UDC 550.4+550.84.09

GEOCHEMICAL BEHAVIOR INVESTIGATION BASED ON K-MEANS AND ARTIFICIAL NEURAL NETWORK PREDICTION FOR TITANIUM AND ZINC, KIVI REGION, IRAN

Adel Shirazy[1],
Adel.shirazy@shahroodut.ac.ir

Mansour Ziaii[1],
mziaii@shahroodut.ac.ir

Ardeshir Hezarkhani[2],
Ardehez@aut.ac.ir

Timofey V. Timkin[3],
timkin@tpu.ru

Valery G. Voroshilov[3],
v_g_v@tpu.ru

[1] Shahrood University of Technology,
Bolvar Daneshka, Shahrood, 3619995161, Iran.

[2] Amirkabir University of Technology (Tehran Polytechnic),
1591634311, Iran.

[3] National Research Tomsk Polytechnic University,
30, Lenin avenue, Tomsk, 634050, Russia.

The relevance. *These are the first studies in the Kivi region. Due to the presence of titanium and zinc in the area, these studies are necessary. Artificial Neural Network and K-means methods for element behavior measurement are new methods in mineral exploration.*
The main aim *of the research is to identify Ti and Zn geochemical behavior for prediction Ti by ANN and K-means methods.*
Object: *Kivi 1:100000 geochemical map in Ardabil province, Iran.*
Methods. *The samples taken from bottom sediments of the Kiwi region, which were analyzed by the ICP-MS method, served as the initial data. Then, the behavior of these elements in relation to each other and their geographical coordinates was analyzed by the K-means clustering method. The amount of titanium was also predicted with the artificial neural network (ANN- GRNN).*
Results. *The Ti and Zn elements relationship was determined using this K-means method taking into account the latitude and longitude of the samples to estimate the grade and more accurate estimation of the appearance and extent of the geochemical halos in the studied area. According to the results obtained during processing of these elements, a regression equation was drawn up to estimate the titanium content based on three parameters: Zn content, the length and width of the sampling points, the correlation coefficient. According to the K-means cluster centers and artificial neural network, the Ti element grade was predicted and the correlation coefficient was reported 0,51. Both methods produce the desired results, but the artificial neural network method has more accurate data. Schematic maps of the initial and predicted Ti content were constructed. The results of the study can be used in the course of geological exploration to forecast and identify new promising areas.*

Key words:
Titanium, zinc, Kivi region, K-means clustering method, artificial neural network, estimation of the elements grade.

Introduction

In recent years, due to the high dependence of mineral projects on the precise determination of the tonnage of mineral materials, various methods have been developed to estimate the grade, such as geometric methods based on distance and geostatistics [1, 2]. Determining the grade of different anomalous communities is very important when calculating mineral reserves, as well as when processing minerals. [3, 4].

Each method has limitations and disadvantages which affect the accuracy of estimation [5]. One of the new methods is the grade estimation using the clustering. The cluster analysis methods are widely used in the earth sciences. The cluster grouping method is used to classify geochemical data [6]. The clustering method is also used in processing satellite images (remote sensing), which is useful in combining exploratory information layers [7].

The cluster analysis relates the observations to each other which together have many similarities, then, the observations consecutively link them which is most similar to previous observations [8]. In other words, in clustering, we try to divide the data into clusters that the similarity is maximized between the data within each cluster and it is minimized between the data within the different clusters [9]. There are no classes in the clustering method and in fact, the variables are not divided independently and dependently, but here, the search is performed to access groups of data which are similar to each other and the behaviors can be better identified by discovering these similarities and it can be operated to achieve a better result based on them [10]. Clustering method is the indirect method; this means that it can be used even when there is no previous information from the internal database structure. This method can be

used to discover hidden patterns and improve the performance of direct methods [11].

The K-means method is one of the techniques for clustering the data in data mining. It is an exclusive and planar method, which has been widely studied by different researchers and attempts to cluster the samples with the specified number of k classes so that the total Euclidean intervals of each sample are minimized from the center of the class [12]. Some of clustering methods applications include: the division of the geological terrain [13], the classification of the effect of vegetation and the recovery of water health in the Mediterranean coast forests [14], the presentation of geochemical patterns in mineral areas [15], the prediction of organic carbon in the intelligent systems [16, 17], and the determination of gas diffusion effect in urban environments [18].

In recent years, considerable attention has been given to developing estimation and forecasting models based on artificial intelligence techniques such as artificial neural networks (ANNs) and expert systems. These techniques have been successfully applied to a wide range of engineering applications by many authors reporting higher accuracy compared to classical estimation methods. ANNs are computer models that are designed to emulate human information processing capabilities such as knowledge processing, speech, prediction, and control. The ability of ANN systems to handle a large number of variables with complex relationships, spontaneously learn from examples, reason over inexact and fuzzy data, and to provide adequate and quick responses to new information has generated increasing acceptance of this technology in different engineering fields.

In this article, the behavior of the titanium and zinc elements has been evaluated using K-means method, MATLAB and SPSS software based on the data collected from the drainage sediments in the Kivi area and then the titanium grade is predicted as well with the artificial neural network. The Kivi area, located in the East Azerbaijan Province of Iran (Fig. 1), has a high mineral potential of metal elements [19].

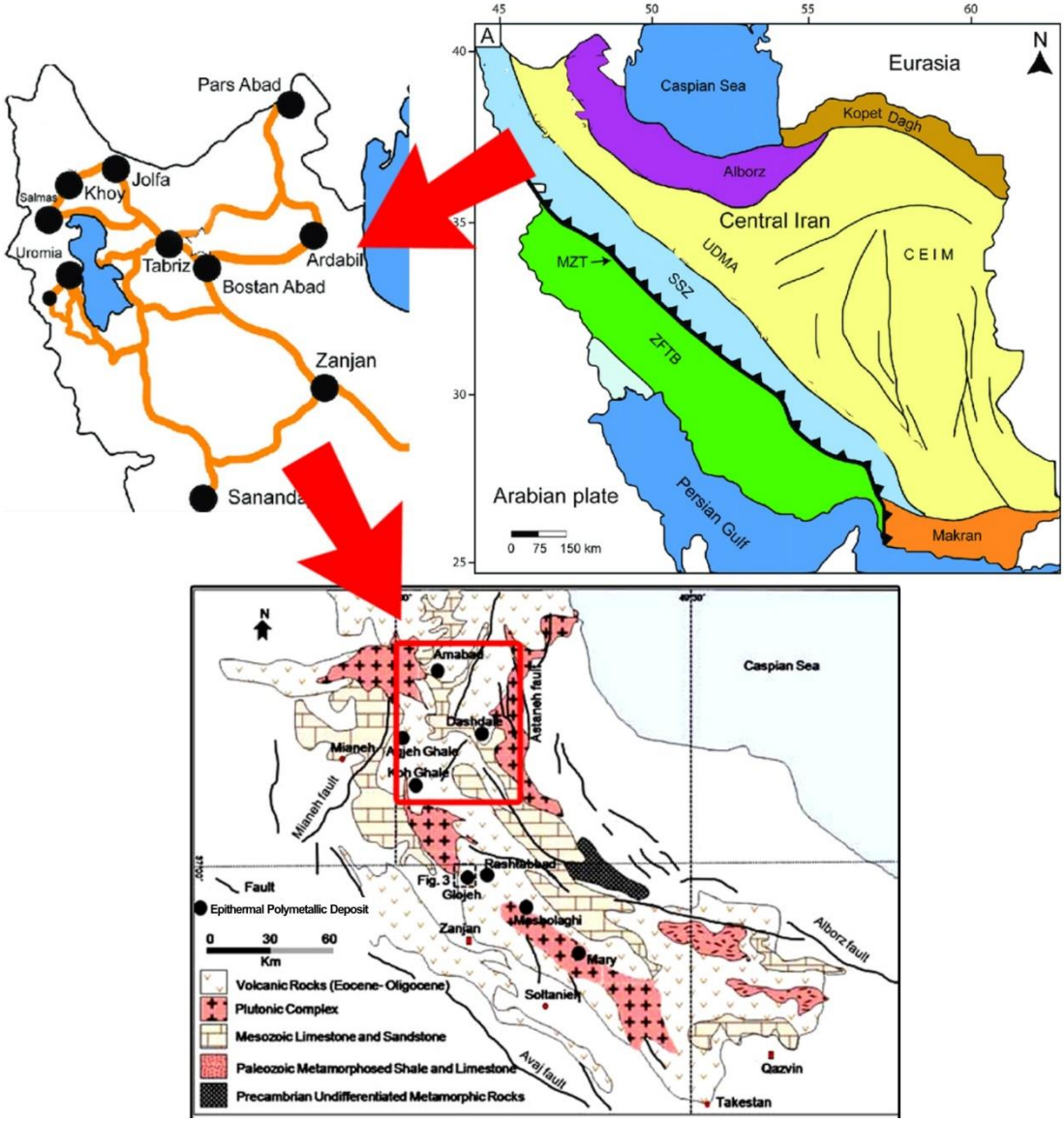

Fig. 1. *Geographical location and simplified geological map of Kivi region in north Iran*

Рис. 1. *Географическое положение и упрощенная геологическая схема региона Киви на севере Ирана*

Geological setting of studied area

The geology map of Kivi is located in Ardabil province among the cities of Ardabil, Khalkhal and Mianeh (Fig. 1). This area is located in the northwestern part of the 1:250000 geological map of Anzali port and its geographical coordinates are 48° 00’ to 48° 30’ Eastern longitude and 37° 30’ to 38° 00’ northern latitude.

