSTIS.

In recent years, the activity of using artificial intelligence tools in solving problems of geology and geoecology has been rapidly increasing. In particular, dozens of articles are published every year on the algorithms and methods of cluster analysis considered in this paper. There are publications that provide solutions to practical problems, methods of preprocessing and interpretation of geophysical data, analysis of results, expert opinions of conclusions and recommendations (for example, [5], [6]).

A number of features of data preparation for computer-based geological and geoecological models from the perspective of the feasibility of using artificial intelligence tools have been regularly discussed at OSTIS conferences since 2019. We will note only the typical difficulties that arise when developing tools and conducting computational experiments for specific practical tasks, which were identified and solutions are presented in subsequent results.

Thus, in the materials of the OSTIS-2019 conference (“Examples of the use of artificial neural networks in the analysis of geodata"), methodological and technical solutions, software tools, results and examples of data processing typical for geophysical methods of studying geological objects, in particular, on observation profiles, were discussed and published. The effectiveness of the use of artificial neural networks to eliminate noise and errors in the measurement results, perform the necessary data preprocessing for mathematical models by smoothing in order to prepare regular digital distributions is illustrated.

In the materials of the OSTIS-2020 conference (“Examples of intelligent adaptation of digital fields by means of the GeoBazaDannych system“ and ”Interactive Adaptation of Digital Fields in the GeoBazaDannych System"), examples of interactive formation of digital models of geological and geoecological objects in computational experiments that meet the intuitive requirements of an expert are considered and given. Methodological and algorithmic solutions effective in processing remote environmental monitoring data, special tools of the GeoBazaDannych system are noted, the results of interactive adaptation and comparison with standard reference solutions in the complex "Generator of the geological model of the deposit" are presented.

The examples show that this way you can significantly improve the quality and adequacy of the digital description. But you need to understand that at that stage of the state of the algorithmic and software complex of GeoBazaDannych, the allocation itself is implemented according to the intuitive suggestions of the expert.

The issues of automatic identification of sites of the “highlighted” type using cluster analysis tools integrated into the GeoBazaDannych system were discussed at the OSTIS-2022 conference; some positive results were published in the proceedings of the conference and in [7]. In particular, the results illustrating the effects of choosing and confirming the best clustering algorithms are presented, and the options for using different clustering methods are compared, moreover, for different ways of setting the metric distance. In the above-mentioned published materials of OSTIS-2022, along with the positive ones, the disadvantages of the described software implementations and the settings used were noted.

A number of additions to the GeoBazaDannych complex, other variations of the settings of the Wolfram Mathematica system tools for data clustering have been tested in calculations, visualized and described below. Nevertheless, it should be understood that further research, improvement of algorithms and modification of software tools are needed.

II. Initial data, a reference distribution for computational experiments

The results below cannot be strictly mathematically justified, but are indicative and adequate due to the rules of their preparation. On the one hand, they are being formed by random number generators and refined by an expert in order to give them the character of comparable data from field observations in the practice of geophysical methods for studying geological, geoecological objects On the other hand, accepted mathematical expressions are used for measurement values in observations (in calculations, this is an imitation). In fact, for each particular processing algorithm, it is known that the basic (reference) digital distribution can be calculated with the necessary accuracy — comparing calculations with the standard, one can judge the advantages and disadvantages of the method used. Below, the reference distribution (in the accepted GeoBazaDannych terminology, the surface) differs slightly from that considered in [4], the location of the fragments of disturbances and their shape and dimensions have been changed for some. Generally speaking, there was no need for changes, because the set of fragments of typical relief elements in [4] was quite representative, but calculations were performed specifically for a reference surface of a different shape to ensure

stable reproduction of the qualitative properties of the results obtained. All the calculation variants described below were performed for both reference distributions, they confirmed that the results are qualitatively the same; they do not change with variations in the size, position, orientation of perturbations.

