Adv Syst Sci Appl 2023; **04**; 179-187
Published online at https://ijassa.ipu.ru.

Machine Learning and Geometric Mathematical Models in Kimberlite Well Classification Problems

Nailia Gabdrakhmanova*, Pavel Klimtsev

Peoples Friendship University of Russia (RUDN University), Moscow, Russia

Abstract: Nowadays, the application of mathematical models in geology becomes more and more relevant. The steady trend towards the global digitalization has led to the possibility of using the most modern computational methods in the construction of mathematical models. Digitalization, further processing of digital data, their analysis and subsequent modeling contributes to the improvement of production efficiency. The purpose of this paper is the development of various methods of classification of kimberlite wells. The paper presents neural network, statistical and geometric mathematical models for solving the problem of kimberlite well classification. The problem was solved using geological and exploration data from wells drilled in the Süldükar and Ulakhan-Kurung-Yuryakh areas located in Western Yakutia. For the constructed models the estimations of the models' qualities were obtained, the comparative analysis of the models was carried out. The analysis of mathematical models showed that the most accurate models are neural network models and models using geometric methods.

Keywords: five factor analysis, neural network, kimberlite wells, classification, persistent diagram, curvature, topological data analysis.

1. INTRODUCTION

Nowadays in geology there is a lot of quantitative data obtained during exploration subjected to digitization. This makes it possible to use geological and exploration data in mathematical modeling to predict certain processes and phenomena of interest to geologists. In particular, such a task arises in the exploration of diamond deposits where kimberlite is present. Since its search is very costly, it is desirable to use mathematical models to predict the presence of kimberlite based on samples taken from the wells.

A kimberlite pipe [1] is understood as a vertical or near-vertical geologic body formed when magma breaks through the Earth's crust. A kimberlite pipe is usually filled with kimberlite [2], a series of magmatic ultramafic rocks of extrusive facies that form explosion tubes as well as dikes and sills. They often contain xenoliths of mantle rocks and sometimes contain diamonds of industrial concentrations.

The purpose of the study is the development of mathematical models to solve the problem of classifying wells by the presence of kimberlite based on the data of geological exploration of wells.

1.1. Data

The material for the study was the data of geological surveys of kimberlite wells drilled in the Süldükar and Ulakhan-Kurung-Yuryakh areas located in Western Yakutia. The object of the study is two data sets of geologic features, where such important parameters as: their occurrence, their presence and their number in each drill hole are indicated. The data were collected from two areas located geographically in Western Yakutia: Süldükar, where kimberlite was found, and Ulakhan-Kurung-Yuryakh, where kimberlite is absent.

During special documentation of core [3,4] – rock samples extracted from wells of two areas in Western Yakutia – with kimberlite (46 wells, Süldükar) and without kimberlite (103 wells, Ulakhan-Kurung-Yuryakh), tectonic and mineralogical features in the host strata of the Lower

* Corresponding author: gabd-nelli@yandex.ru

Paleozoic were recorded. The total area of both sites under consideration is 4 km^2. The core is a cylindrical column (pillar) of rock strong enough to maintain monolithicity. The density of drilling networks is correlated, and for a competent comparison of areas, wells of near-tubular space were excluded on a 20x20m network. Also, for correct mathematical analyses, the parameter of feature occurrence per drilled meter of rock was adopted.

1.2. Method

To solve the problem of object classification, several parallel models were built: a model using factor and cluster analysis, neural network models, models using methods of differential geometry and topological data analysis. All models showed a satisfactory solution to the problem. According to the analysis of solutions, all models have an acceptable calculation error.