The trend of the mountain ranges in this region is mostly north-south and very harsh morphology. The region includes Eocene and young volcanic rocks in the northwest and along the Ardabil–Kivi road, there are outcrops of Neogene loose marl sediments, flat ground and hill-like marshes. In general, there are two main mountain ranges in this area. The mountains of the eastern part of the region, the highest elevation of which exceeds 2600 meters, and the mountain range of the western part, with the highest elevation of about 2500 meters.

The Kivi area consists of three sedimentary, igneous and metamorphic units. The oldest existing sedimentary unit is the pre-Cretaceous rocks, and the youngest one is Quaternary sediments. The geological characteristics of the rock units are as follows:

Pre-Cretaceous Rocks

In the western part of Haji Yousef Village along the Sangabad Road, the outcrops of metamorphic rocks with inclusions of sericite schiste, andalusite micaschiste.

Upper Cretaceous

Cretaceous limestone outcrops are found only in the southeastern part of the area, which are more spread to the east. These formations are massive and thick layers silica limestone, and is found between lime shales, calcareous shales, and pyrite shales in it. Its color is gray.

Eocene

In the region, it is a sedimentary and igneous unit that covers most of the area which includes conglomerate, green tuffs, tuffaceous sandstone, lithic tuff, sand limestone and tuff limestone, andesitic lava, andesitic basalt and basalt, rhyolitic and rhyodacitic yellowish tuffs [20, 21].

Methodology and materials

Exploratory Geochemistry

The sampling method, which was used in this area, is drainage sediments sampling type. 714 samples of drainage sediments were collected from the Kivi area and analyzed on 63 elements (Ag, Al, As, Au, Ba, B, Be, Bi, Br, Ca, Cd, Ce, Co, Cr, Cs, Cu, Eu, F, Fe, Ga, Ge, Hf, Hg, In, Ir, K, La, Li, Mg, Mn, Mo, Na, Nb, Nd, Ni, Os, P, Pb, Pd, Pr, Pt, Rb, Re, Ru, S, Sb, Sc, Se, Si, Sn, Sr, Ta, Te, Th, Ti, Tl, U, V, W, Y, Yb, Zn, Zr) using ICP-MS method. The location of the samples can be seen in Fig. 2.

Raw Data Preparation Methods

Before using raw data the censored and outlier data must be identified and replaced. Censored data is said to be the data among which, due to the high sensitivity limit of measuring devices, a number of data are found to be smaller than the device sensitivity limit. Such data can make statistical problems, because, firstly, statistical methods require a complete set of non-censored data, and secondly, in some cases, such as anomaly separation from background and relative measurements, the existence of censored data leads to inappropriate evaluations. If the censored data are identified and replaced, the amount of background and intensity of the anomalies will be calculated more accurately [22–25].

As the existence of censored data among geochemical data leads to errors, outlier data also have the same effect on the results. The release of statistics is the point of observation, remote from other observations [26, 27]. Outlier may be due to measurement variability or may indicate experimental error, the latter are sometimes excluded from the dataset [28]. Emissions can cause major problems in the statistical analysis. Outliers can occur randomly in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution [29].

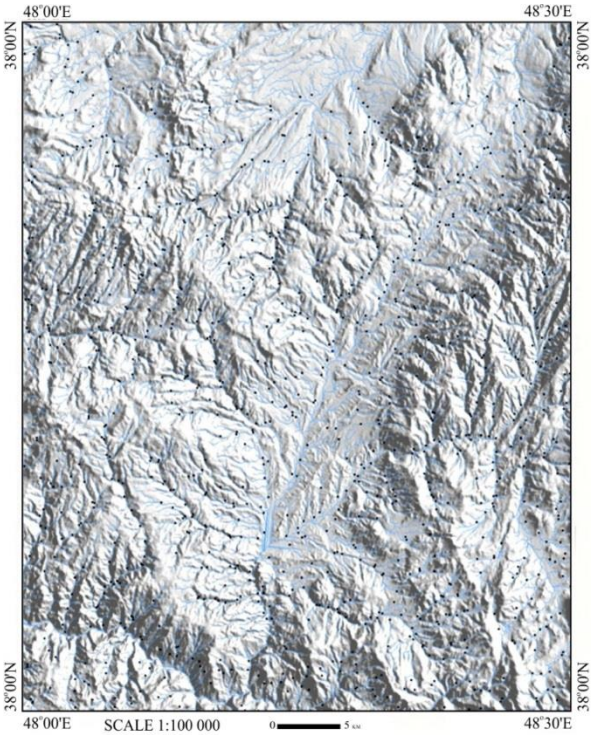

Fig. 2. *Location of geochemical samples on Kivi area topography map [30]*

Рис. 2. *Расположение геохимических точек отбора проб на топографической карте района Киви [30]*

Several methods can detect and replace censored and outlier data. In this study, a simple method is used to replace sensor data. In this method, the values lower than sensitivity limit are replaced by 3/4 of data value. The main problem of this method is that it is by no means influenced by the statistical parameters of the data society and is merely a function of the sensitivity limit of the measurement method [31].

In order to identify and replace the outlier data, the Doerffel method was used [29]. Using this method, a graph for determining the threshold of outlier data values, which is provided for two levels of significance of 5 and 1 % (Fig. 3).

To perform the Doerffel test, the average ($\bar{x}$) and standard deviation of the data (s) is calculated regardless of the largest amount of data. Then the largest amount of data (x_A) is considered to be outside of the row if it is true in the following equation (1):

$$x_A \geq \bar{x} + s.g. \quad (1)$$

Where «g» is the outlier threshold, which can be calculated from the graph shown in Fig. 3.

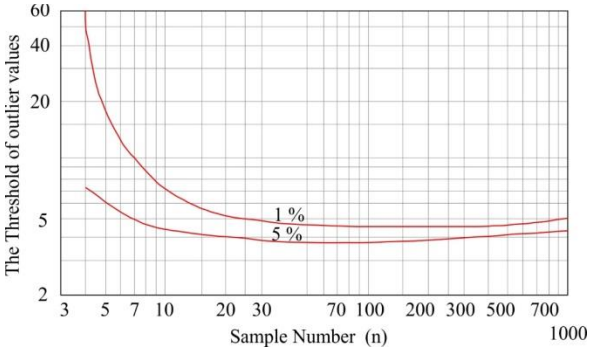

Fig. 3. *Threshold of outlier values (g) as a function of the sample number (n) and the level of trust*

Рис. 3. *Пороговое значение выбросов (g) в зависимости от номера выборки (n) и уровня доверия*

K-Means

The K-means algorithm starts with a given value for K (number of classes) and tries to estimate the following cases:

- Finding the points as centers of clusters, in fact, these points are the same average points of each cluster.
- Assigning each sample data to a cluster that data has the smallest distance to the center of that cluster [32]. In the simple form of this method, first, the points are selected randomly as much as needed clusters. Then, the data is assigned to one of these clusters according to the similarity and so, new clusters are obtained [33].
- New centers can be calculated for them in each of iterations by repeating the same steps and averaging of data and again the data can be attributed to new clusters [34].

The important steps of this algorithm are summarized as follows [35]:

1. First, k members randomly are selected as the number of clusters among the n members (k is the number of clusters).
2. Z_j vector is calculated based on equation (2) which represents the center of each class C_j.

$$z_j = \frac{\sum_{x \in c_j} x}{\#c_j} for\ j = 1. \ldots . k. \quad (2)$$

3. In this equation, x represents the vector of a sample which is a member of C_j and $\#C_j$ represents the number of samples which are members of the C_j class. It should be noted that relation (2) is used to calculate the center of each class during solving and usually, k samples are randomly selected at the start of the algorithm and are considered as the center of each class [9].
4. Computing of the target function of the classification $\{C_1, C_2, \ldots, C_k\}$ is based on equation (3) which calculates the total distance of samples from the center of the classes.

$$f(C_1. C_2. \cdots . C_k) = \sum_{j-1}^{k} \sum_{X \in C_j} \left|X - z_j\right|^2. \quad (3)$$

5. Minimize the objective function of equation (3) and find the proper classification on the M set with the number k of classes.

To speed up the operation above the authors introduced a software [36–38].

Artificial Neural Network

ANN are a class of flexible non-linear models to simulate biological neural systems. ANN are widely used to solve many complex problems in various fields, including pat-tern recognition, signal processing, language learning, and etc. Typically, a biological neural system consists of several layers, each of which consists of a large number of neural units (neurons) that can process information in parallel. The models with these features are known as ANN models. A typical neuron structure is shown in Fig. 4.

