Interactive and intelligent tools of the GeoBazaDannych system

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

***Abstract*—In this paper, from the perspective of creating and maintaining geological or geoecological models, methodological and technical issues, ways of developing the system GeoBazaDannych [1], expanding its functionality by including data mining modules of the Wolfram Mathematica computer algebra system are considered. In particular, the examples illustrate the tools for preparing control sets of geodata for validation, testing and evaluation of related neural network models. It is shown how the adopted architecture, the implemented concept of constructing the system GeoBazaDannych, allow expanding the functionality by including additional software components. Examples illustrate the variants for choosing the best clustering algorithms.**

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

I. Introduction

The features of solving the problems of developing and implementing computer-based geological and geoecological models with the means of their adaptation and self-adjustment, the main approaches to processing, analysis, interpretation of the data used and obtained are noted in [2] – [5]. The mentioned publications provide several basic solutions to the issues of preprocessing, intelligent analysis of geodata by means of the computer system GeoBazaDannych. It is emphasized that at this stage, data mining is among the priority areas of research and development, the corresponding classes of systems for its implementation are listed. The results and methodological recommendations of cluster analysis of geodata obtained with the environment of the system GeoBazaDannych are discussed below.

II. General information about cluster analysis

The solution to the problem of cluster analysis (segmentation) [6], [7] is the partitions that satisfy the accepted criterion. The criterion is usually a functionally formalized set of rules for determining the levels of differences in partitions and groupings (the objective function). In data mining, segmentation can be used as an independent tool for making decisions about data distribution, for monitoring characteristics and subsequent analysis of data sets of certain clusters. Alternatively, cluster analysis can serve as a preprocessing stage for other algorithms. Segmentation is also used to detect atypical outlier objects (values that are “far” from any cluster), in other words, it is a novelty detection, such objects may be more interesting than those included in clusters. In addition, cluster analysis, unlike most mathematical and statistical methods, does not impose any restrictions on the type of source data under consideration. An important advantage of cluster analysis is that when it is performed, it is possible to divide objects not only by one parameter, but by a set of features. In addition, cluster analysis, unlike most mathematical and statistical methods, does not impose any restrictions on the type of source data under consideration. It is well known that cluster analysis is widely used in many fields, in particular, in computer systems for pattern recognition, image analysis, information retrieval, data compression, computer graphics, bioinformatics, machine learning. The following are representative examples and the cluster analysis tools implemented in the system GeoBazaDannych environment are noted.

III. Brief information about the software system GeoBazaDannych

The interactive computer system GeoBazaDannych is the 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’s 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. The main components of the system GeoBazaDannych [1]:

- the data generator Gen_DATv;
- the generator and editor of thematic maps and digital fields Gen_MAPw;

- modules for organizing the operation of geographic information systems in interactive or batch modes;
- the software package Geo_mdl – mathematical, algorithmic and software tools for building geological models of soil layers, multi-layer reservoirs; modules for three-dimensional visualization of dynamic processes of distribution of water-soluble pollutants in active soil layers;
- software and algorithmic support for the formation and maintenance of permanent hydrodynamic models of multiphase filtration in porous, fractured media;
- the integrated software complex of the composer of digital geological and geoecological models (GGMD).

IV. What is the novelty of the presented results

To explain the novelty of the results presented in this paper, we note that [4], [5] provide examples of interactive formation of digital models of geological objects in computational experiments that meet the intuitive requirements of the expert. Examples of approximation and reconstruction of the digital field, its interactive adaptation by means of the system GeoBazaDannych were discussed. The examples of approximation and reconstruction of the digital field, its interactive adaptation by means of the system GeoBazaDannych and evaluation of the accuracy of results using the tools of the GGMD complex illustrate the unique capabilities of the developed methods and software. In [2], [3], the results of the use of artificial neural networks in the analysis and interpretation of geospatial data are presented and discussed, the possibilities of obtaining and visualizing errors are described. This paper discusses variants and provides tools for implementing cluster analysis of geodata in the environment of the system GeoBazaDannych; recommendations are given for choosing the optimal parameters of classification algorithms when dividing the studied objects and features into groups that are homogeneous in the accepted sense. Particular additions are discussed and illustrated with examples.