Methods and models of factor analysis, as well as the method of principal components [6], are aimed at compression of information, i.e. reduction of the dimensionality of the feature space (this premise of the possibility of reducing the feature space in factor analysis is based on the mutual correlation of the initial features). Within the development of the factor analysis model, the methods proceed from a common basic idea, in which the structure of relationships between p analyzed features $(X^{(1)}, X^{(2)}, \dots, X^{(p)})$, can be explained in such a way that all these variables depend (either linearly or otherwise) on a smaller number of other, not directly measurable (latent) factor $\left(f^{(1)}, f^{(2)}, \dots, f^{(r)}\right), (r < p)$, which are called common and which in most models are constructed so that they turn out to be mutually uncorrelated. Cluster analysis was performed on the constructed factors using the k-means method. The action of the clustering algorithm is such that it seeks to minimize the total quadratic deviation of cluster points from the centers of these clusters:

$$V = \sum_{i=1}^{k} \sum_{z \in S_i} (z - \mu_i)^2,$$

where k is the number of clusters, S_i is the clusters obtained, $i = 1,2, \dots, k$, μ_i is the centers of masses of all vectors z from the cluster.

Neural networks [7] are mathematical models that do well in the task of classification. Classification is the process of finding a function that helps to divide a set of data into classes. The task of a classification algorithm is to find a mapping function to map an input vector X to a discrete output y. In neural network modeling, a neural network is first trained on a training dataset and then, the trained neural network can be used to classify new data. The quality of the solution depends on the choice of neural network architecture, the amount of data, and the degree of data noise. To evaluate the accuracy of the solution of the problem, the solution of the neural network on the test set is used.

Geometric mathematical models are based on the idea that topology and geometry provide a powerful approach to obtaining reliable qualitative and sometimes quantitative information about the structure of data. They are developed from the incorporation of geometric and topological methods, dealing with point clouds, i.e. finite sets of points equipped with a distance function. Point clouds can be viewed as finite samples taken from a geometric object, possibly with noise. Mathematical models adapt tools from different sections of geometry to study point clouds.

The classical Ricci curvature plays an important role in the geometric analysis of Riemannian manifolds. The Ricci curvature at a given point characterizes the average curvature of sectional curvatures in all directions. The notion of Ricci curvature for metric spaces of general form was first introduced by Bakri and Emery [8], Ollivier [9] in 2009 gave a definition of rough Ricci curvature on Markov chains, which can be used for metric spaces generated by graphs. And in 2011, Lin, Lu and Yau [10] modified Ollivier's definition for the Ricci curvature of Markov chains on metric spaces.

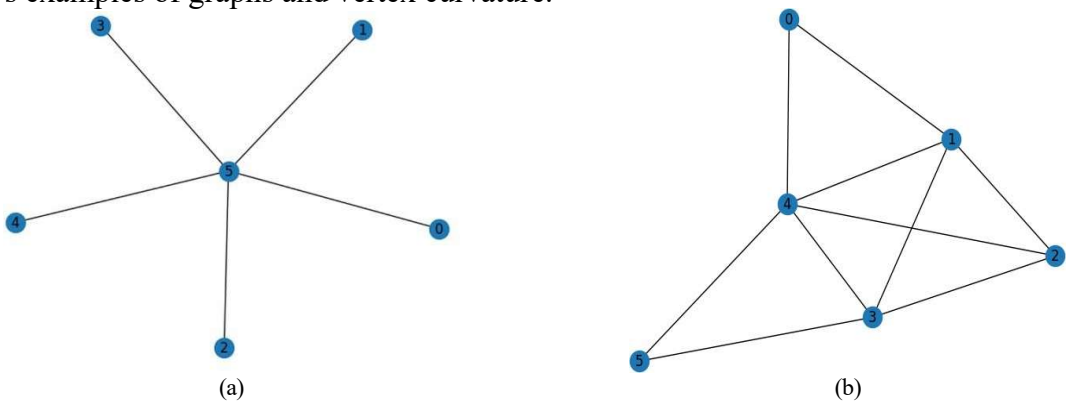
This paper uses a method that combines network analysis techniques with classical correlation, graph theory and local clustering coefficient. This provides a novel graphical representation of the features that solves the problem. By analogy with curvature in Riemannian geometry, we

interpret Ricci curvature as the amount of overlap between the neighborhoods of two neighboring vertices.