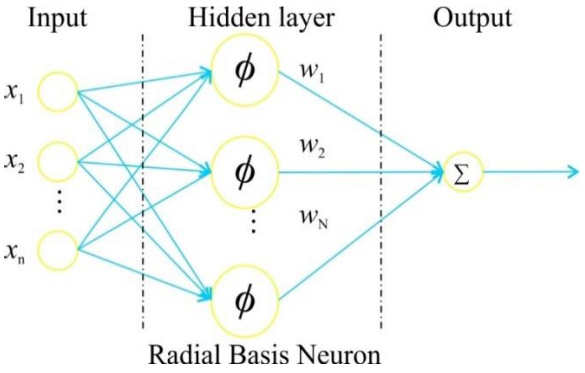

Fig. 4. *Typical neuron of radial basis network*

Рис. 4. *Типичный нейрон радиальной опорной сети*

The basic information processing occurs in the following manner: input (P), coming from another neuron are multiplied by their individual weight ($w_{1,i}$), and weighted input connections are combined within the neuron and the displacement member (b_i) is added to the summation in the neuron to increase or decrease the input (n_j) which included in the activation function (Fig. 5).

When these neurons are combined into a neural network, the system goes through a training stage, in which pairs of inputs and outputs are introduced into the network, and the weights of the neurons are changed to make the network.

The outputs match the desired outputs as closely as possible. According to the learning algorithm, several types of neural networks such as Backpropagation Neural Network

(BPNN), Probabilistic Neural Network (PNN) and general regression neural network (GRNN) have been designed in MATLAB software. Since GRNN method is used in this study, it is therefore described below briefly [41].

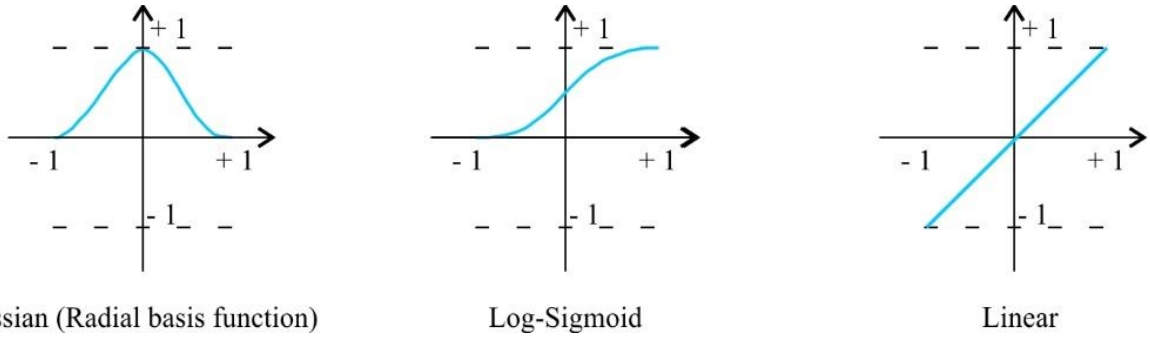

Fig. 5. *Three examples of activation functions [39, 40]*
Рис. 5. *Три примера функций активации [39, 40]*

GRNN is a three-layer supervised network (input, hidden and output layer), where there is one hidden neuron for each training pattern in the hidden layer. A memory-based network that provides estimates of continuous variables and converges to basic surface regression is fast to learn and can model nonlinear functions [39]. It can be thought of as a normalized RBF (Radial Basis Functions) network that has a hidden unit centered in each learning case. These RBF units are usually probability density functions such as Gaussian. The only weights that need to be examined are the RBF block widths. These widths are called «smoothing parameters (r)». The main disadvantage of GRNN is that it cannot ignore irrelevant inputs without major changes to the underlying algorithm. Thus, GRNN is unlikely to be the best choice if there are more than 5 or 6 redundant inputs [42–44].

Regression of the dependent variable Y to the independent variable X is the calculation of the most likely Y value for each X value based on a finite number of possible noisy X dimensions and associated Y values. The X and Y variables are usually vectors. To carry out system identification, it is usually necessary to take some functional form. In the case of linear regression, for example, it is assumed that the output Y is a linear function of the input, and the unknown parameters, a_i, are linear coefficients. The procedure does not have to take a particular functional form. The Euclidean distance (d^2_i) it measured between the input vector and the weights, which are then, scaled using the smoothing coefficient. Output radial exponent basis is negatively weighted distance. The GRNN equations (4), (5) are as follow:

$$d_i^2 = (X - X^i)^T (X - X^i); \quad (4)$$

$$Y(X) = \frac{\sum_{i=1\ to\ n} \exp\left(-\frac{d_i^2}{2\sigma^2}\right)}{\sum_{i=1\ to\ n} \exp\left(-\frac{d_i^2}{2\sigma^2}\right)}. \quad (5)$$

Grade $Y(X)$ can be visualized as a weighted average of the observed values Y_i, where every observed value is exponentially weighted according to its distance from Euclidean X. $Y(X)$ is simply the sum of the Gaussian distributions, centered on each training sample. In this theory, r denotes a smoothing coefficient and smoothing optimum coefficient can be determined after several runs in accordance with a mean square error of the estimated values, which should be minimal. This process is called network training. If several iterations go through without improving the mean square error, the smoothing factor is determined to be optimal for this dataset. During production, the smoothing factor is applied to datasets that the network has not previously seen. When applying a mesh to a new dataset, increasing of smoothing factor will decrease the range of the output values. There are no learning parameters in GRNN such as learning rate, momentum, optimal number of neurons in the hidden layer, and learning algorithms like in BPNN. In addition, GRNN has a high evaluation rate relative to BPNN. The GRNN structure has a smoothing factor, the optimal value of which is achieved by trial and error. The smoothing factor should be greater than 0 and can usually range from 0,1 to 1 with acceptable results. The number of neurons in the input layer is the number of entries in the task, and the number of neurons in the output layer corresponds to the number of outputs. The number of neurons in the hidden layer is these training patterns. Since GRNNs evaluate each output independently of the other outputs, GRNNs can be more accurate than BPNNs when there are multiple outputs. GRNNs work by measuring how far a given sample is from the samples in the training set. The network predicted output is the proportional value of all outputs in the training set. The proportion depends on how far the new template from the set of patterns in the training set [45].

Results and discussion

In various studies, such as the research of relationship of altered diorite with magnetite mineral in the Chilean Iron Belt [46] showed that, the relationship between copper and molybdenum of porphyry copper ore, and the relationship between elements of the platinum group of porphyry copper ore, the behavior of elements has been measured relative to each other in various methods. This article is a purely mathematical study and these methods are selected to create a new perspective on the science of behaviorism and estimation of elements in geochemical data. The K-means clustering method is one of the new ones in behavioral measurement and the artificial neural

network method is also emerging in the world. For this reason, we think that the combination of these two methods can be very efficient and attractive. In the current study, *k* optimum value has been calculated using the K-means method for clustering the drainage sediment data in the Kivi area with three grade value of elements of titanium and zinc (taking into account the coordinates of the sampling points), because the zinc element is important in determining the geochemical halos of the titanium element. Using the coordinates for predicting, transforms our method from a numerical to a structural one and brings our results closer to geological structures.

In this study, two appropriate criteria have been used to calculate the appropriate value of *k* to determine the number of clusters. The first used benchmark is the *S*(*i*) that the number of clusters is changed from 3 to perfect number based on it and then, the obtained results are analyzed to select optimal *k* using the above benchmark.

An appropriate benchmark has been calculated according to equation (6) for determining optimum k. The obtained classifications are measured based on the benchmark.

$$S(i) = \frac{\text{Min}\,(\text{Ave}g\_\text{Between}(i.k)) - \text{Aveg\_Within}(i)}{Max\,[\text{Aveg\_Within}(i).\text{Min}(\text{Aveg\_Between}(i.k))]}\ . \quad (6)$$

In the above equation, *S*(*i*) expresses the utility rate of the i^{th} sample in its class, the parameter Aveg_within (*i*) represents the average distance between the i^{th} sample and the other samples in that class and the parameter Aveg_Between (*i.k*) represents the average distance between the i^{th} sample and the other samples which are members of another class such as *k*.

The results are analyzed by calculating the utility rate as an average utility. The utility rate varies between –1 and +1; as this value approaches to +1, the sample is a member of more appropriate classification and as it approaches –1, it has an inappropriate classification and the zero number means that the presence of the sample is not very important in the current classification or another classification. So, the value of equation (10) is calculated for each sample and then the obtained results are analyzed by calculating the average numbers as the average utility rate of the classification.

The second used benchmark is the quality function. According to the information, the best cluster maximizes the total similarity between the cluster center and all cluster members and minimizes the total similarity between cluster centers. First, a range is determined for the number of clusters to select the best cluster which is between 3 and 10 in this research. Then *p*(*k*) is calculated for each value *k*.

