



Methodological and Technical Solutions for the Implementation of Clustering Algorithms in the GeoBazaDannych System

Valery Taranchuk^(✉) 

Belarusian State University, Nezavisimosti avenue, 4, 220030 Minsk, Belarus
Taranchuk@bsu.by

Abstract. In this paper, from the perspective of creating and maintaining geological or geocological models, methodological and technical issues, ways of developing the system GeoBazaDannych (GBD), expanding its functionality are considered. The new functionality provided by the inclusion of executable data mining modules of the Wolfram Mathematica computer algebra system into the GBD is noted. In particular, it illustrates tools for preparing benchmark geodata sets for validating, testing, and evaluating related neural network models. Examples on representative data sets illustrate options for choosing the best clustering algorithms.

Examples of representative datasets illustrate the options for choosing the best clustering algorithms. A series of calculations illustrate the effects of choosing the number of clusters, the clustering method, and metrics. Separately, options are considered when clustering takes into account not only the coordinates of observation points, but also the values in them.

Keywords: System GeoBazaDannych · Intelligent adaptation of digital fields · Clustering

1 Introduction

Digital geological, geocological 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. 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 geocological models are a representative part of the geodata, which classify, summarize information about processes and phenomena on the earth's surface [1].

The features of solving the problems of developing and implementing computer-based geological and geocological 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]. 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 [6]. The mentioned publications [2–5] provide several basic solutions to the issues of preprocessing, intelligent analysis of geodata by means of the computer system GeoBazaDannych. The results and methodological recommendations of cluster analysis of geodata obtained with the environment of the system GeoBazaDannych are discussed below.

The solution to the problem of cluster analysis (segmentation) [7,8] 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. 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 educational aspect should be noted separately. For example, in [9] it is emphasized that any activity in the field of artificial intelligence combines a high degree of research intensity, the complexity of engineering work, and involves highly qualified performers. The combination of fundamental scientific and engineering-practical training of specialists is a complex educational and pedagogical problem. When learning specialists in the field of artificial intelligence, it is necessary to simultaneously form their research and engineering-practical skills, an understanding of the high demands on the quality and reliability of the results and conclusions obtained. Representative examples are given below and cluster analysis tools implemented in the GeoBazaDannych system environment [10] are noted.

2 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.

By means of the system GeoBazaDannych, it is possible to generate and visualize digital descriptions of spatial distributions of data on sources of contamination, on the geological structure of the studied objects; graphically illustrate solutions to problems describing the dynamic processes of multiphase filtration, fluid migration, heat transfer, moisture, and mineral water-soluble compounds in rock strata; design and implement interactive scenarios for visualization and processing the results of computational experiments 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

The main components of the system GeoBazaDannych [4,5,10]:

- 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;
- software and algorithmic support for the formation and maintenance of permanent hydrodynamic models of multiphase filtration in porous, fractured media;
- modules for three-dimensional visualization of dynamic processes of distribution of water-soluble pollutants in active soil layers;
- modules for organizing and supporting the operation of geographic information systems in interactive or batch modes;
- software and algorithmic support for the formation and maintenance of permanent hydrodynamic models of multiphase filtration in porous, fractured media;
- the Generator of the geological model of a deposit (GGMD) – the integrated software complex of the composer of digital geological and geoecological models, which includes software components:
 - • tools and patterns for preparation of reference (calibration) model of digital field, which corresponds to the specified properties (“Digital field constructor”);
 - • tools and several options of “distortion” of reference model;
 - • tools for data capture simulation, which are used in simulation practice (“Generator of profile observer”);
 - • modules for calculation, visualization, comparison of digital fields approximation by several different methods (“Approximation component”);
 - • tools and adaptation modules for digital model being formed (“Adaptation component”);
 - • clustering tools.

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.

3 Preparation of Source Data

The examples below are calculated with the data [5] for surface $zSurfF$. Explanations of why exactly such a surface is representative of the corresponding class of models are given in [5]. It is noted which disturbances of the base surface are added, and how the possibilities of their numerical description are analyzed by approximation based on the results of measurements on a scattered set of points. Figure 1 illustrates the used expression, it shows: the surface on the left and the volume bounded by it on the right.