For the solution we use the notion of local clustering coefficient, which shows the density of triangular relations. Studies show that curvature is usually extremely low in random graphs. Clusters of high curvature have a very non-random structure. Depending on the problem to be solved, different methods of Ricci curvature estimation are used. To solve the clustering problem, the Watts-Strogatz formula [11] is used in this paper:

$$curv(A) = {}^{t}\Big/{}_{\left(\frac{v(v-1)}{2}\right)}.$$

Here $curv(A)$ is the vertex curvature, v is the numbers of vertices and t is the number of triangles that are formed by the edges of the graph containing vertex A. This function is a function of two variables. Note that the value $v(v-1)/2$ is the maximum number of triangles that can be formed by all vertices of the graph, hence $curv(A)$ lies between 0 and 1. Figure 1 shows examples of graphs and vertex curvature.



(a) (b)

Fig 1. (a) The graph has vertices $V = 6$, for vertex 5 : $curv(5)$=0;
(b) The graph has vertices V = 6, for vertex 3 : $curv(3) = {}^{1}/_{5}$.

Recently, there has been an increasing interest in topological data analysis (TDA) [12- 17]. Topological data analysis is a valuable tool for data analysis and visualization, which can be applied in various fields such as computer vision, bioinformatics and graph theory. The tools of topological analysis include barcodes, Battie numbers and persistence diagrams. We propose a method to compare two data representations of two clusters in feature space. To investigate the point cloud in this problem, we used the Vietoris-Rips complex. The Vietoris-Rips complex is a symplectic complex whose simplexes are all possible sets of points of a given metric space in which pairwise distances between points do not exceed a given positive number. The constructed persistence diagram can be used to understand the features and relationships between features in the data.

2. RESULTS AND DISCUSSION

Mathematical statement of the problem. We are given 10 attributes of 46 objects of class ‘0’and 10 attributes of 107 objects of class ‘1’. Objects of class ‘0’ – wells where there is kimberlite; objects of class ‘1’ – wells where there is no kimberlite. It is necessary to build a mathematical model to classify objects by 10 features.

The features investigated were:

0) tectonic fractures (r)
1) extensive pyritization (Q)
2) fluidizite breccias (L)
3) gaping fractures (m)
4) recrystallization (T)

5) drag folds (X)
6) microfossils (A)
7) subhorizontal stylolite seams (V)
8) ogling (d)
9) slip mirrors with subvertical grooves (F).

We will use the number in the list later in the article as a feature code.

2.1. Statistical Models

Initially we have 10 geological features under consideration, but only 4 of them, namely tectonic fractures (r), extensive pyritization (Q), fluidisite breccias (L) and drag folds (X), were used in the factor analysis because they occur in both areas.

The factor analysis is based on the correlation matrix between variables. Tables 2.1 and 2.2 present the correlation matrices calculated using IBM SPSS Statistics 23 program. The table 2.3 and 2.4 below presents two different tests: the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's test of Sphericity and Bartlett's test of Sphericity for Süldükar and Ulahan respectively, found using Statistics.

Table 2.1 Correlation matrix of indicators (Süldükar)

Variable	r	Q	L	X
r	1	0.02	-0.07	0.2
Q	0.02	1.0	-0.05	-0.07
L	-0.07	-0.05	1.0	0.11
X	0.2	-0.07	0.11	1.0

Table 2.2 Correlation matrix of indicators (Ulahana)

Variable	r	Q	L	X
r	1.00	-0.06	-0.02	-0.13
Q	-0.06	1.00	-0.06	-0.09
L	-0.02	-0.06	1.00	-0.01
X	-0.13	-0.09	-0.01	1.00

The table 2.3 and 2.4 below presents two different tests: the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's test of Sphericity and Bartlett's test of Sphericity for Süldükar and Ulahan respectively, found using Statistics.