The *k* value is selected as the optimal number of clusters which maximizes *p*(*k*). In this way, the number of clusters can be selected, the distance is maximized between cluster centers and the similarity of cluster centers with the members within each cluster. The quality of clustering results is determined using *k* clusters according to the equations (7)–(12) [43]:

$$O = \{c^n | n = 1. \ldots . k\}; \quad (7)$$

$$O^n = \{c_i | i = 1. \ldots . \|T^c - O\|\}; \quad (8)$$

$$\rho(k) = \frac{1}{k}\sum_{n=1}^{k}\left(m_i n\left\{\frac{\eta_n + \eta_m}{\delta_{nm}}\right\}\right); \quad (9)$$

$$\eta_n = \frac{1}{\|\sigma^n\|}\sum_{c_i \in 0^n} Sim(c_j . c^n); \quad (10)$$

$$\eta_m = \frac{1}{\|\sigma^m\|}\sum_{c_i \in 0^m} Sim(c_j . c^m); \quad (11)$$

$$\delta_{nm} = Sim(C^n . C^m). \quad (12)$$

In these equations, *O* is the set of cluster centers; C^n is the center of clusters; O^n is the set of elements which has not been selected as cluster centers; T^c is the set of all elements which is clustered; η_n is the average similarity between the center of the cluster C^n and all the cluster elements of O^n; η_m is the average similarity between the center of the cluster C^m and all the elements of the cluster O^m, and δ_{nm} is defined as the similarity of C^n and O^n [44].

The monitoring of Ti and Zn elements relative to each other

In order to study the behavior of the elements relative to each other, the cluster profile and the utility rate of each sample were determined in pair for classifications *k*=3 and *k*=20 for the elements of titanium and zinc, and the results of the utility rate of the classes have been compared, the best class is determined according to the utility rate of the classes, and then the centers of the clusters of each class are determined according to it.

As the Fig. 6 shows, class 6 is selected as the best one according to the class profile diagrams and the utility rates of best class for the two Ti and Zn elements, since if the utility factor is close to 1, the samples are more correctly located in the class. According to the diagram, the little negative values are also found in this classification.

The utility rate average in this classification is equal to 0,6964 which is greater than the average utility rate of other classes.

The value of *k* has been increased to 100 due to the changes in the utility rates and to ensure the results, but the utility rate has not exceeded the rate of the best classification of each class, and it has decreased for more than 20.

The diagram of validation value *S*(*i*) can be shown in Fig. 6 based on changing the number of clusters to select the optimal cluster number which is easier to compare. In other words, a cluster is selected as the optimal number of clusters which has the highest value of *S*(*i*). Fig. 7 shows the value of *S*(*i*) for two elements of Ti and Zn which has the highest value in the cluster 6.

Also, the proper number of clusters is determined according to the quality function and using the value of *p*(*k*). The value of *p*(*k*) has been calculated using the equation (6) for different k values to determine the number of clusters. As it is stated, the maximum value of *p*(*k*) represents the proper number of clusters. Table 1 shows the values of *p*(*k*) corresponding to the number of clusters. The highest value is 0,6845 in monitoring the two elements of Ti and Zn. Consequently, the most suitable number of clusters is 6 and therefore the number of clusters is 4 for two elements of Ti and Zn with location of the samples. As can be seen, the proper number of clusters is obtained from the quality function which is consistent with the standard results of *S*(*i*).

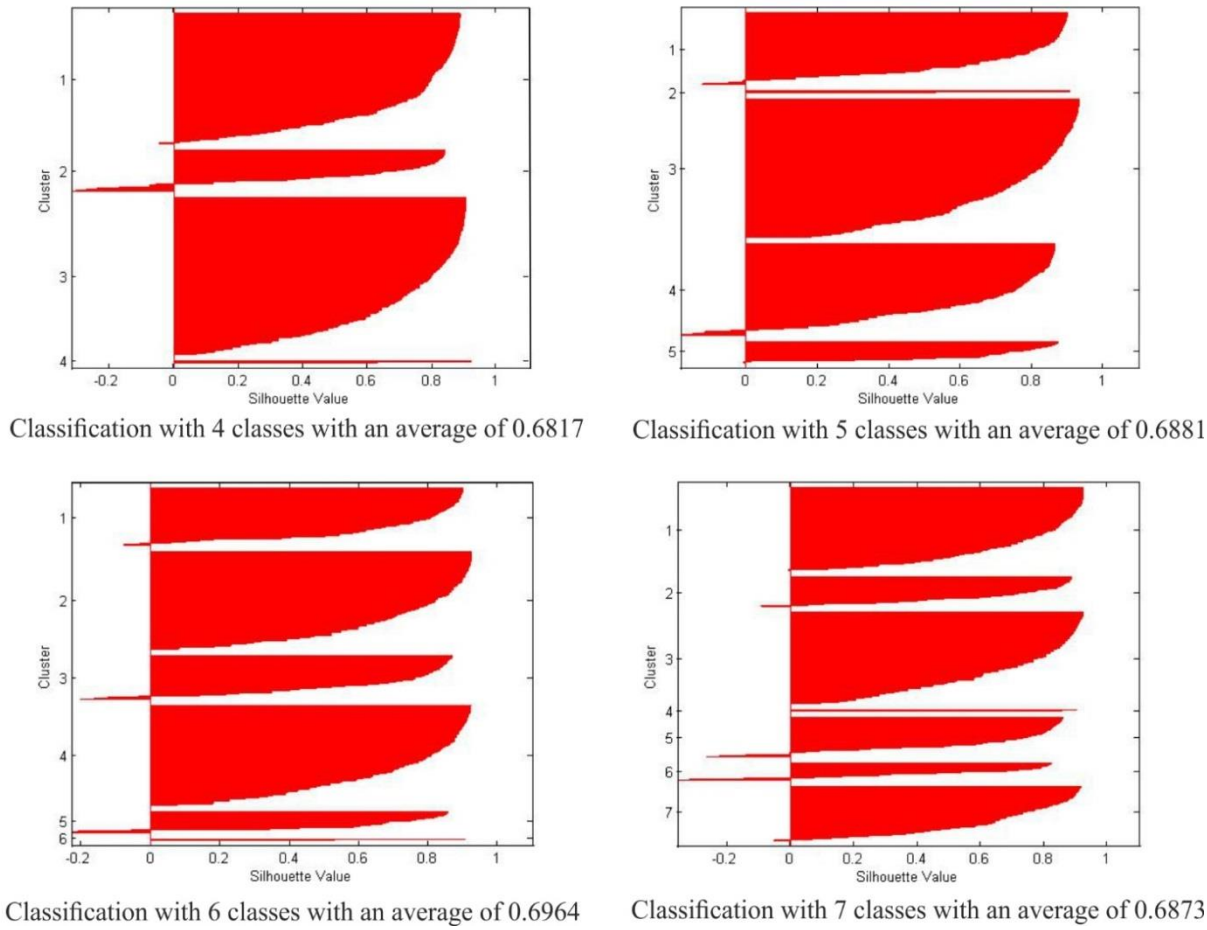

Fig. 6. *Profile of clusters and utility rates from 4 to 7 classes related to two elements of Ti and Zn*

Рис. 6. *Профиль кластеров и коэффициенты полезности с 4 по 7 классы, относящиеся к двум элементам Ti и Zn*

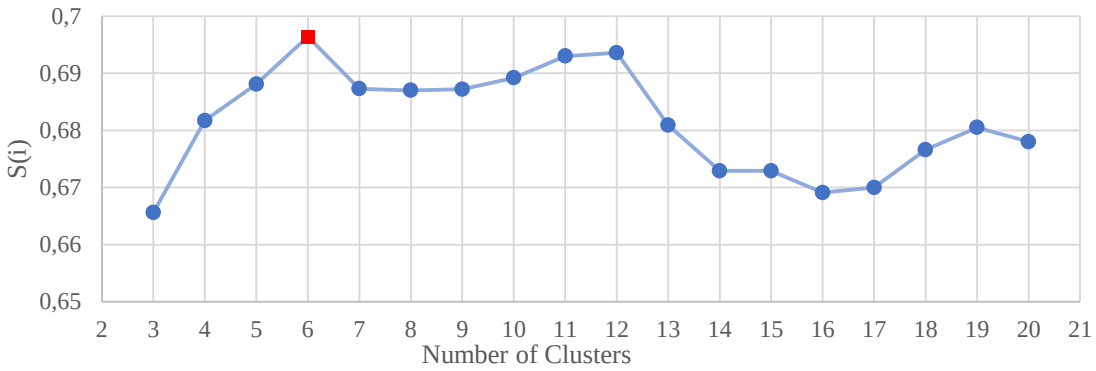

Fig. 7. *Validation of S(i) value based on the number of clusters (Ti and Zn)*

Рис. 7. *Проверка значений S(i) на основе количества кластеров (Ti и Zn)*

Based on the classification shown in Fig. 8 due to the behavior of Ti relative to Zn, Ti increases at Zn growth, for this reason, we see the direct relationship of these elements. The fitted line equation is Y=0,0132X+6,717 and the correlation coefficient of the equation fitted to the center of the classes is equal to R^2=0,9774.