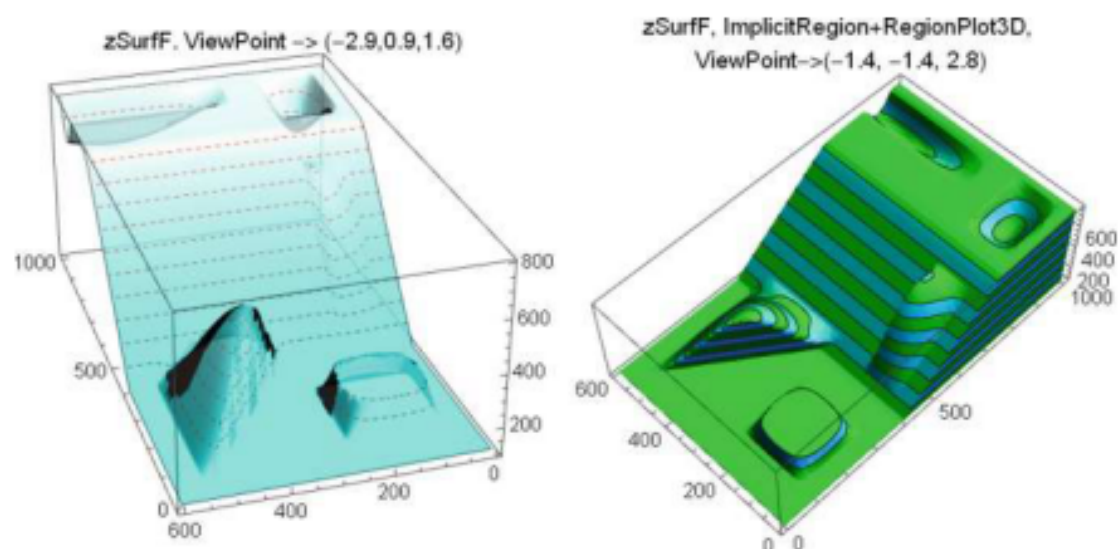


Fig. 1. Visualization of the $zSurfF$ reference surface in the Plot3D and RegionPlot3D variants.

For clarity, Fig. 2 shows the isolines (contour lines) of the $zSurfF$ levels. Also, signatures of the form (1), ..., (5) indicate the positions of perturbations of the base surface – they will be the objects of search during clustering.

Data for demonstration of methods and algorithms of intellectual analysis are obtained by simulation of measurements – the main is shown in Fig. 3.

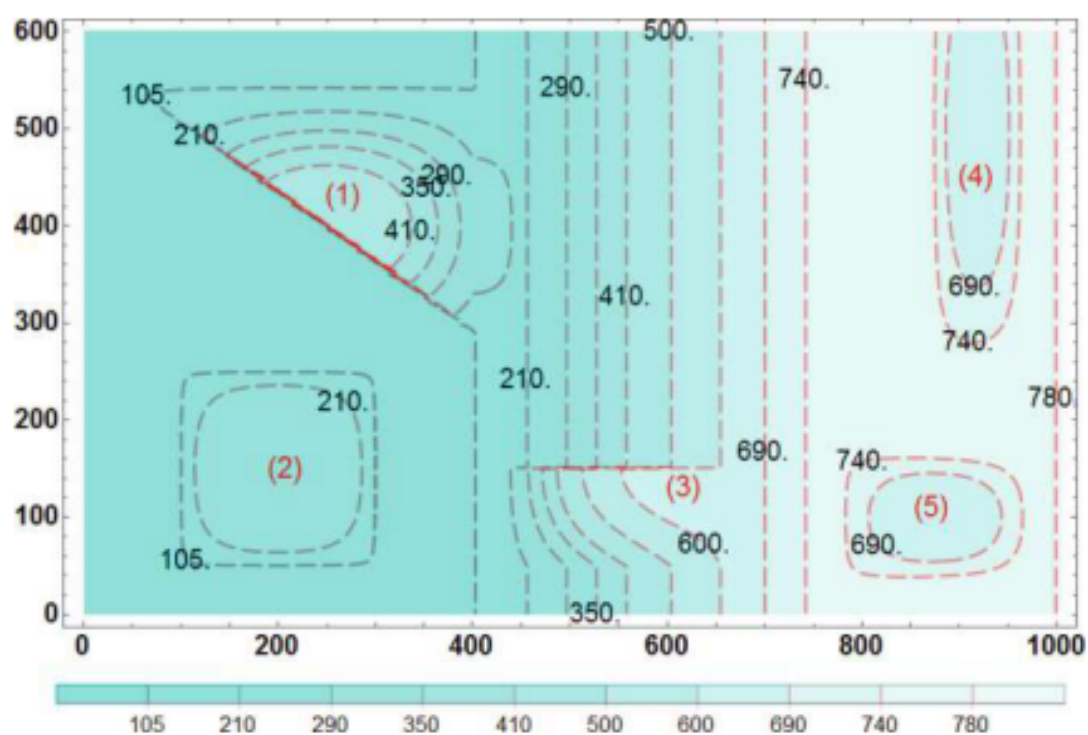


Fig. 2. Contour map of the reference surface z_{SurfF} .