The results presented in this paper are obtained using the numerical values of the level marks of the (reference) surface according to the formula (1):

$$\begin{gathered} zSurfH(x,y) = fOriginF(x,y)+ \\ +400 \cdot fHill6(0.005 \cdot (x-250), 0.007 \cdot (y-400))+ \\ +600 \cdot fHill3(0.01 \cdot (x-150), 0.01 \cdot (y-150))- \\ -200 \cdot fHill(0.01 \cdot (x-880), 0.015 \cdot (y-500))- \\ -150 \cdot fHill(0.02 \cdot (x-920), 0.004 \cdot (y-100))+ \\ +200 \cdot fHill5(0.006 \cdot (x-450), 0.001 \cdot (y-150)), \\ fOriginF(x,y) = zBasicF(x). \end{gathered} \tag{1}$$

The visualization of the zSurfH reference surface is shown in Figures 1 and 2, where 3D views are shown in the surface and volume variants; Figure 2 shows a map of isolines. Additionally, the numbers of fragments of disturbances are added to the image on the contour map (in parentheses in blue).

The corresponding scheme of their placement is shown in Figure 3, where the isolines of the reference surface and the one reconstructed in Wolfram Mathematica are also given (the Interpolation method, InterpolationOrder = 1).

The initial data for demonstrating methods and algorithms of mining and clustering were obtained using a random number generator, in which the following were set: the number of observation profiles, points on each profile, and coordinates of the beginning and end of the profile were generated in the specified ranges of values. The values in the points were calculated using the formula (1) — simulation of measurements of the level of the surface being restored. Note that in fact we have a scattered set of points.

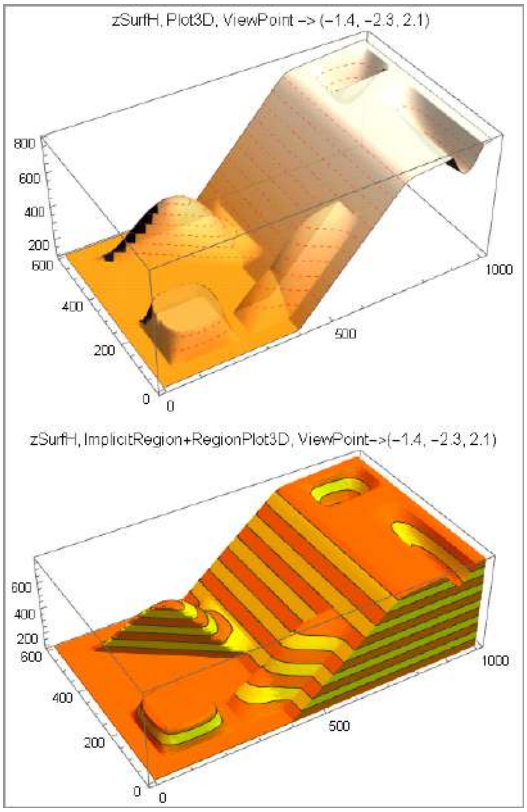

Figure 1. 3D views of the zSurfH reference surface.

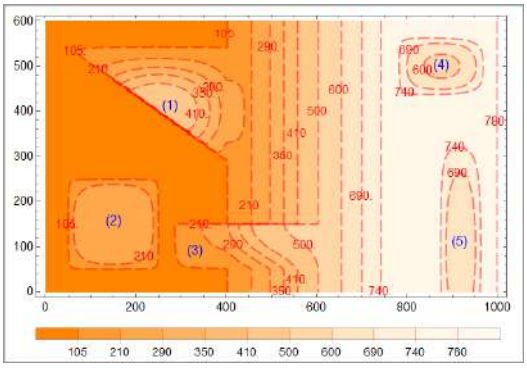

Figure 2. Contour map of the reference surface zSurfH.

III. Illustrations, comparison of the results of the refined clustering methods

Cluster analysis allows for many different types of clustering techniques/algorithms to determine the final result [8], [9]. We note the development, a new way of processing data, and present the results of the comparison Without going into the details of the algorithmic, software implementation, we recall that in [7] clustering was actually performed according to two parameters, namely, grouping was carried out according to the criterion of proximity of points with measurements. In the presented results, grouping is performed according to a combination of three values, namely, for each point, their coordinates (x,y) and the z value (surface level) were taken into account. Here are several interpretations of representative calculation results using the Wolfram Mathematica FindClusters function with different criteria of Criterion Function, we will explain the illustrations.