Investigating the Ti behavior regarding the grade of Zn and coordinates

Therefore, in order to obtain center characteristics of classes, all input values must be in a standard interval to prevent the error in calculations and obtain accurate estimation value given the fact that coordinates are considered as input features besides the grade of Ti and Zn and the interval of coordinates variation and grade values are different (Fig. 9, 10). For this reason, all inputs were selected in the interval [1, 0] using equation (13).

The characteristics of cluster centers are given with Forth classes in Table 2.

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \quad (13)$$

Table 1. *Values of p(k) for the number of different clusters*

Таблица 1. *Значения p(k) для количества различных кластеров*

p(k)	Number of clusters	Elements	
0,4337	3	Ti–Zn	
0,5298	4		
0,5898	5		
0,6845	6		
0,5146	7		
0,5674	8		
0,5390	9		
0,5252	10		
0,6053	3	Ti–Zn	According to the location of the samples
0,7423	4		
0,6709	5		
0,6625	6		
0,6584	7		
0,6314	8		
0,6125	9		
0,6033	10		

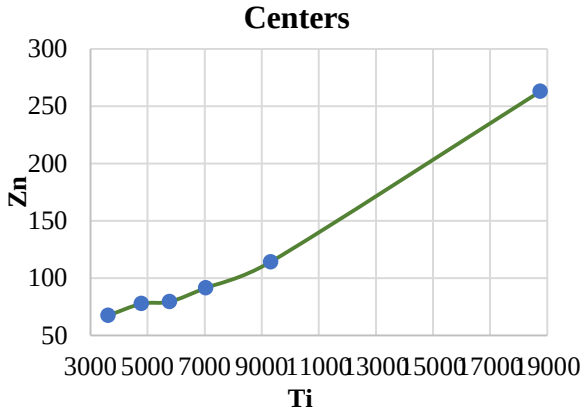

Fig. 8. *The best line fitted to the centers of classes per six classes for Ti and Zn elements*

Рис. 8. *Лучшая линия, подходящая к центрам классов по шести классам для элементов Ti и Zn*

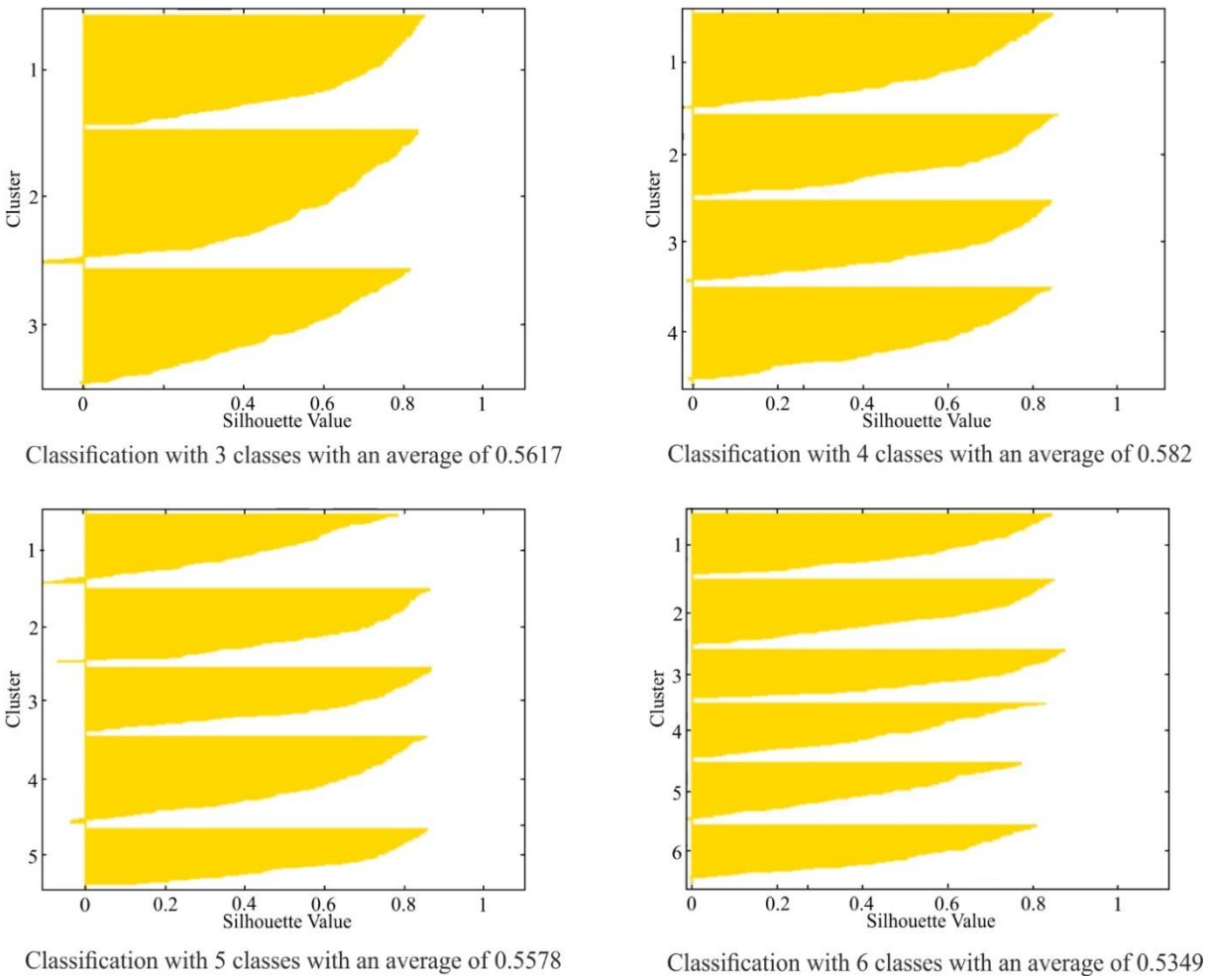

Fig. 9. *Profile of clusters and utility rates from 3 to 6 classes of Ti and Zn (with coordinates)*

Рис. 9. *Профиль кластеров с 3 по 6 классы Ti и Zn (с координатами)*

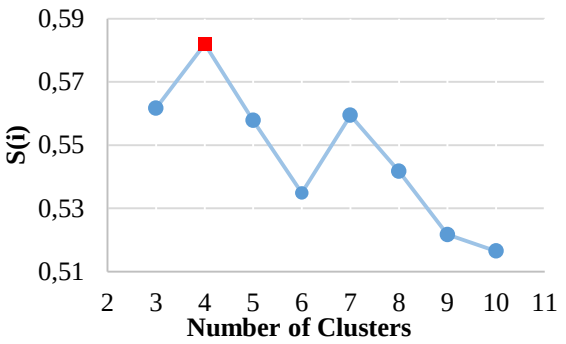

Fig. 10. *Change in the validation of S(i) value based on the number of clusters (for Ti and Zn elements with coordinates)*

Рис. 10. *Изменение значения проверки S(i) в зависимости от количества кластеров (для элементов Ti и Zn с координатами)*

Table 2. *Normalized characteristics of cluster centers*

Таблица 2. *Нормированные характеристики кластерных центров*

Length	Width	Zn	Ti	Class
1	0,105046	0,874296	0,09511	First
0,087734	0	1	0	Second
0,899769	1	0,709867	1	Third
0	0,824349	0	0,338263	Forth

The prediction of the titanium grade

In this section, a relationship between Ti and Zn and coordinates using GRNN that is a type of ANN has been determined according to the length and width of the samples selected from all samples, so that the elements grade can be estimated using obtained relationship.

The titaniun value is introduced to the Matlab 2014 as an output variable and the values of zinc and length, and width of the points are introduced as input variables. The ratio of training data to experimental (test) data is 30 % which are randomly selected from all data.

The regression method with the K-means cluster centers was used to predict Ti. For the multi-variable regression (equation (10)) in the SPSS software according to the length and width of the samples selected from cluster centers, so that the element grade can be estimated using obtained relationship.

The titaniun K-means cluster centers value is introduced to the SPSS software as a dependent variable and the cluster centers values of zinc and length, and width of the points are introduced as independent variables. Then, the results of Table 3 were reported as characteristics and multi-variable regression coefficients of equation (14) were calculated.

The estimation with all samples

Due to the need to determine the optimal radius of estimation in this method, different values from 0 to 1 were selected experimentally. The optimal value of 0,015 was selected for the impact radius.

Fig. 11 shows the estimated continuous line and actual point scores in the training data. Also, Fig. 12 shows the same thing on test data.