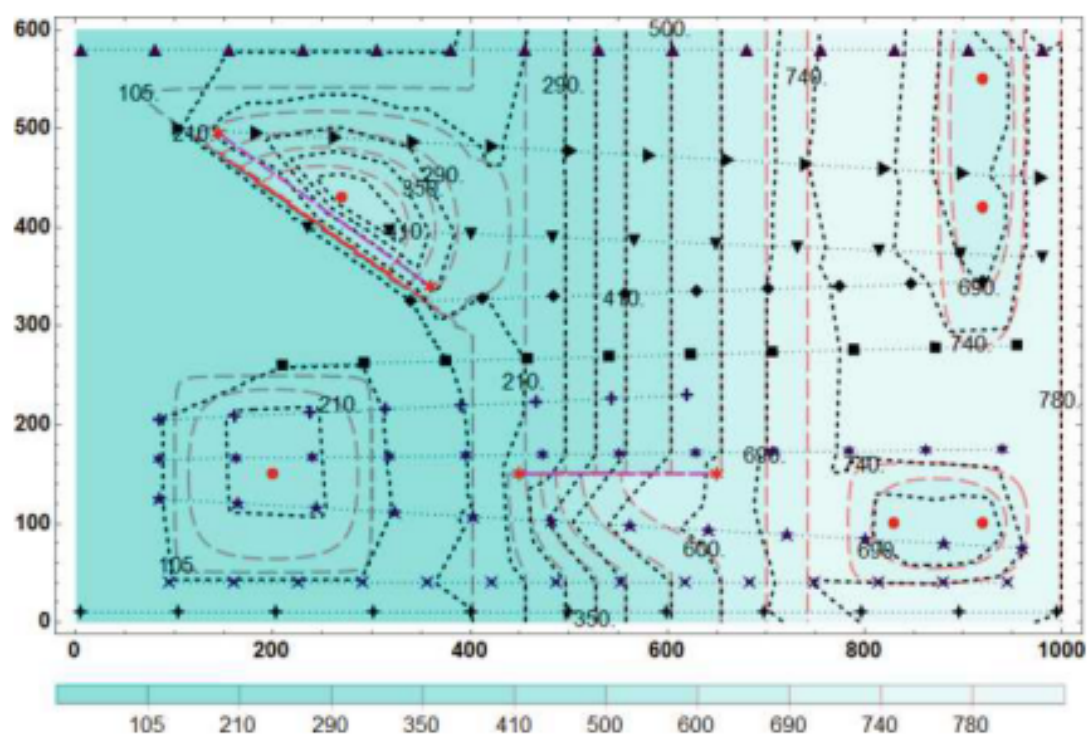


Fig. 3. A scheme of points with level measurements, a map of isolines of the reference and reconstructed surfaces. (Color figure online)

The data shown and used below in computational experiments, the calculated set of triples of numbers (coordinates and values) are points with the values of the level of the reference surface, representing (in fact) a scattered set of points. In this approach, they are interpreted as data on observation profiles, which are shown in Fig. 3 by dotted lines, the points themselves on different profiles are shown by different primitives. The isolines of the reference surface (thin long dashed red lines) and the isolines of the surface reconstructed in Wolfram Mathematica are also shown (Interpolation method, InterpolationOrder = 1, line format – black dashed lines).

4 Tools, Examples of Cluster Analysis of Geodata

Cluster analysis allows for many different types of clustering techniques/algorithms to determine the final result [11]. Below are the results that illustrate the features of the most commonly used clustering algorithms.

4.1 Effects of the Number of Clusters

One of the most important problems of segmentation is determining the number of clusters. 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.

The series of illustrations in Fig. 4 shows the results calculated with default settings using the Wolfram Mathematica FindClusters function (details below).

A comparison of the presented variants gives grounds to assert that additional actions are needed to select the method, metric and other parameters of clustering algorithms.

4.2 Effects of the Accepted Clustering Method

In the examples below, a priori information is used, the number of clusters is set to 6. Why so much – it is taken into account that in the initial data, measurements were carried out for a surface that included 5 different distortions of it with individual positioning of perturbations.

The effects of the accepted clustering method (Possible settings for Method) are illustrated by the schemes in Fig. 5. In the illustrations (to remind the data source), the isolines of the reference surface are given in long dashed red lines. 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. 5 was calculated using the DistanceFunction EuclideanDistance metric.

Representative clustering options are shown, namely: KMeans [12], k-medoids [13], Spectral [14], Optimal.