It is important that the corresponding additions to the system GeoBazaDannych, new instrumental content are implemented within the framework of the concept of computer model development adopted and actively developed in recent years (see, for example, [1], [8]), the basis of which is integration into software packages, complexes of modules of computer algebra systems, ensuring functioning in one environment, a single interface of current software modules and extensions.

V. Tools, examples of cluster analysis of geodata

The examples below are calculated with the data [3], from the two surfaces considered there, *zSurfB* is selected for illustrations. Recall that the simulated surface (a reference for evaluating the accuracy of numerical experiments, approximate calculations by various methods) has a complete mathematical description, for clarity, Fig. 1 shows the isolines (contour lines) of the *zSurfB* levels.

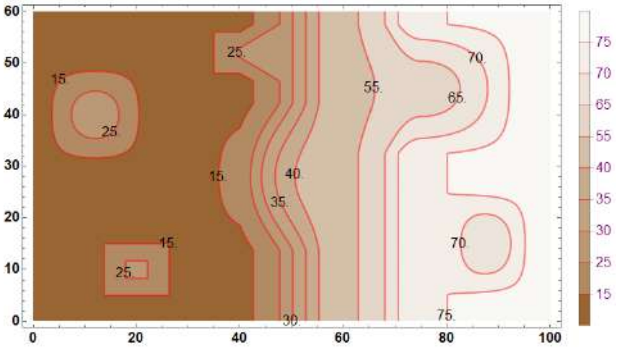


Figure 1. Contour map of the reference surface *zSurfB*.

The corresponding scheme of their placement is shown in Fig. 2, where the isolines of the reference surface and the one reconstructed in Wolfram Mathematica are also given (the Interpolation method, InterpolationOrder = 1). Data for demonstrations of methods and algorithms of intellectual analysis are obtained by simulation of measurements, the corresponding data set – the points of measurements of the level of the restored surface, representing (in fact) a scattered set of points, are interpreted as data on observation profiles.

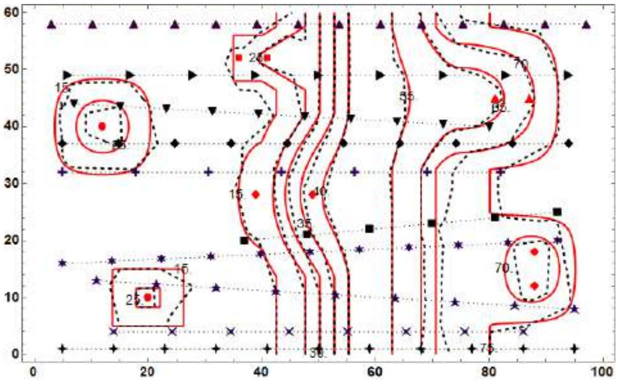


Figure 2. A scheme of points with level measurements, a map of isolines of the reference and reconstructed surfaces.

VI. Effects of the accepted clustering method

Cluster analysis allows for many different types of clustering techniques/algorithms to determine the final result. Determining the number of clusters is one of the most important problems of segmentation. In a broader sense, this is the problem of initializing the algorithm:

selection of optimal values of control parameters, evaluation functions used, metrics, stopping conditions, etc. In the examples below, a priori information is used, the number of clusters is initially set to 7. Why so much – it is taken into account that in the initial data, measurements were carried out for a surface that included: the base surface and 6 different distortions of it with individual positioning of perturbations. In the illustrations (to remind the data source), the isolines of the reference surface are given in red dotted lines. Below are the results that illustrate the features of the most commonly used clustering algorithms.

The effects of the accepted clustering method are illustrated by the schemes in Fig. 3. Clustering in the examples of this series was considered only for pairs of coordinates, i.e. the relative position of the points of the scattered set was taken into account, moreover, the FindClusters function with different criteria was used in the program module, the norm in the examples of the series Fig. 3 was calculated using the DistanceFunction EuclideanDistance metric.