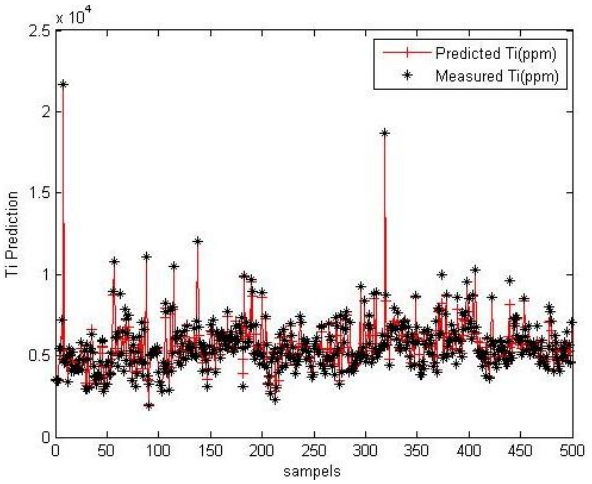

Fig. 11. *Titanium estimation line with real values in training data*

Рис. 11. *Линия оценки Ti с реальными значениями в обучающих данных*

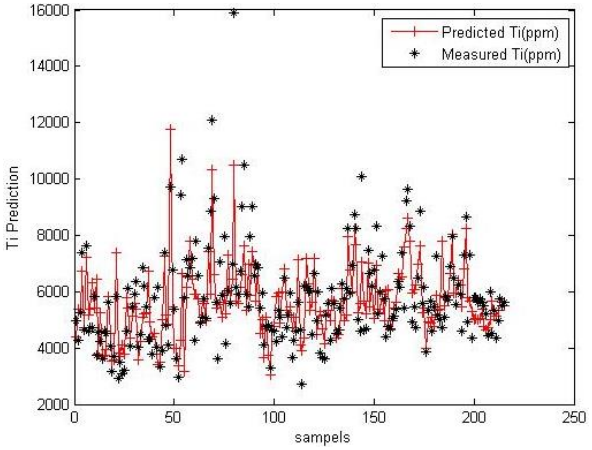

Fig. 12. *Titanium estimation line with real values in test data*

Рис. 12. *Линия оценки Ti с реальными значениями в тестовых данных*

In the better view of the estimate accuracy, the estimated values versus the actual ones in two categories of educational and experimental data are given in Fig. 13, 14, respectively. The accuracy of these estimates in educational data was set at 0,97 and for experimental data at 0,73.

The estimation with K-means cluster centers

The formula of the multiple regression line is determined (equation (14)) according to the coefficients in Table 3.

$$y=a_1x_1+a_2x_2+a_3x_3+b. \quad (14)$$

Table 3. *Regression coefficient*

Таблица 3. *Коэффициент регрессии*

a_1	a_2	a_3	b
–0,0089	–0,0811	399,22	318994

$$Ti = –0,0089X–0,0811Y+399,22Zn+318994. \quad (15)$$

The obtained R-value represents a parabola equation that the regression model has been able to explain the variations in terms of y (titanium). Here, the R-value is 70

and therefore, 70 % of variations of titanium values are due to Xs (i. e. Zn, length and width of the sampling points). To validate the titanium grade estimation based on the obtained equation (15), a number of actual data should be compared with the values obtained from the equation to measure the accuracy of the estimator. Therefore, 30 % of the samples are randomly separated before the multi-variable regression based on the values of the Zn element and the length and width of the sampling points and replacing it in equation (15). Then the values of the element Ti are estimated and compared with its actual values according to the samples. The results are shown in Fig. 15 as a dispersion diagram. Fig. 15 shows the correlation between the real and estimated grades with a correlation coefficient of 51 % which indicates the relative accuracy of the used method.

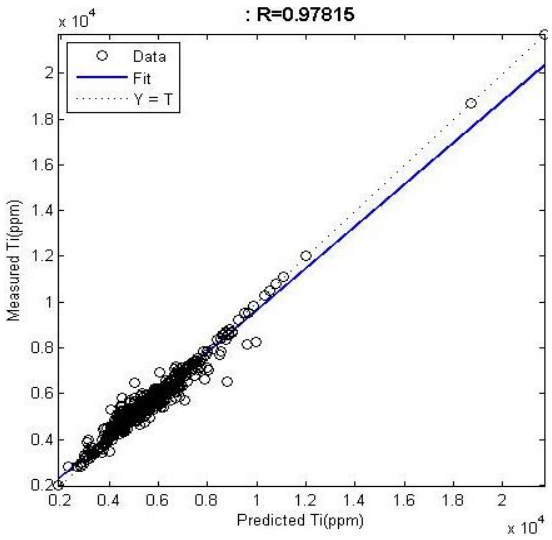

Fig. 13. *Estimated vs. actual data (test)*
Рис. 13. *Расчетные и фактические данные (тест)*

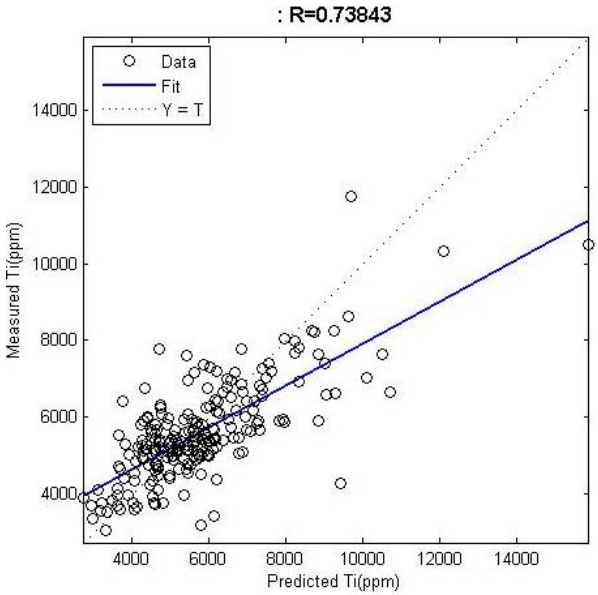

Fig. 14. *Estimated vs. actual data (training)*
Рис. 14. *Расчетные и фактические данные (обучение)*

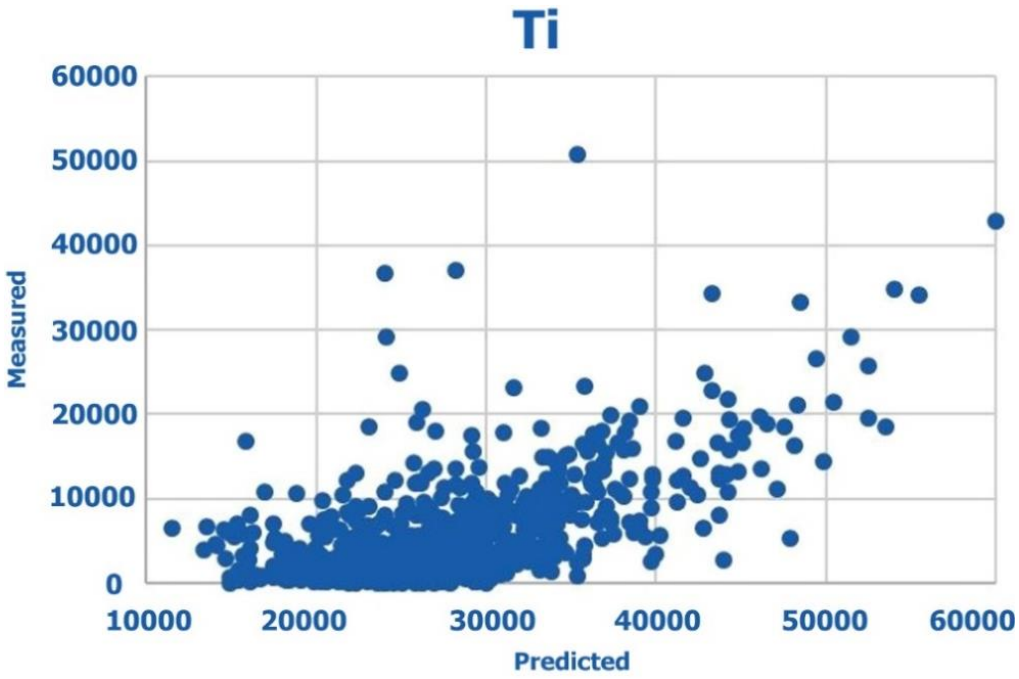

Fig. 15. *Estimated versus actual value of Ti*
Рис. 15. *Расчетное значение Ti по сравнению с фактическим значением*

Considering the actual titanium amount with the Kriging method, the area map (Fig. 16) is drawn to compare with the resulting map with the grades obtained from equation (12) (Fig. 17). The point of these maps is the normalization of all of the parameters. For this reason, it can only be used visually for the accuracy of the estimation.