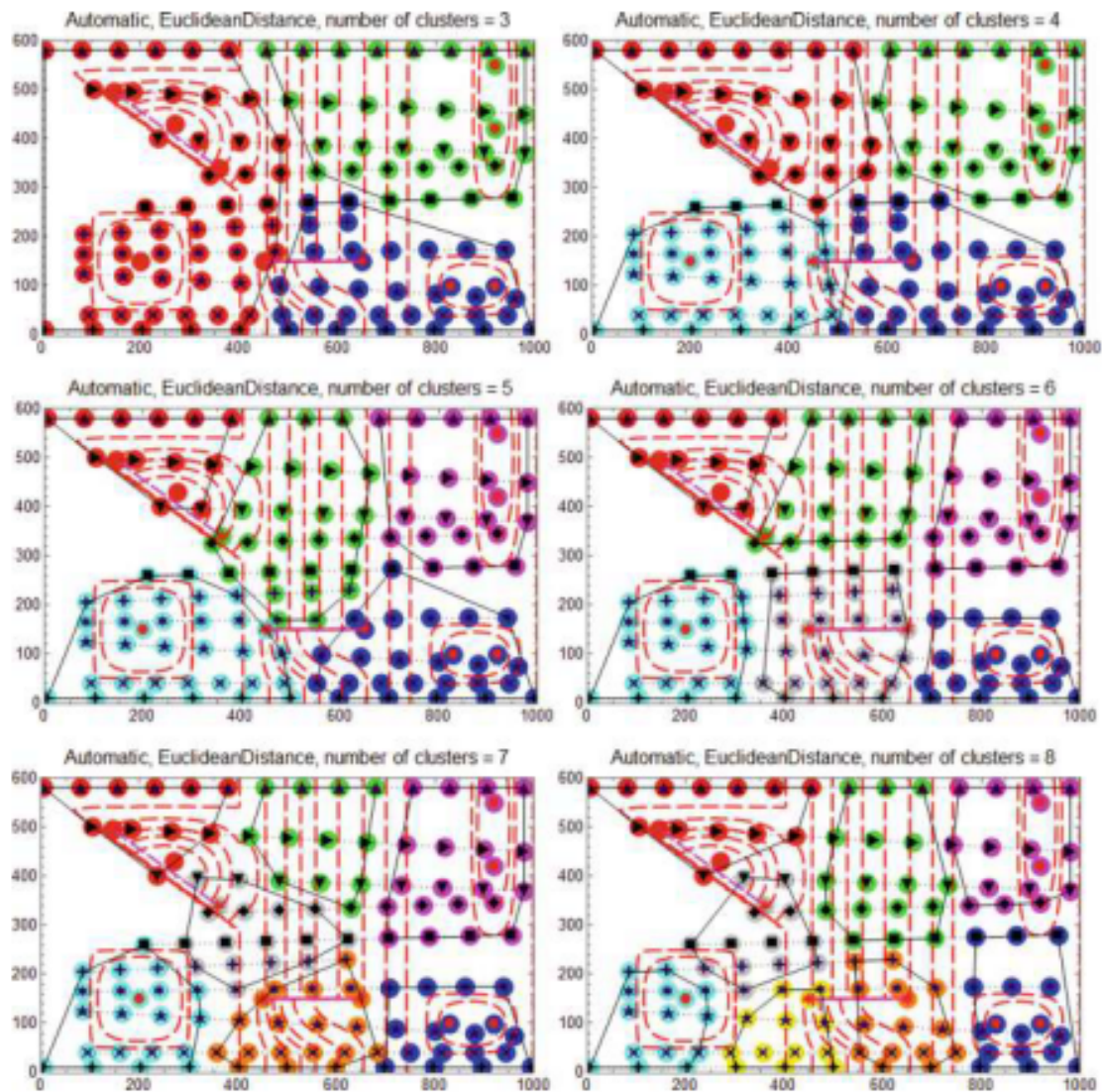


Fig. 4. Influence of the number of clusters.

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 (find clustering hierarchically), Optimize (find clustering by local optimization), DBSCAN (density-based spatial clustering of applications with noise), GaussianMixture (variational Gaussian mixture algorithm), JarvisPatrick (Patrick clustering algorithm), KMeans (k-means clustering algorithm), KMedoids (partitioning around medoids), MeanShift (mean-shift clustering algorithm), NeighborhoodContraction (displace examples toward high-density region), SpanningTree (minimum spanning tree-based clustering algorithm), Spectral (spectral clustering algorithm). What segmentation methods are used in the calculations are recorded in the headers of the diagrams.