The schemes in Figure 4 repeat the graphic layers from the maps above, they are useful for interpretation and explanation. The schema at the top details the reference, clearly marked sections of fragments-perturbations. The schema below clearly shows areas where there is no reproduction of the standard, which is explained by the lack of measurements. The contour map with data points at the bottom is useful for understanding that in some areas the field cannot be restored to the standard

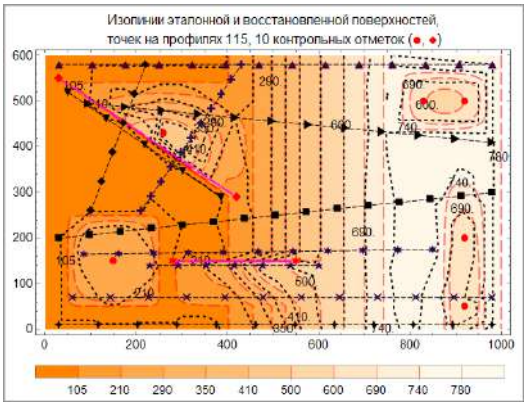

Figure 3. A scheme of the measurement points of the levels, a map of the isolines of the reference and reconstructed surfaces.

because there are no measurements. It is clear that in parts of the area where the digital field differs from the reference, classification / binding to a fragment-disturbance is unlikely.

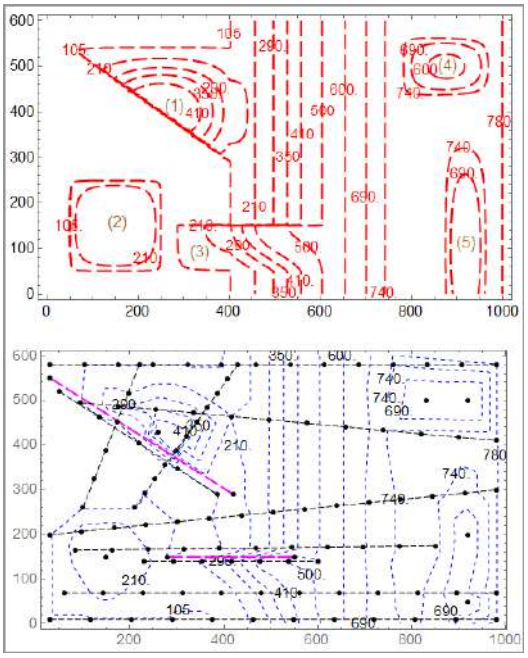

Figure 4. Schemas for understanding the differences between the processed data and the reference.

Figure 5 shows the results of clustering by two parameters (rProfXY) in the upper part, and by three (rProfXYZ) at the bottom. The upper illustration shows the results of calculations with the settings of the KMeans method, as in [7]. Below is the current implementation for the same method; in both versions, the metric is "default", the number of clusters is 6.

The proposed clustering method for three parameters is clearly preferable to the method for two, in particular, in terms of localization of fragments 1, 4 and 5. The

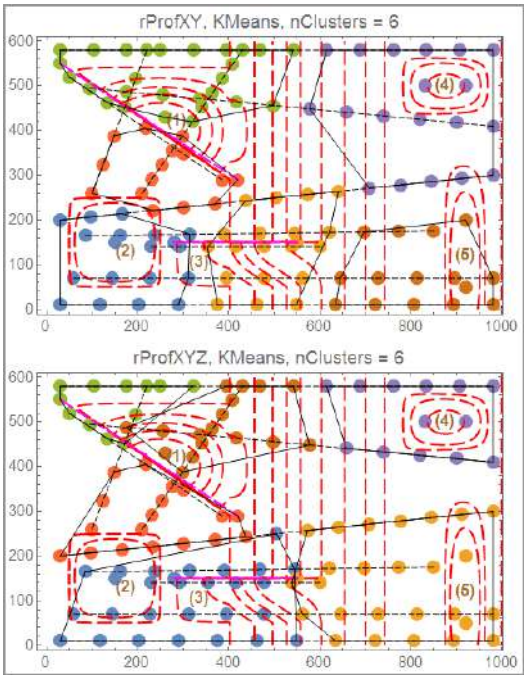

Figure 5. Clustering results for two (rProfXY) and three (rProfXYZ) parameters.

results near fragments 2 and 3 are not indicative due to the discontinuous distribution on different sides of the horizontal dotted magenta line (projection onto the horizontal plane of the fragment section-perturbation 3 by the vertical plane).