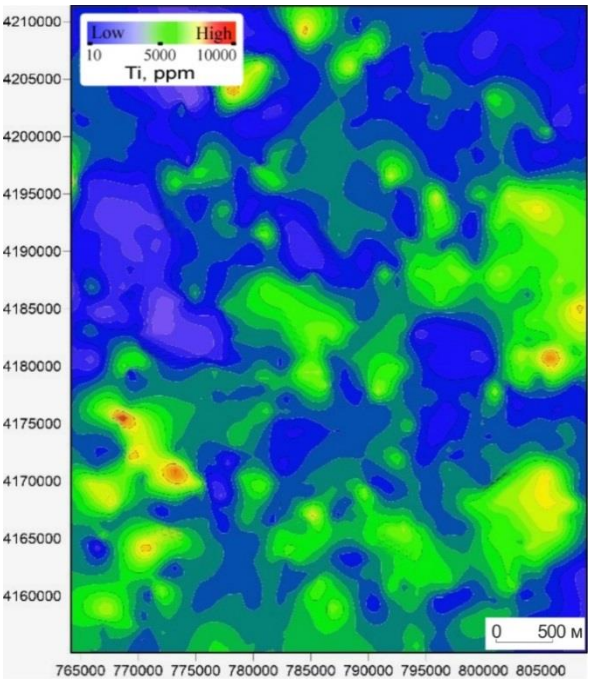

Fig. 16. *Schematic representation of actual Ti grades with the Kriging method on the map*

Рис. 16. *Схематическая карта аномалий фактического содержания Ti по методу Кригинга*

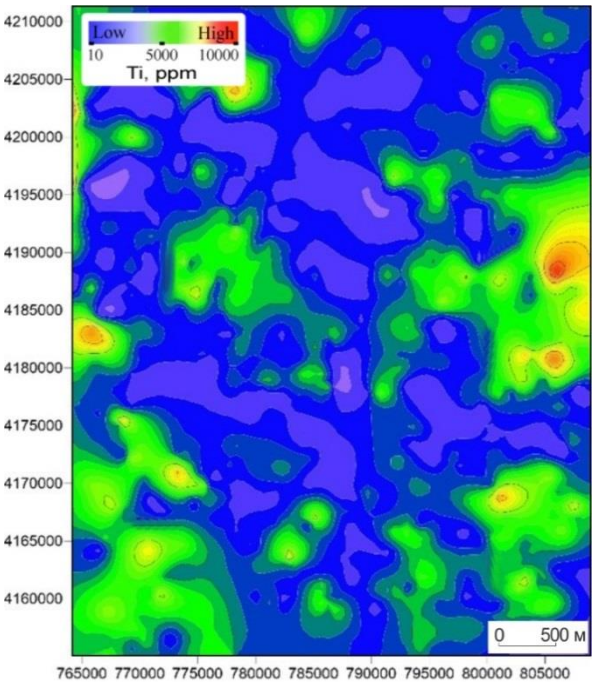

Fig. 17. *Schematic representation of predicated Ti grades on the map*

Рис. 17. *Схематическая карта распространения предполагаемых классов Ti*

Conclusion

The relationship between the elements Ti and Zn was determined using this K-means method, taking into account latitude and longitude samples taken to more accurately assess the appearance and extent of geochemical halos in the study area. According to the results obtained during processing of these elements, a regression equation was drawn up to estimate the titanium content based on three parameters: Zn content, the length and width of the sampling points, the correlation coefficient. According to the K-means cluster centers and artificial neural network, the Ti element grade was predicted and the correlation coefficient was reported 0,51. Both methods produce the desired results, but the artificial neural network method has more accurate data. Schematic maps of the initial and predicted Ti content were constructed. The results of the study can be used in the course of geological exploration to forecast and identify new promising areas.

The research was financially supported by the RFBR grant no. 18-45-700019 and within the framework of a Competitiveness Enhancement Program Grant at Tomsk Polytechnic University.

REFERENCES

1. Osterholt V., Dimitrakopoulos R. Simulation of orebody geology with multiple-point geostatistics – application at Yandi Channel Iron Ore Deposit, WA, and implications for resource uncertainty. *Advances in Applied Strategic Mine Planning*. Cham, Springer International Publ., 2018. pp. 335–352.
2. Ozkan E., Iphar M., Konuk A. Fuzzy logic approach in resource classification. *International Journal of Mining, Reclamation and Environment*, 2019, vol. 33, no. 3, pp. 183–205.
3. Alahgholi S., Shirazy A., Shirazi A. Geostatistical studies and anomalous elements detection, Bardaskan Area, Iran. *Open Journal of Geology*, 2018, vol. 8, no. 7, pp. 697–710.
4. Khakmardan S., Shirazi A., Shirazy A., Hosseingholi H. Copper oxide ore leaching ability and cementation behavior, mesgaran deposit in Iran. *Open Journal of Geology*, 2018, vol. 09, no. 8, pp. 841–858
5. Carranza E.J.M., Zuo R. Introduction to the thematic issue: analysis of exploration geochemical data for mapping of anomalies. *Geochemistry: Exploration, Environment, Analysis*, 2017, vol. 17, no. 3, pp. 183–185.
6. Shirazy A., Shirazi A., Ferdossi M.H., Ziaii M., Adel S., Aref S. Mohammad H.F., Mansour Z. Geochemical and geostatistical studies for estimating gold grade in tarq prospect area by k-means clustering method. *Open Journal of Geology*, 2019, vol. 9, no. 6, pp. 306–326.
7. Shirazi A., Shirazy A., Karami J. Remote sensing to identify copper alterations and promising regions, Sarbishe, South Khorasan, Iran. *International Journal of Geology and Earth Sciences*, 2018, vol. 4, no. 2, pp. 36–52.
8. Abbas S., Rahimi G.H., Najmodini M. Recognition of Cu-porphyry mineralization areas by using one and multivariate integration methods on drainage geochemical data in Ghale Askar area, Kerman province. *Journal of Analytical and Numerical Methods in Mining Engineering*, 2013, vol. 3, no. 6, pp. 69–82.
9. Ghannadpour S.S., Hezarkhani A., Farahbakhsh E. An investigation of Pb geochemical behavior respect to those of Fe and Zn based on k-Means clustering method. *Journal of Tethys*, 2013, vol. 1, no. 4, pp. 291–302.
10. Malyszko D., Wierzchon S.T. Standard and genetic k-means clustering techniques in image segmentation. *6th International Conference on Computer Information Systems and Industrial Management Applications (CISIM'07)*. NW Washington, United States, 2007. pp. 299–304.
11. Abolhassani B., Salt J. A simplex K-means algorithm for radioport placement in cellular networks. *Canadian Conference on Electrical and Computer Engineering*. Canada, University of Saskatchewan, 2005. pp. 2117–2121.
12. Chen T.-W., Chien S.-Y. Bandwidth adaptive hardware architecture of K-Means clustering for intelligent video processing. *IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan, 2009. pp. 573–576.