The results shown in Fig. 4 and Fig. 5 (Automatic, KMeans, KMedoids, Spectral, Optimal) are markedly different. At the same time, given the digital

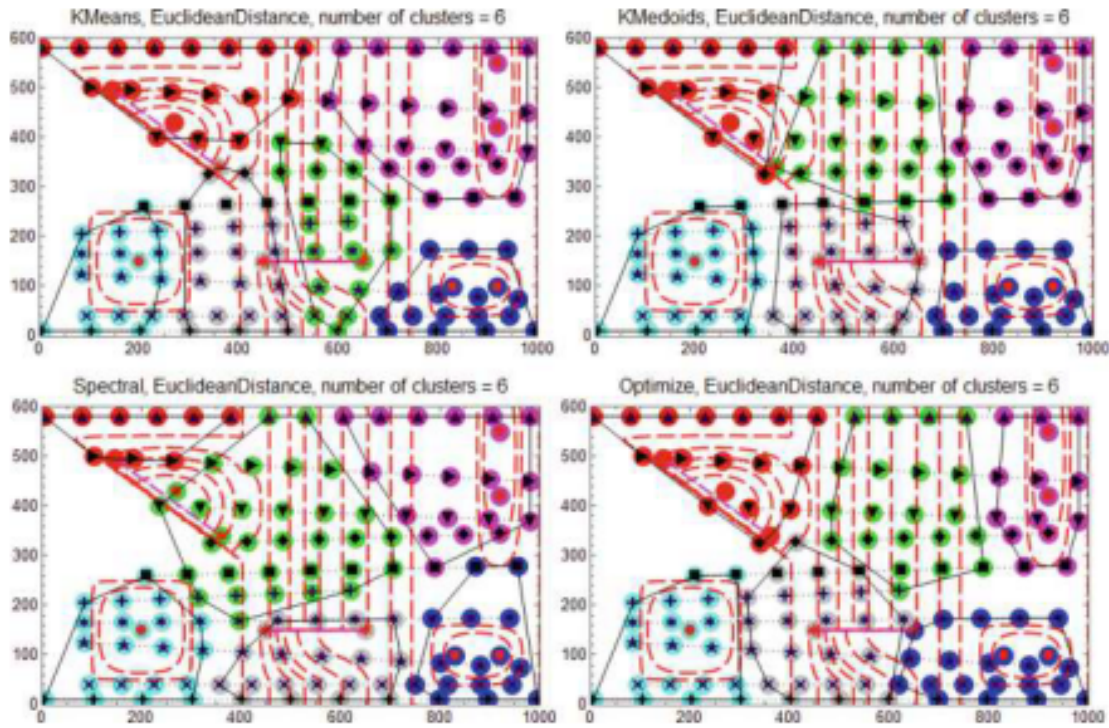


Fig. 5. Clustering methods. (Color figure online)

field of the original, it is difficult to name the preferred variant, we need to refine the algorithms.

The results shown were obtained in calculations when the number of clusters was set by the user of the program. By default, the FindClusters function tries various methods and selects the best clustering. However, there are possibilities [15] to specify additional options, in particular, you can determine: the average radius of the neighborhood of a point (NeighborhoodRadius), the average number of points in the neighborhood (NeighborsNumber), initial centroids/medoids (InitialCentroids). ClusterDissimilarityFunction specifies the intercluster dissimilarity. The methods KMeans and KMedoids determine how to cluster the data for a particular number of clusters k . The methods DBSCAN, JarvisPatrick, MeanShift, SpanningTree, NeighborhoodContraction, and GaussianMixture determine how to cluster the data without assuming any particular number of clusters. The methods Agglomerate and Spectral can be used in both cases.

The results of calculations using the MeanShift function and two representative options for setting NeighborhoodRadius values are shown in Fig. 6. It should be noted that in the case of NeighborhoodRadius = 10, the automatically selected number of clusters was 5, in the variant 20–8. And the same number of clusters for the data set in question is obtained by increasing NeighborhoodRadius to 100.

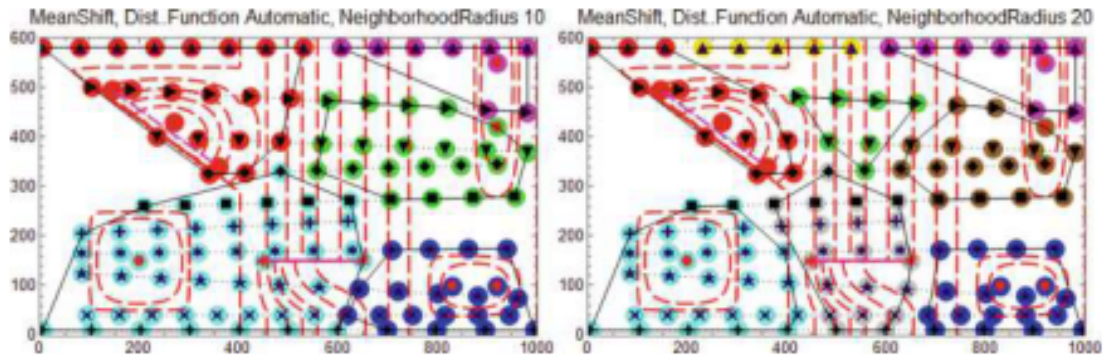


Fig. 6. Clustering method MeanShift, different NeighborhoodRadius.