Such situations (gaps, offset) are not identified at all in conventional digital field restoration systems without special a priori additional conditions. Note that in the GeoBazaDannych, such conditions can be set interactively by correcting the initial information on the map [3].

The figure shows the results of the variant when the number of clusters is set to 6. Why. Determining the number of clusters is one of the most important segmentation problems. In a broader sense, this is the problem of initializing the algorithm. The results of the variants for the number of clusters from 4 to 9 were calculated and compared. According to the results of the comparison, the variant of 6 clusters seems preferable, and the explanation for this may probably be that the distribution is reproduced when there is a basic continuous surface-a ribbon and 5 fragments-perturbations, i. e. 6 different shapes.

A. The impact of the clustering method

The variants of clustering results using different methods and metrics are illustrated in Figures 6 and 7, the names of the methods and metrics are written in the headings of the diagrams The corresponding software application included in the GeoBazaDannych from the Wol-

fram Mathematica system allows variants of the clustering method (Criterion Function): Automatic, Agglomerate, DBSCAN, GaussianMixture, JarvisPatrick, KMeans, KMedoids, MeanShift, NeighborhoodContraction, Optimize, SpanningTree, Spectral [10]. What segmentation methods are used in the calculations are written in the headings of the diagrams. Representative clustering options are shown, namely K Means (k-means clustering algorithm), k-medoids (partitioning around medoids), Optimal (Wolfram Mathematica method). The effects of the accepted clustering method (Possible settings for Method) are illustrated by the schemes in Figure 6. Clustering in the examples of this series was considered for three parameters, the FindClusters function was used, the norm in the examples of the series in Figure 6 was not set, but was determined by the default calculation module. These results are quite indicative. At the same time, taking into account the reference and the digital field of the original, we can consider the clustering options by the KMeans and Optimal methods as preferable.

B. The impact of the metric

The issues of measuring the proximity of objects have to be solved with any interpretation of clusters and various classification methods, moreover, there is an ambiguity in choosing the method of normalization and determining the distance between objects. The influence of the metric (DistanceFunction) is illustrated by the diagrams in Figure 7. The results presented in this series are obtained by means of the corresponding software application included in the GeoBazaDannych from the Wolfram Mathematica, which allows different options for setting DistanceFunction.

The Wolfram Language provides built-in functions for many standard distance measures, as well as the capability to give a symbolic definition for an arbitrary measure. In particular, the following metric variant sare available for analyzing digital data [10]: EuclideanDistance, SquaredEuclideanDistance, NormalizedSquaredEuclideanDistance, ManhattanDistance, ChessboardDistance, BrayCurtisDistance, CanberraDistance, CosineDistance, CorrelationDistance, BinaryDistance, WarpingDistance, CanonicalWarpingDistance. What methods of DistanceFunction are used in calculations is recorded in the headers of the schemes. Representative variants are shown, namely: EuclideanDistance (the length of a line segment between the two points), ChessboardDistance, SquaredEuclideanDistance, BrayCurtisDistance, ChebyshevDistance (a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension), ManhattanDistance. It follows from the above results that for the considered configuration of data points, taking into account the digital field of the original, clustering options using Spectral EuclideanDistance methods can be considered preferable.