13. Yang J., Zhuang Y., Wu F. ESVC-based extraction and segmentation of texture features. *Computers & Geosciences*, 2012, vol. 49, pp. 238–247**.**
14. Mora J.L., Armas-Herrera C.M., Guerra J.A., Rodríguez-Rodríguez A., Arbelo C.D. Factors affecting vegetation and soil recovery in the Mediterranean woodland of the Canary Islands (Spain). *Journal of Arid Environments*, 2012, vol. 87, pp. 58–66**.**
15. Meshkani S.A., Mehrabi B., Yaghubpur A., Alghalandis Y.F. The application of geochemical pattern recognition to regional prospecting: a case study of the Sanandaj–Sirjan metallogenic zone, Iran. *Journal of Geochemical Exploration*, 2011, vol. 108, no. 3, pp. 183–195**.**
16. Sfidari E., Kadkhodaie-Ilkhchi A., Najjari S. Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. *Journal of Petroleum Science and Engineering*, 2012, vol. 86–87, pp. 190–205**.**
17. Isaeva E.R., Voroshilov V.G., Timkin T.V., Mansour Ziaii. Geochemical criteria to identify reservoirs and to forecast their oil and gas content in terrigenous deposits in Pur-Tazovskoy oil-bearing field. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering*, 2018, vol. 329, no. 4, pp. 132–141. In Rus.
18. Wegner T., Hussein T., Hämeri K., Vesala T., Kulmala M., Weber S. Properties of aerosol signature size distributions in the urban environment as derived by cluster analysis. *Atmospheric Environment*, 2012, vol. 61, pp. 350–360**.**
19. Shirazy A., Ziaii M., Hezarkhani A., Timkin T. Geostatistical and remote sensing studies to identify high metallogenic potential regions in the Kivi area of Iran. *Minerals*, 2020, vol. 10, pp. 1–25.
20. *Geological Report of Kivi 1:100000 sheet*. Tehran, GSI, IRAN, 2006. In Persian.
21. Zohrab E. *Geological Map of Kivi 1:100000 (on scale), Sheet 5665*. Geological Survey and Mineral Exploration of Iran (GSI). Tehran, GSI, IRAN, 2006. In Persian.
22. *Geochemical map of Kivi 1:100000 (on scale)*. Tehran, GSI, IRAN, 2006. In Persian.
23. Hassanipak A.A., Sharafeddin M. *Exploration Data Analysis*. Tehran, Iran, Tehran University Press, 2005. Vol. 1, 352 p. In Persian.
24. Timkin T., Voroshilov V., Yanchenko O., Suslov J., Korotchenko T. Geology, geochemistry and gold-ore potential assessment within Akimov ore-bearing zone (the Altai Territory). *IOP Conference Series: Earth and Environmental Science*, 2016, vol. 43, pp. 1–5.
25. Timkin T., Voroshilov V., Askanakova O., Cherkasova T., Chernyshov A., Korotchenko T. Estimating gold-ore mineralization potential within Topolninsk ore field (Gorny Altai). *IOP Conference Series: Earth and Environmental Science*, 2015, vol. 27, pp. 1–5.
26. Ziaii M., Safari S., Timkin T.V., Voroshilov V., Yakich T. Identification of geochemical anomalies of the porphyry–Cu deposits using concentration gradient modelling: a case study, Jebal-Barez area, Iran. *Journal of Geochemical Exploration*, 2019, vol. 199, pp. 16–30.
27. Grubbs F.E. Procedures for detecting outlying observations in samples. *Technometrics*, 1969, vol. 11, no. 1, pp. 1–21.
28. Madala G.S. *Introduction to Econometrics*. 2nd ed. New York, NY, USA, Maxmillan Publ. Company, 1992. 631 p.
29. Kalisch M., Michalak M., Sikora M., Wróbel Ł., Przystalka P. Influence of outliers introduction on predictive models quality. *Proc. of the 12th International Conference, BDAS 2016.* Ustroń, Poland, 2016. Vol. 613, pp. 79–93.
30. Shirazi A., Shirazy A., Saki S., Hezarkhani A. Geostatistics studies and geochemical modeling based on core data, sheytoor iron deposit, Iran. *Journal of Geological Resource and Engineering*, 2018, vol. 6, pp. 124–133**.**
31. Hassani Pak A.A. *Geostatistical*. 5th ed. Tehran, University of Tehran press, 2013. 328 p. In Persian.
32. Shirazy A., Shirazi A., Hezarkhani A. Predicting gold grade in Tarq 1:100000 geochemical map using the behavior of gold, Arsenic and Antimony by K-means method. *Journal of Minerals Resources Engineering*, 2018**,** vol. 2, no. 4, pp. 11–23.
33. Zhou S., Zhou K., Wang J., Yang G., Wang S. Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. *Frontiers of Earth Science*, 2018, vol. 12, no. 3, pp. 491–505**.**
34. Jain A.K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010, vol. 31, no. 8, pp. 651–666**.**
35. Saha S., Bandyopadhyay S. A generalized automatic clustering algorithm in a multiobjective framework. *Applied Soft Computing*, 2013, vol. 13, no. 1, pp. 89–108**.**
36. Hezarkhani A., Ghannadpour S.S. *Geochemical behavior investigation based on K-means clustering: basics, concepts and case study*. Tehran, LAP LAMBERT Academic Publ., 2015. 60 p.
37. Mahvash M.N., Hezarkhani A. Estimation of grade gold in khooni deposit using the behavior of gold, arsenic and antimony elements by clustering k-means method. *Journal Analytical and Numerical Methods in Mining Engineering*, 2015, vol. 5, no. 10, pp. 77–92. In Persian.
38. Shirazi A., Shirazy A., Saki S., Hezarkhani A. Introducing a software for innovative neuro-fuzzy clustering method named NFCMR. *Global Journal of Computer Sciences: theory and research*, 2018, vol. 8, no. 2, pp. 62–69.
39. Specht D.F. A general regression neural network. *IEEE Trans. Neural Networks*, 1991, vol. 2, no. 2, pp. 568–576**.**
40. Beale M.Y., Hagan M.T., Demuth H.B. *Neural network toolbox user's guide*. Natick, The MathWorks, 2010, 846 p.
41. Rooki R. Application of general regression neural network (GRNN) for indirect measuring pressure loss of Herschel–Bulkley drilling fluids in oil drilling. *Measurement*, 2016, vol. 85, pp. 184–191.
42. Grosan C., Abraham A. *Artificial Neural Networks*. Springer, Berlin, Heidelberg, 2011. pp. 281–323.
43. Menard J.J. Relationship between altered pyroxene diorite and the magnetite mineralization in the Chilean Iron Belt, with emphasis on the El Algarrobo iron deposits (Atacama region, Chile). *Mineralium Deposita*, 1995, vol. 30, pp. 268–274 **.**
44. Xu L., Bi X., Hu R., Zhang X., Su W., Qu W., Hu Z., Tang Y. Relationships between porphyry Cu–Mo mineralization in the Jinshajiang–Red River metallogenic belt and tectonic activity: Constraints from zircon U–Pb and molybdenite Re–Os geochronology. *Ore Geology Reviews*, 2012, vol. 48, pp. 460–473**.**
45. Shin H.W., Sohn S.Y. Segmentation of stock trading customers according to potential value. *Expert Systems with Applications*, 2004, vol. 27, no. 1, pp. 27–33**.**
46. Tarkian M., Stribrny B. Platinum-group elements in porphyry copper deposits: a reconnaissance study. *Mineralogy and Petrology*, 1999, vol. 65, pp. 161–183**.**

Received: 23 December 2020.
Поступила 23.12.2020.

Information about the authors

Adel Shirazy, PhD, assistant, Shahrood University of technology.

Mansour Ziaii, PhD, associate professor, Shahrood University of Technology.

Ardeshir Hezarkhani, PhD, full professor, Mining Engineering Department, Amirkabir University of Technology.

Timofey V. Timkin, Cand. Sc., associate professor, National Research Tomsk Polytechnic University.

Valery G. Voroshilov, Dr. Sc., professor, National Research Tomsk Polytechnic University.

УДК 550.4+550.84.09

ИССЛЕДОВАНИЕ ГЕОХИМИЧЕСКОГО ПОВЕДЕНИЯ ТИТАНА И ЦИНКА НА ОСНОВЕ МЕТОДА К-СРЕДНИХ И ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ НОВЫХ ПЛОЩАДЕЙ, РЕГИОН КИВИ, ИРАН

Адель Ширази[1],
Adel.shirazy@shahroodut.ac.ir

Мансур Зиаии[1],
mziaii@shahroodut.ac.ir

Ардешир Хезархани[2],
Ardehez@aut.ac.ir

Тимкин Тимофей Васильевич[3],
timkin@tpu.ru

Ворошилов Валерий Гаврилович[3],
v_g_v@tpu.ru

[1] Шахрудский технологический университет,
Иран, 3619995161, Шахруд, Болвар Данешка.

[2] Технологический университет им. Амир Кабира (Тегеранский политехнический институт),
Иран, 1591634311, Тегеран, Авеню Хафеза.

[3] Национальный исследовательский Томский политехнический университет,
Россия, 634050, г. Томск, пр. Ленина, 30.

Актуальность. *Это первые геохимические исследования в горнорудном районе Киви. Они необходимы в виду возможного наличия в районе перспективных месторождений титана и цинка. Сложность геологического строения определяет необходимость применения нетрадиционных методов исследования и прогнозирования – искусственных нейронных сетей и методов кластеризации – для оценки поведения химических элементов.*

Цель *заключается в определении геохимического поведения Ti и Zn для прогнозирования новых рудоносных площадей и перспективных участков.*

Объект: *район Киви в провинции Ардебиль, Иран (Иранский Азербайджан), геохимическая карта масштаба 1:100000.*

Методы. *Исходными данными послужили отобранные пробы из донных отложений района Киви, которые были проанализированы методом ICP-MS. Интерпретация геохимических данных проводилась с использованием одномерных и многомерных статистических методов, включая кластеризацию методом К-средних. Содержания Ti также были предсказаны с использованием искусственных нейронных сетей.*

Результаты. *Согласно результатам, полученным в процессе обработки геохимических данных, было составлено уравнение регрессии, которое представляет собой функцию для оценки содержания титана на основе трех параметров: содержания цинка, длины и ширины точек отбора проб, коэффициента корреляции. Согласно результатам исследования, были предсказаны концентрации Ti; коэффициент корреляции между исходными и предсказанными значениями составил 0,51. Метод искусственных нейронных сетей дает более точные данные, чем кластеризация методом К-средних. Были построены схематические карты исходных и предсказанных содержаний Ti. Результаты исследования можно использовать в процессе проведения геологоразведочных работ для прогнозирования и выявления новых перспективных площадей.*

Key words:
Титан, цинк, регион Киви, метод кластеризации К-средних, искусственные нейронные сети, предсказание концентраций элементов.

Работа выполнена при финансовой поддержке РФФИ (грант № 18-45-700019) и в рамках гранта Программы повышения конкурентоспособности Томского политехнического университета.

Информация об авторах

Ширази А., PhD, ассистент Шахрудского технологического университета.

Зиаии М., PhD, доцент Шахрудского технологического университета.

Хезархани А., PhD, професор Технологического университета им. Амир Кабира (Тегеранский политехнический институт).

Тимкин Т.В., кандидат геолого-минералогических наук, доцент отделения геологии Инженерной школы природных ресурсов Национального исследовательского Томского политехнического университета.

Ворошилов В.Г., доктор геолого-минералогических наук, профессор отделения геологии Инженерной школы природных ресурсов Национального исследовательского Томского политехнического университета.