Table 2.3 KMO and Bartlett’s (Süldükar)

KMO		0,461
Bartlett’s test of Sphericity	Approx.Chi-Square	2.956
	df	6
	Sig.	0.814

Table 2.4 KMO and Bartlett’s (Ulahana)

KMO		0.456
Bartlett’s test of Sphericity	Approx.Chi-Square	3.336
	df	6
	Sig.	0.766

Both tables do not show high values of correlations, therefore, the variables are not correlated with the same factors. We accept the null hypothesis in both cases according to Bartlett's criterion of sphericity. The approximate values of χ^2 statistics are 2.956 and 3.336 respectively with 6

degrees of freedom, they are insignificant at 0.05 level of significance. The values of KMO statistics (0.461 and 0.456) are also smaller (< 0.5); significance 0.814 and 0.766. Thus, it is questionable to consider factor analysis as an acceptable method for analyzing correlation matrix data.

Tables 2.5, 2.6 show the eigenvalue tables and Figures 2.1, 2.2 stony scree plots for Süldükar and Ulahan, respectively, plotted using Statistica program.

Table 2.5 Table of eigenvalues (Süldükar)

	The initial eigenvalues		
Component	Total	% variance	Total %
1	1,213	30,328	30,328
2	1,101	27,526	57,853
3	0,951	23,786	81,640
4	0,734	13,360	100,000

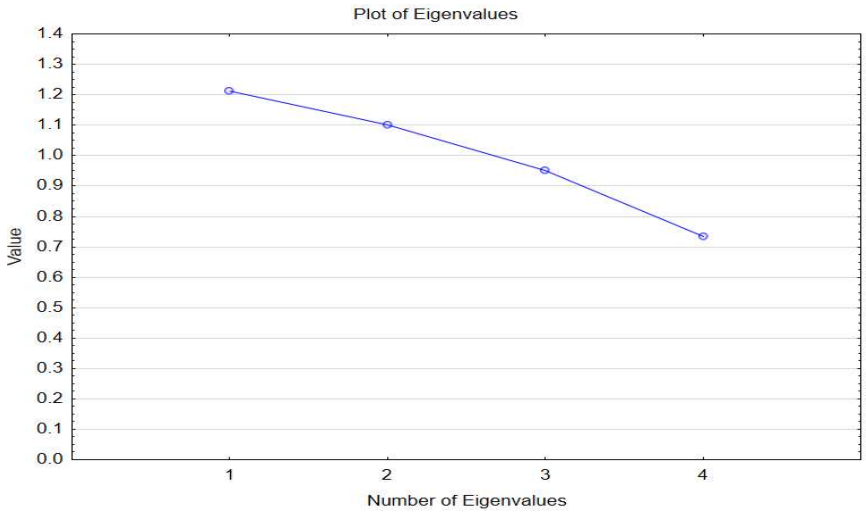


Fig. 2.1. Graph of rocky scree (Süldücar)

Table 2.6 Table of eigenvalues (Ulahana)

	The initial eigenvalues		
Component	Total	% variance	Total %
1	1,133	28,321	28,321
2	1,077	26,919	55,240
3	0,988	24,688	79,927
4	0,803	20,073	100,000