Generally speaking, the corresponding software application included in the system GeoBazaDannych from the Wolfram Mathematica allows variants of the clustering method (Criterion function): Automatic, Agglomerate, DBSCAN, Gaussian mixing, Jarvispatrick, KMeans, KMedoids, Neighborhood, Optimization, SpanningTree, Spectral [9]. What segmentation methods are used in the calculations are recorded in the headers of the diagrams. Representative clustering options are shown, namely Automatic (Wolfram Mathematica automatically selects the method, the Wolfram Language will automatically try to pick the best method for a particular computation), k-means (k-means clustering algorithm [10]), k-medoids (splitting into medoids [11]), Spectral (spectral clustering algorithm [12]). 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 option by the Spectral method as preferable.

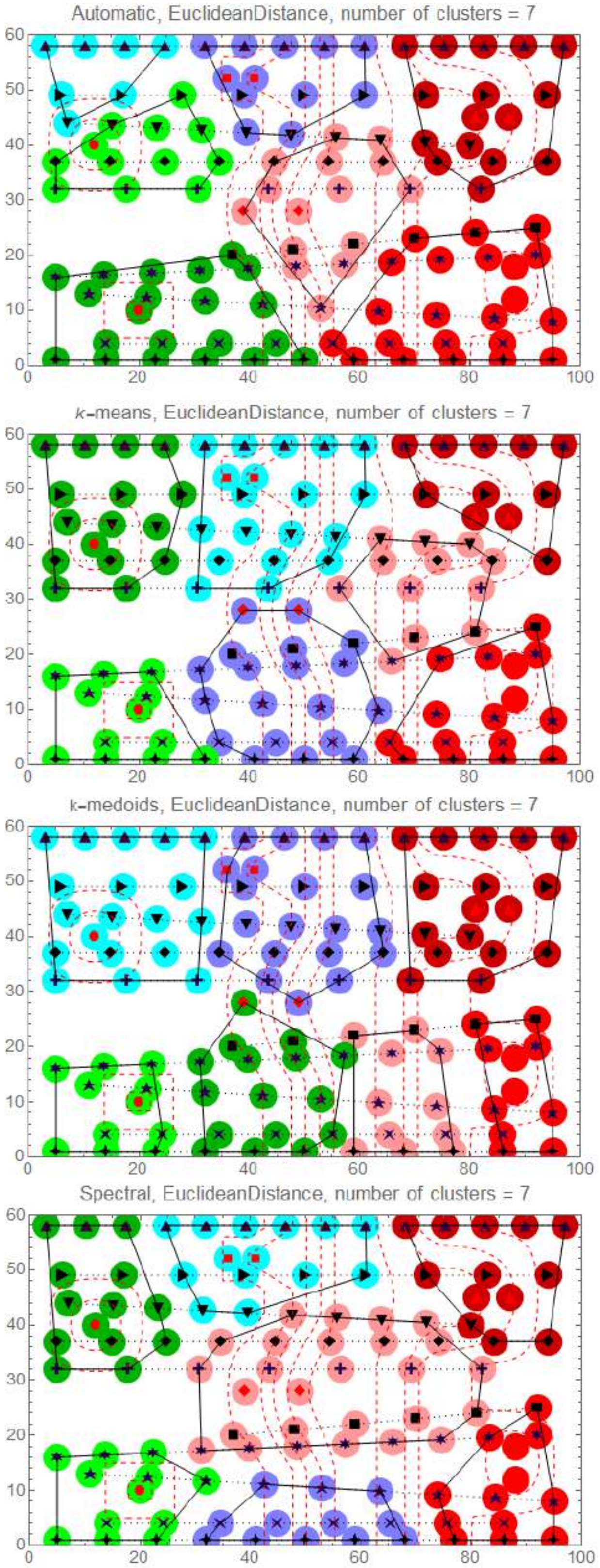


Figure 3. Clustering methods.

VII. The impact of the metric

In the examples discussed above, as well as in this series of results, the similarity or difference between the classified objects is established depending on the metric distance between them. 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 Fig. 4. 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 (Possible settings for Method). In the Wolfram

Mathematica system, different measures of distance or similarity are convenient for different types of analysis. 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 [13]. The algorithmic features of the listed metrics can be clarified in the articles [14], [15]. As in the examples above, clustering algorithms were considered only for pairs of coordinates, i.e. the relative position of the points of the scattered set was taken into account, the Spectral method was used.