4.3 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. 7.

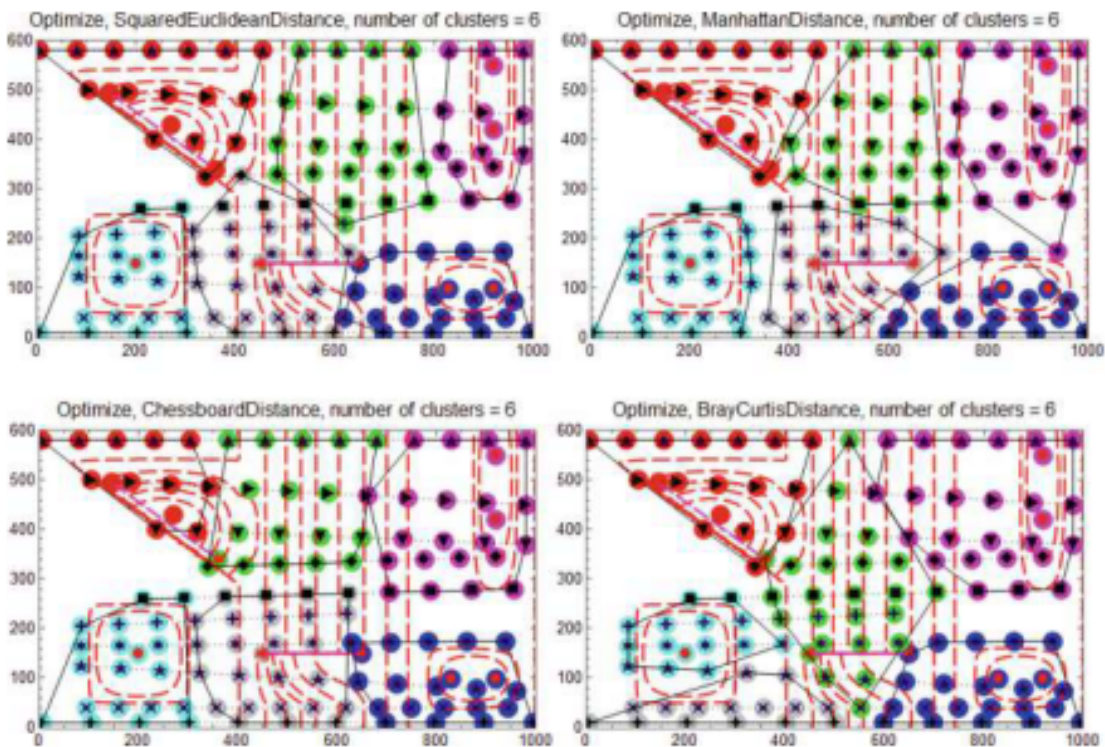


Fig. 7. Influence of DistanceFunction.

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 were available for analyzing digital data [15]: `EuclideanDistance`, `SquaredEuclideanDistance`, `NormalizedSquaredEuclideanDistance`, `ManhattanDistance`, `ChessboardDistance`, `BrayCurtisDistance`, `CanberraDistance`, `CosineDistance`, `CorrelationDistance`, `BinaryDistance`, `WarpingDistance`, `CanonicalWarpingDistance`. The algorithmic features of the listed metrics can be clarified in the articles [16–18]. 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 k-medoids method was used. What methods of `DistanceFunction` are used in calculations is recorded in the headers of the schemes. Representative variants are shown, namely `SquaredEuclideanDistance`, `ManhattanDistance`, `ChessboardDistance`, `BrayCurtisDistance`.

5 Influence of XYZ-Accounting for Values in Points

In the results considered and shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7, the similarity or difference between the classified objects is established depending on the metric distance between them. Figure 8 shows classification options using the Wolfram Mathematica `ClusterClassify` function. This function uses data to group into clusters and works for various types of data, including numerical, textual, and image, as well as dates and times and combinations of these. Options of `ClusterClassify` function [15]: `DistanceFunction` (how to compute distances between elements), `ClusterDissimilarityFunction` (how to compute dissimilarity between clusters), `Weights` (weights for different data elements), `PerformanceGoal` (whether to optimize for speed, memory, quality, training time, etc.), `CriterionFunction` (how to assess automatically selected methods), `Method` (manual override for automatic method selection).