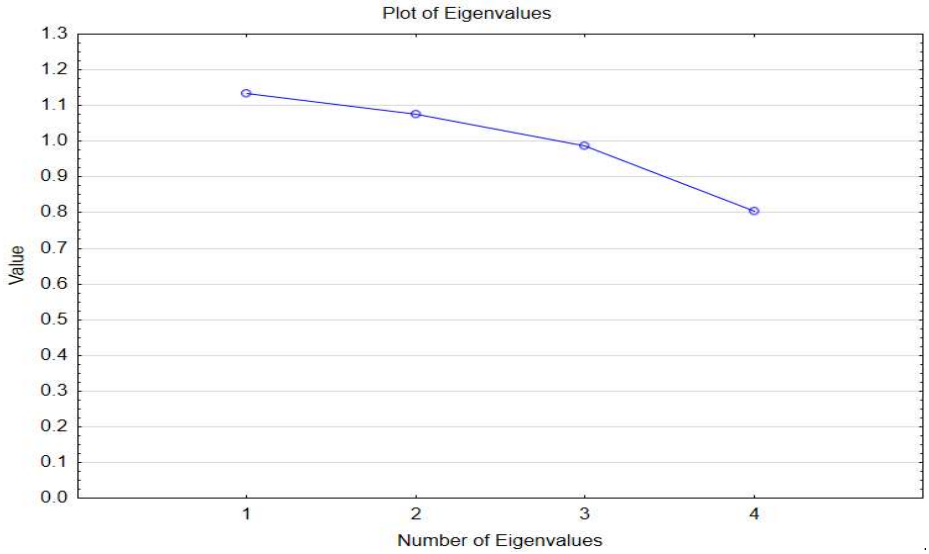


Fig. 2.2. Graph of rocky scree (Ulahana)

The result of this clustering method is as follows:
Süldükar 32 out of 46, Ulahan 37 out of 103.
In summary, the classification accuracy of the factor analysis method was approximately 46.3%.

2.2. Neural Network Models

Tuples were used to build and implement the neural network model <X,Y> : input vector X consisting of 10 elements (geologic features), one output Y producing 0 or 1 (class number – Süldükar and Ulahan respectively). Neural networks of multilayer perceptron (MLP) architecture were chosen to solve the problem. Deep learning neural network was not considered because the data sample size is small. The architecture and parameters of the constructed MLP: two hidden layers of dimensionality 20 and 15 neurons respectively, one output producing 0 or 1 (class number - Süldükar and Ulahan respectively), number of epochs - 200, learning rate - 0.2 (const), activation function - logic function (sigmoid). 65% of the total data sample was used for training the neural network, while the remaining 35% (53 wells) was used for testing the network. The model was implemented using the open-source web application Jupyter notebook in the Python programming language. The MLP multilayer perceptron model performed better than the factor analysis method, namely, the total percentage of correctly classified objects, i.e. wells, in the two classes was approximately 85 % while the factor model performed approximately 46 %.

2.3. Modelling Using Differential Geometry Methods

To solve the problem, we used methods of graph curvature estimation using the Watts–Strogatz formula. Curvature estimates were found for each cluster separately. First, we switched from continuous space to discrete space and constructed a complete graph. Each vertex of the graph was assigned one of 10 features and connected all vertices in pairs by edges. Each edge has a weight equal to the sample correlation coefficient of adjacent vertices. The constructed graph is called a correlation network.

Algorithm 1

Step 1. Construct the correlation network.
Step 2. Remove edges with weight less than h, where h is the selected threshold value.
Step 3. For the resulting connected graph, compute the curvature estimate for all vertices using the Watts-Strogatz formula.

An example of correlation network for Süldükar cluster is presented in Fig. 2.3.

An example of correlation network for Ulahan cluster is presented in Fig. 2.4.

The arithmetic mean of the vertex curvature is taken as the curvature of graph G:

$$curv(G_i) = \frac{1}{n_i} \sum_{j=G_i} curv(j) \quad (1)$$

where i is the cluster number, j is the vertex (feature) number, n_i is the number of vertices of the graph G_i.

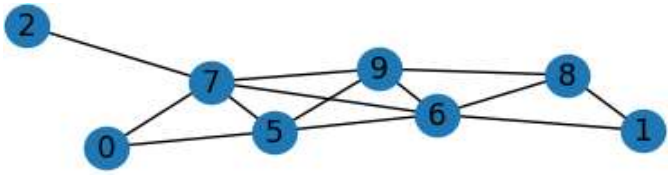


Fig. 2.3. The graph of cluster '0' (Süldükar)

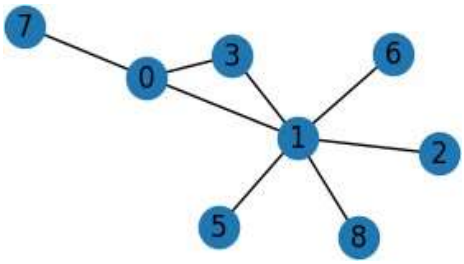


Fig. 2.4. The graph of cluster '1' (Ulahana)

Curvature estimates calculated by formula (1): $curv(G_0) = 0.1$, $curv(G_1) = 0$.