What methods of DistanceFunction are used in calculations is recorded in the headers of the schemes in Fig. 4. Representative variants are shown, namely ChessboardDistance [16], CosineDistance (a measure of similarity between two non-zero vectors of an inner product space), ChebyshevDistance (a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension), EuclideanDistance (the length of a line segment between the two points):

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.

VIII. Influence of the number of clusters

As noted above, one of the most important problems of segmentation is determining the number of clusters. The series of illustrations in Fig. 5 shows the results calculated by Spectral EuclideanDistance methods with the number of clusters 6 and 8; 7 clusters are shown in Fig. 4.

IX. The effect of accounting for values in points

In the results considered and shown in Fig. 3, Fig. 4 and Fig. 5, the similarity or difference between the classified objects is established depending on the metric distance between them. In other words – in the results presented in this series, the algorithms take into account not pairs *Xi*, *Yi*, but triples – *Xi*, *Yi*, *Zi*. Fig. 6 shows classification options using the Wolfram Mathematica ClusterClassify function (use data to create a function to classify new data into clusters), which allows clustering not only taking into account the coordinates of the points of the scattered set, but also the values in them.

From the above results, it follows that for the data class under consideration, taking into account the values at points does not give an additional positive effect in the implementation of clustering.

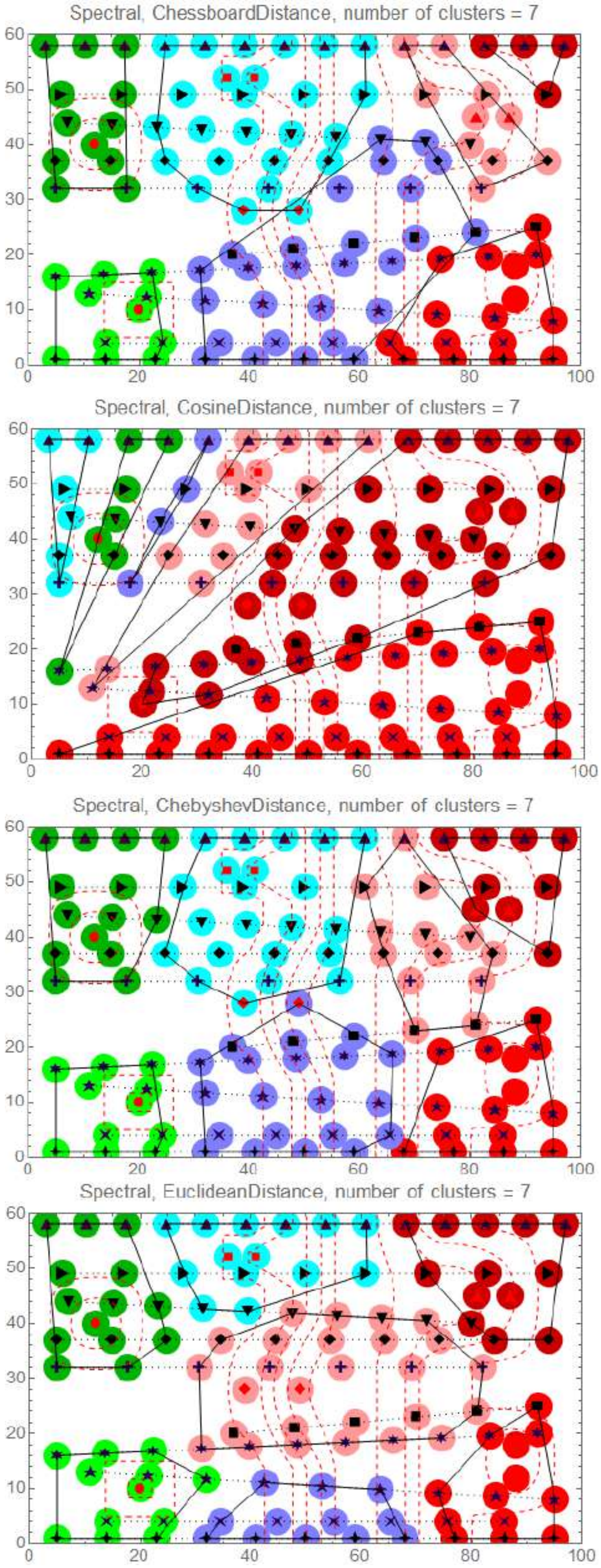
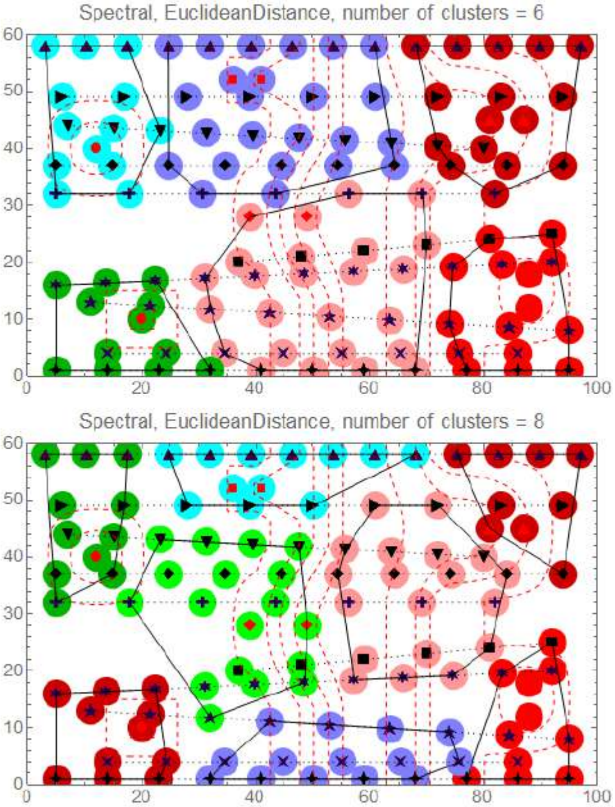
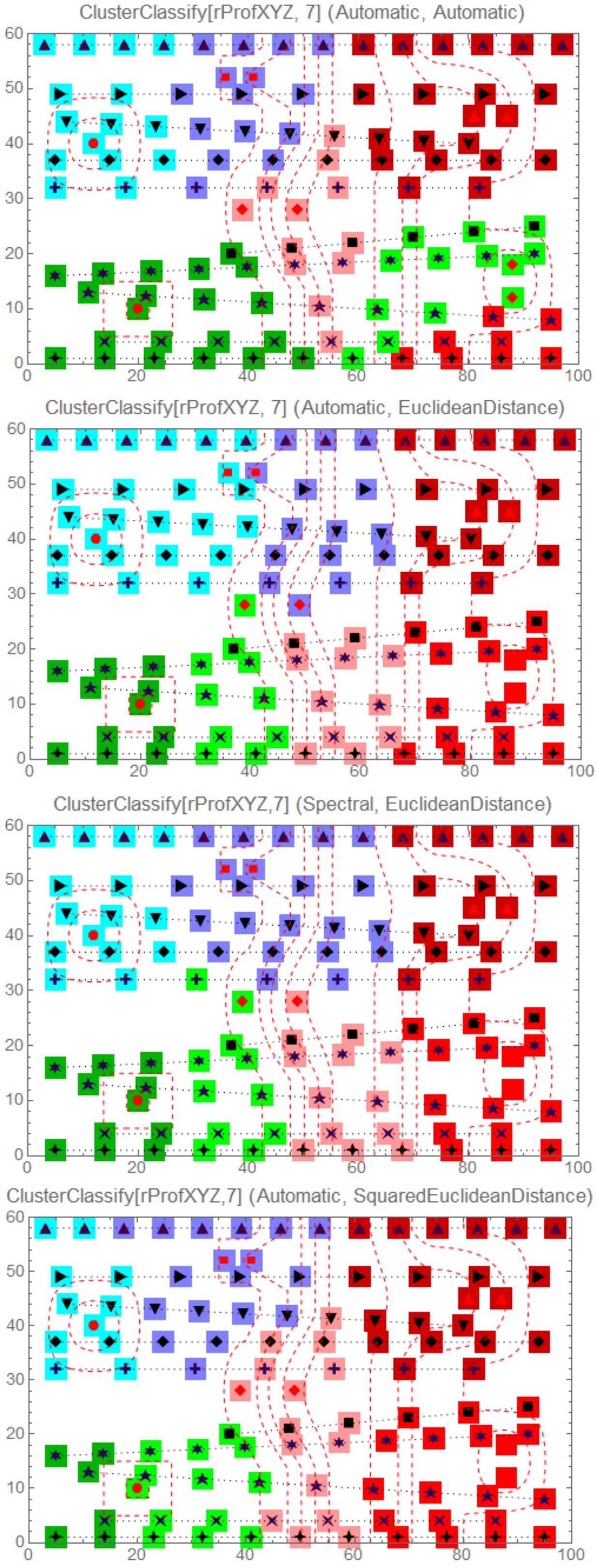


Figure 4. Influence of DistanceFunction.