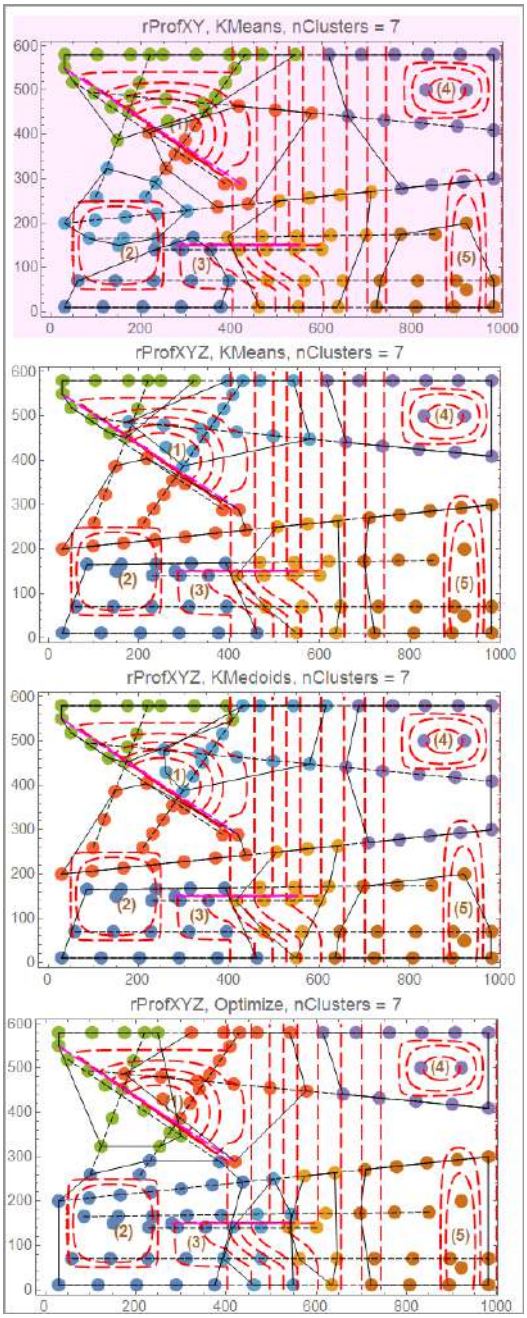

Figure 6. The effects of the accepted clustering method.

IV. Conclusion

The issues of instrumental filling and use of the interactive computer system GeoBazaDannych, expansion of its functionality through integration with the Wolfram Mathematica computer algebra system are considered. A modification of the typical clustering method is proposed, and computational experiments have confirmed the advantages in comparison with traditional methods.

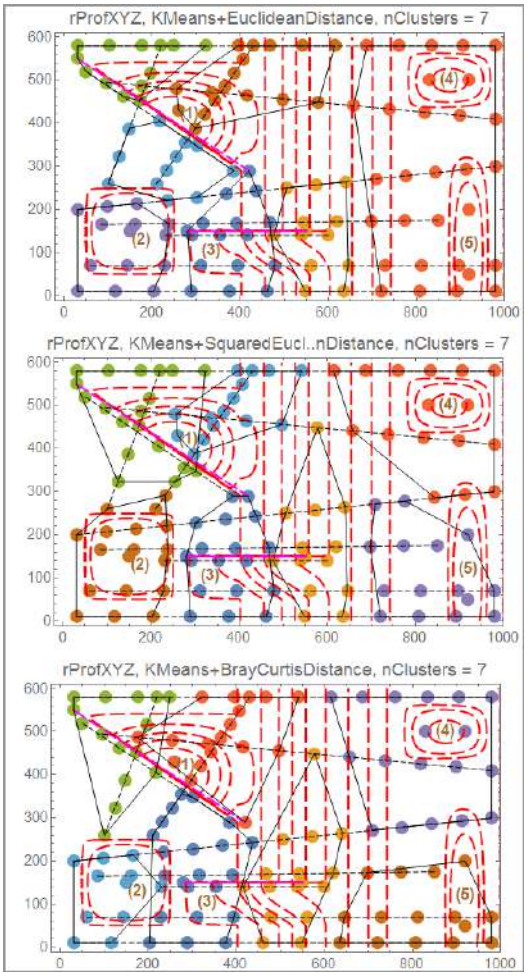

Figure 7. The effects of the accepted metric (EuclideanDistance, SquaredEuclideanDistance,BrayCurtisDistance).