In the results presented below, the `ClusterClassify` function is used to perform clustering not only taking into account the coordinates of the points of the scattered set, but also the values in them. In other words, in the results presented in this series, the algorithms take into account not pairs X_i, Y_i , but triples - X_i, Y_i, Z_i . The results are shown in Fig. 8. The illustrations on the left show the results of clustering by pairs (coordinates only), and on the right – by three values; at the same time, calculations are performed using the same methods and metrics.

It follows from the results that for the data set under consideration, taking into account the values at the points does not give an additional obviously positive effect in the implementation of clustering. But such results are useful and

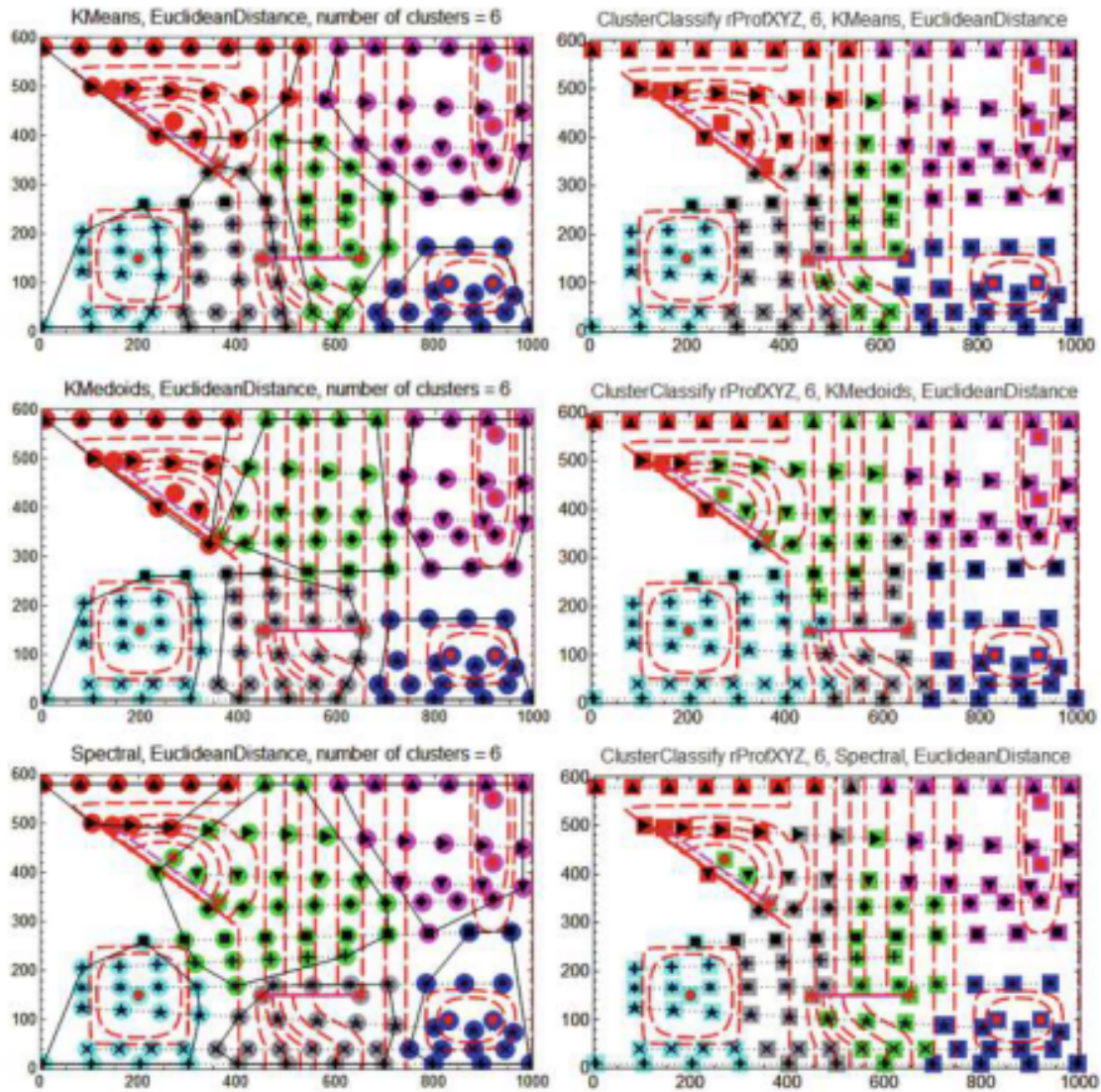


Fig. 8. Results of using the ClusterClassify function.

important, since it is clear from the comparison where additional source data is needed.