We obtained that $curv(G_0) > curv(G_1)$.

This indicates that the feature space of cluster '1' is flat, while the feature space of cluster '0' is convex.

The graphs in Fig. 2.3, 2.4. give a visualization of the observation points of each cluster. For the study, it is important which vertices are included in the connected graph and the shape of the vertex connectivity. The research conducted using this method identified the features of cluster '0' that have the most stable connection with each other. These features were used in the topological analysis of the data.

Algorithm 2 is a well classification algorithm.

Algorithm 2

1. Identify the vertices (features) that form a connected graph for each cluster.
2. Compute the centers of each cluster
3. For new features, calculate distances to the centers
4. Assign the object to the cluster for which the distance to the center is smaller.

Classification of objects that were not used in the solution of the problem showed the correctness of the solution 88%.

2.4. Topological Data Analysis

Let X be the space to which the observation points belong, and let X have a Euclidean metric. Vjetoris–Rips complexes with different radii have been constructed for the considered metric space *(X,d)*. The symplectic complexes are filtered by successively increasing the radius. The algorithm for their construction from the point cloud and the metric contains information about the filtering of the complex, that is, the increasing chain of embeddings of subcomplexes. The filtering thus contains, among other things, geometric information about the original point cloud, which is encoded in the form of a so-called persistence diagram computed on it [7]. We have

constructed persistence diagrams using the package Ripser. Figures 2.5, 2.6 shows the persistence diagram. In the persistence diagram, the abscissa axis marks the time of birth of the complex (Birth), the ordinate axis - time of death (Death). The difference between the time of death and birth of the simplex is taken as the lifetime of the complex. Fig. 2.5 shows the persistence diagram for cluster '0' (Süldükar). Fig. 2.6 shows the persistence diagram for cluster '1'(Ulahana). Barcodes were constructed from the persistence diagrams. Let us denote by L_i the total length of barcodes for homologs H_i. Then the average value of the Euler characteristic can be defined by the formula $\chi = L_0 - L_1 + L_2$.

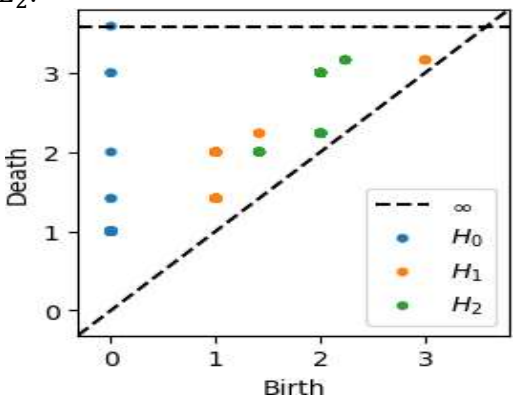


Fig. 2.5. The persistence diagram for cluster '0' (Süldükar)

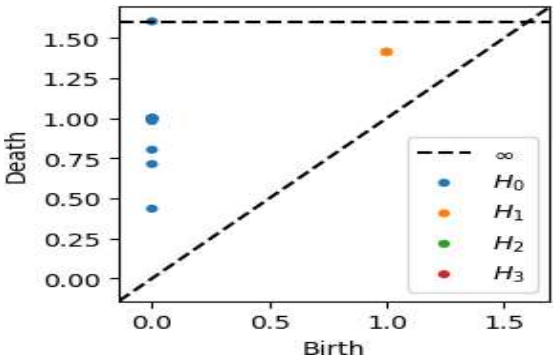


Fig. 2.6. The persistence diagram for cluster '1' (Ulahana)

As a result of the calculations, we obtained that the Euler characteristic is larger for cluster '0', than for cluster '1'. The results of calculating barcode lengths are summarized in Table 2.7.