Spectral, EuclideanDistance, number of clusters = 6

Spectral, EuclideanDistance, number of clusters = 8

Figure 5. Influence of the number of clusters.


ClusterClassify[rProfXYZ, 7] (Automatic, Automatic)

ClusterClassify[rProfXYZ, 7] (Automatic, EuclideanDistance)

ClusterClassify[rProfXYZ,7] (Spectral, EuclideanDistance)

ClusterClassify[rProfXYZ,7] (Automatic, SquaredEuclideanDistance)

Figure 6. Influence of DistanceFunction.

X. Conclusion

The article deals with the issues of instrumental filling and the use of the interactive computer system GeoBazaDannych. The results of clustering of a representative data set of a typical model of a geological object are presented and discussed.

References

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

[2] V. B. Taranchuk, "Examples of the use of artificial neural networks in the analysis of geodata", Open semantic technologies for intelligent systems, vol. 3, pp. 225–230, 2019.

[3] V. Taranchuk, "Tools and examples of intelligent processing, visualization and interpretation of GEODATA", Modelling and Methods of Structural Analysis. IOP Conf. Series: Journal of Physics: Conf. Series Vol. 1425 (2020) 012160. – P. 9.

[4] V. B. Taranchuk, "Examples of intelligent adaptation of digital fields by means of the system GeoBazaDannych", Open Semantic Technologies for Intelligent Systems, vol. 4, pp. 243–248, 2020.

[5] V. Taranchuk, "Interactive Adaptation of Digital Fields in the System GeoBazaDannych", Communications in Computer and Information Science. Book series Springer, vol. 1282, 2020, pp. 222–233, https://doi.org/10.1007/978-3-030-60447-9_14.

[6] 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.

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

[8] V. V. Sobchuk, O. V. Chichurin, I. V. Kalchuk , T. V. Zhigallo. Solving problems of analysis and Differential Equations by means of computer algebra Mathematica: textbook. Lutsk: Volyn National Academy of Sciences. Lesya Ukrainka University, 2021, 414 p., (in Ukrainian).

[9] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek, “Density-based Clustering”, WIREs Data Mining and Knowledge Discovery, 2011, 1 (3): 231–240. doi:10.1002widm.30.

[10] H. Bock, “Clustering methods: a history of k-means algorithms”, in: Selected contributions in data analysis and classification. Springer, Berlin, 2007, pp. 161–172.

[11] H. Park, C. Jun, “A simple and fast algorithm for k-medoids clustering”, Expert Syst. Appl., 2009, 36(2), 3336–3341.

[12] U. von Luxburg, M. Belkin, O. Bousquet, “Consistency of Spectral Clustering”, Annals of Statistics. 2008, 36(2), 555–586.

[13] Distance and Similarity Measures. https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures, (accessed 2021, Jun).

[14] E. Amigó, J. Gonzalo, J. Artiles, et al., “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, Inf Retrieval 12, 2009, 461–486. https://doi.org/10.1007/s10791-008-9066-8.

[15] P. Grabusts, “The Choice of Metrics for Clustering Algorithms”, Environment Technology. Resources, DOI:10.17770/etr 2011vol 11 70–76.

[16] W. Zhu, C. Ma, L. Xia and X. Li, “A Fast and Accurate Algorithm for Chessboard Corner Detection”, 2009, pp. 1-5, doi: 10.1109/CISP.2009.5304332.

Интерактивные и интеллектуальные средства системы ГеоБазаДанных

В.Б. Таранчук

В статье рассматриваются вопросы инструментального наполнения и использования интерактивной компьютерной системы ГеоБазаДанных. Представлены и обсуждаются результаты кластеризации представительного набора данных типичной модели геологического объекта.

Received 31.05.2021