6 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 digital model of a spatial object are presented and discussed.

References


1. Savinyh, V.P., Tsvetkov, V.Y.: Geodannye kak sistemnyi informacionnyi resurs. Vestnik Rossiiskoi akademii nauk **84**(9), 826–829 (2014)

2. Taranchuk, V.B.: Examples of the use of artificial neural networks in the analysis of geodata. *Open Semantic Technologies for Intelligent Systems. Research Papers Collection* **3**, 225–230 (2019)
3. Taranchuk, V.: Tools and examples of intelligent processing, visualization and interpretation of GEODATA. In: *Modelling and Methods of Structural Analysis*. IOP Publishing (2020). *IOP Conf. Ser. J. Phys. Conf. Ser.* **1425**(012160), 1–9 (2020)
4. Taranchuk, V.B.: Examples of intelligent adaptation of digital fields by means of the system GeoBazaDannych. *Open Semantic Technologies for Intelligent Systems. Research Papers Collection* **4**, 243–248 (2020)
5. Taranchuk, V.B.: Interactive adaptation of digital fields in the system GeoBazaDannych. In: Golenkov, V., Krasnoproshin, V., Golovko, V., Azarov, E. (eds.) *OSTIS 2020. CCIS*, vol. 1282, pp. 222–233. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60447-9_14
6. Shaitura, S.V.: Intel'ektual'nyi analiz geodannyh. *Perspektivy nauki i obrazovaniya* **18**(6), 24–30 (2015)
7. Charles, D.T.: Concepts of clustering. In: *Indexing, and Structures Data Architecture*, pp. 241–253 (2011). <https://doi.org/10.1016/B978-0-12-385126-0.00013-9>
8. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: *Cluster Analysis*, 5th edn. Wiley, Hoboken (2011)
9. Golenkov, V., Guliakina, N., Golovko, V., Krasnoproshin, V.: Methodological problems of the current state of works in the field of artificial intelligence. *Open Semantic Technologies for Intelligent Systems. Research Papers Collection* **5**, 17–32 (2021)
10. Taranchuk, V.B.: *Komp'yuternye modeli podzemnoi gidrodinamiki*. BGU, Minsk (2020)
11. Kriegl, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. *WIREs Data Min. Knowl. Discov.* **1**(3), 231–240 (2011). <https://doi.org/10.1002/widm.30>
12. Bock, H.: Clustering methods: a history of k -means algorithms. In: Brito, P., Cucumel, G., Bertrand, P., Carvalho, F. (eds.) *Selected Contributions in Data Analysis and Classification. STUDIES CLASS*, pp. 161–172. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73560-1_15
13. Park, H., Jun, C.: A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36**(2), 3336–3341 (2009)
14. von Luxburg, U., Belkin, M., Bousquet, O.: Consistency of spectral clustering. *Ann. Stat.* **36**(2), 555–586 (2008). <https://doi.org/10.1214/0090536070000000640>
15. Distance and Similarity Measures. <https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures>. Accessed 24 Oct 2021
16. Amigó, E., Gonzalo, J., Artiles, J., et al.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retrieval* **12**, 461–486 (2009). <https://doi.org/10.1007/s10791-008-9066-8>
17. Grabusts, P.: The choice of metrics for clustering algorithms. In: *Environment Technology. Resources. Proceedings of the International Scientific and Practical Conference*, pp. 70–76 (2011). <https://doi.org/10.17770/etr2011vol2.973>
18. Zhu, W., Ma, C., Xia, L., Li, X.: A fast and accurate algorithm for chessboard corner detection. In: *2nd International Congress on Image and Signal Processing*, pp. 1–5 (2009). <https://doi.org/10.1109/CISP.2009.5304332>

Editors

Vladimir Golenkov 
Belarusian State University of Informatics
and Radioelectronics
Minsk, Belarus

Vladimir Golovko 
Brest State Technical University
Brest, Belarus

Viktor Krasnoproshin 
Belarusian State University
Minsk, Belarus

Daniil Shunkevich 
Belarusian State University of Informatics
and Radioelectronics
Minsk, Belarus

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-15881-0 ISBN 978-3-031-15882-7 (eBook)
<https://doi.org/10.1007/978-3-031-15882-7>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Vladimir Golenkov
Viktor Krasnoproshin
Vladimir Golovko
Daniil Shunkevich (Eds.)

Communications in Computer and Information Science

1625

Open Semantic Technologies for Intelligent Systems

11th International Conference, OSTIS 2021
Minsk, Belarus, September 16–18, 2021
Revised Selected Papers

 Springer