Table 2.7. The Euler characteristic

	Cluster '0'	Cluster '1'
$L(H_0)$	7.4	3.8
$L(H_1)$	2.8	0.4
$L(H_2)$	2.6	0
χ	7.1	3.4

Algorithm 2 is chosen as the classification algorithm.

CONCLUSION

Neural network models and mathematical models constructed using the tools of statistical analysis, differential geometry and topology are the main tools in the ongoing research.

Algorithms for solving the problem by different methods have been developed and comparative analysis has been carried out.

At the first stage, the classification problem was solved using mathematical statistics and neural networks. The comparison of the accuracy of the problem solution showed the advantage of

neural network models. However, neural network models do not always allow to interpret the results of the problem solution. In contrast to NS models, geometric models allowed not only to qualitatively solve the classification problem, but also gave visual and numerical description of the observation points in the feature space. As numerical characteristics of the observation point arrays of the two clusters, we used the estimates of the Ricci curvatures of the correlation networks and the estimates of the Euler features computed from barcodes. The solution to the problem of computing the Ricci curvature estimates identified the most significant relationships in the feature space for wells with the presence of kimberlite. The study showed that the integration of digitalization, neural network modeling, and geometric mathematical models can qualitatively solve the classification problem on exploration data.

REFERENCES

1. Ignatov, P. A. & Novikov, K. V. (2019). *Field diagnostics of tectonic disturbances and fluid fracture formations in kimberlite-bearing sediments of the Lower Paleozoic: Methodical Manual*. Mirny, Russia: ALROSA.
2. Milashev, V. A. (1984). *Explosion tubes.* Leningrad, USSR: Nedra.
3. Zorina, T. G. & Slonimskaya, M. A. (2010). *Marketing research: textbook*. Minsk, Belarus: BSEU.
4. Ilupin, I. P., Vaganov, V. I., Prokopchuk, B. I. (1990). *Kimberlites: Handbook*. Moscow, USSR: Nedra.
5. Afonin, P. N., Afonin, D. N. (2017). *Statistical analysis with the application of modern software tools: textbook.* Saints-Petersburg, Russia: 1C "Intermedia".
6. Aivazyan, S. (2010). *Methods of Econometrics: Textbook*. Moscow, Russia: INFRA-M.
7. Haikin, S. (2006) *Neural networks: a complete course, 2nd edition*. Moscow, Russia: Williams.
8. Bakry, D. & Emery, M. (1985). Diffusions hypercontractives [hypercontractive diffusions], *Séminaire de probabilités de Strasbourg*, **19**, 177–206, [in French].
9. Ollivier, Y. (2009). Ricci Curvature of Markov chains on metric spaces, *J. Funct. Anal.*, **256**(3), 810–864.
10. Lin, Y., Lu, L. Y. & Yau, S. T. (2011). Ricci curvature of graphs, *Tohoku Mathematical Journal*, **63**, 605–627.
11. Watts, D. J. & Strogatz, S. H. (1998). Collective Dynamics Of 'Small-World' networks, *Nature* **393**, 440–442.
12. Edelsbunner, H., Letscher, D., Zomorodian, A. (2002). Topological persistence and simplification, *Discrete Comput. Geom.*, **28**, 511–533.
13. Carlsson, G., Zomorodian, A. *Computing persistent homology* // In “Proc. 20th Ann. Sympos.Comput. Geom, 2004, 347–356.
14. Cohen-Steiner, D., Edelsbrunner, H. & Harer, J. (2007). Stability of Persistence Diagrams, *Discrete & Computational Geometry*, **37**(1), 103–120.
15. Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes, *The Journal of Machine Learning Research*, **16**(1), 77–102.
16. Gabdrakhmanova, N. & Pilgun, M. (2021). Neural Network Technologies and Topological Analysis of Social Media Data, *Advances in Systems Science and Applications*, **21**(3), 101–112.
17. Gabdrakhmanova, N., Fedin, V. & Matsuta.B. (2020). The modeling of forecasting new situations in the dynamics of the economic system on the example of several financial indicators, *Proc. of the 14th International Symposium Intelligent System* (Moscow, Russia).