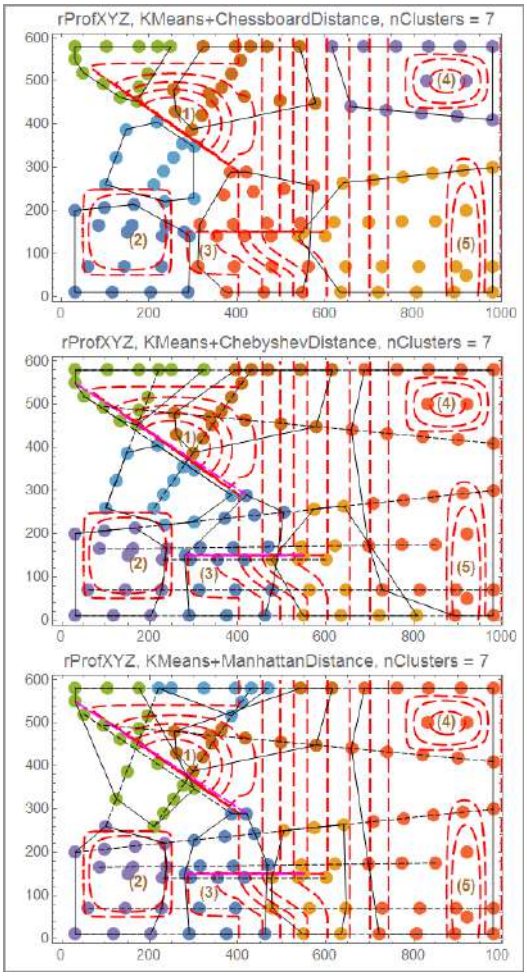

Figure 8. The effects of the accepted metric (ChessboardDistance, ChebyshevDistance, ManhattanDistance).

References

[1] V. P. Savinyh, V. Ya. Cvetkov, "Geodannye kak sistemnyi informacionnyi resurs", Vestnik Rossiiskoi akademii nauk, vol. 84, no. 9, pp. 826–829, 2014, (in Russian).

[2] V. Golenkov, N. Guliakina, and D. Shunkevich, Otkrytaja tehnologija ontologicheskogo proektirovanija, proizvodstva i jekspluatacii semanticheski sovmestimyh gibridnyh intellektual'nyh komp'juternyh sistem [Open technology of ontological design, production and operation of semantically compatible hybrid intelligent computer systems], V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2021. 690 p.

[3] V. Taranchuk, Komp'yuternye modeli podzemnoi gidrodinamiki. Minsk: BGU, 2020. 235 p., (in Russian)

[4] V. Taranchuk, "Interactive Adaptation of Digital Fields in the System GeoBazaDannych", Communications in Computer and Information Science. Book series Springer, vol. 1282, –2020, P. 222–233.

[5] M. Abdideh, A. Ameri, "Cluster Analysis of Petrophysical and Geological Parameters for Separating the Electrofacies of a Gas Carbonate Reservoir Sequence. Nat Resour Res 29, 1843–1856 (2020). https://doi.org/10.1007/s11053-019-09533-1

[6] M. A. Tkachenko, Y. V. Karelina "A cluster analysis as a method of identifying potential chromite mineralisation within the Voykar-Synyinsky massif". International Research Journal. — 2023. — 10 (136). DOI: 10.23670/IRJ.2023.136.63

[7] V. Taranchuk, "Integration of computer algebra tools into OSTIS applications". Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2022, no 6, pp. 369-374.

[8] D. Tupper Charles, "Concepts of Clustering, Indexing, and Structures", Data Architecture, 2011, pp. 241–253, https://doi.org/10.1016/B978-0-12-385126-0.00013-9.

[9] B. S. Everitt, S. Landau, M. Leese, D. Stahl. Cluster Analysis. 5th Edition, John Wiley & Sons, 2011, 360 p.

[10] Distance and Similarity Measures. https://reference.wolfram.com/language/ guide/ DistanceAndSimilarityMeasures.html/ (accessed 2024, Feb).

ПРИМЕРЫ ИНТЕГРАЦИИ МОДУЛЕЙ ИНТЕЛЛЕКТУАЛЬНЫХ ВЫЧИСЛЕНИЙ И СИСТЕМЫ ГЕОБАЗАДАННЫХ

Таранчук В.Б.

Обсуждаются методические и технические решения интеграции модулей интеллектуальных вычислений системы Wolfram Mathematica и инструментов программного комплекса ГеоБазаДанных в задачах формирования, интерпретации, обработки, визуализации цифровых полей при компьютерном моделировании объектов геологии, геоэкологии. Предложена модификация типового способа кластеризации, расчетами на представительных данных подтверждены преимущества.

Received 04.04.2024