



Journal of Mining and Environment (JME)

journal homepage: www.jme.shahroodut.ac.ir



Application of Probabilistic Clustering Algorithms to Determine Mineralization Areas in Regional-Scale Exploration Studies

Hamid Geranian^{1*} and Zahra Khajeh Miry²

1. Department of Mining Engineering, Birjand University of Technology, Birjand, Iran

2. Industry, Mine & Trade Organization of South Khorasan Province, Birjand, Iran

Article Info

Received 6 July 2020

Received in Revised form 10 October 2020

Accepted 13 October 2018

Published online 4 November 2020

DOI: [10.22044/jme.2020.9867.1910](https://doi.org/10.22044/jme.2020.9867.1910)

Keywords

Probabilistic clustering algorithms

Composite geochemical anomaly

Geochemical potential mapping

Hydrothermal alterations

Deh-Salm quadrangle

Abstract

In this work, we aim to identify the mineralization areas for the next exploration phases. Thus, the probabilistic clustering algorithms due to the use of appropriate measures, the possibility of working with datasets with missing values, and the lack of trapping in local optimal are used to determine the multi-element geochemical anomalies. Four probabilistic clustering algorithms, namely PHC, PCMC, PEMC, PDBSCAN, and 4138 stream sediment samplings, are used to divide the samples into the three clusters of background, possible anomaly, and probable anomaly populations. In order to determine these anomalies, ten and eight metal elements are selected as the chalcophile and siderophile elements, respectively. The results obtained show the areas of approximately 500 and 5,000 km² as the areas of the probable and possible anomalies, respectively. The composite geochemical anomalies of the chalcophile and siderophile elements are mostly dominant in the metamorphic-acidic-intermediate rock units and the alkaline-metamorphic-intermediate rock units of the studied area, respectively. Besides, the obtained anomalies of the four clustering algorithms also cover about 65% of the mineralized areas, all mines, and almost 60% of the alteration areas. The validity criterion of the clustering methods show more than 70% validity for the obtained anomalies. The results obtained indicate that the probabilistic clustering algorithms can be an appropriate statistical tool in the regional-scale geochemical explorations.

1. Introduction

Geochemical sampling of stream sediments at the 1:100000 scale is one of the most important regional exploration operations, especially in countries with dry climates such as Iran. The goal is to determine the promising areas for the prospecting phase of metallic minerals. It is clear that these areas are distinguished by the integration of the geological, geophysical, geochemical, and remote sensing data [1]. The anomaly areas, especially geochemical anomalies, are the results of the integration. If the locations and extents of geochemical anomalies are determined more exactly, the risk of exploration will be smaller in the next phase.

Threshold determination to specify an anomaly is a statistical-geological process. As a whole, the

threshold determination methods are broadly classified into two categories. The first one comprises the non-structural methods that focus on the frequency distribution of element concentration. Statistical parameter methods [2], the probability of the appearance multiplied by the sample number method [3], the Mahalanobis distance method [4], the graphical methods such as probability plot, box plot, and chi-square plot [5], and the disjunctive statistics method [6] belong to this category. The second category comprises the structural methods that in addition to emphasizing the frequency distribution of element concentration, consider the spatial variability and correlation of the element concentration. The methods such as the U-spatial statistics [7], spatial

Corresponding author: h.geranian@birjandut.ac.ir (H. Geranian).

filtering [8], moving average [9], fractal models [10], and kriging [11] are examples of the second category. If the formation of minerals occurs through simple geological phases and the mineralization area is not subsequently influenced by different geological phenomenon, the methods in the first category are more reliable in identifying the anomalous areas; otherwise, the methods in the second category will be more exact in this respect.

The threshold determination methods, in terms of the number of elements, can also be divided into two groups: (1) threshold determination methods for uni-elements; and (2) threshold determination methods for two elements and multi-elements. While all of the methods listed in the previous paragraph belong to the first group, multivariate statistical, clustering, and fractal methods belong to the second group [12-14]. Since the geochemical samples are often multi-element analyzed, the use of the methods from the second group is superior in this respect. In this case, the anomaly obtained will be a multi-element or composite geochemical anomaly, indicating an exploration potential or a mineralization area for several elements. There are a great number of published scientific papers dealing mostly with the first group of methods but few publications on the second group.

The clustering methods were first used by Collyer and Merriam (1973) for the resolution of similar deposits based on the geological variables [15]. Then Roy (1981) used the clustering analysis for the separation of elements related to the mineralization area [16]. Hierarchical clustering [17], fuzzy clustering [19], k-means clustering [18], and mixture-model clustering [20] algorithms are the most important clustering methods that have been used in the geochemical exploration research works. These algorithms have also been used in both the R and Q models.

The aim of this paper is to recommend new probabilistic clustering algorithms for the determination of composite geochemical anomalies. Four probabilistic clustering algorithms, namely the probabilistic hierarchical clustering (PHC), probabilistic c-mean clustering (PCMC), probabilistic expectation maximization clustering (PEMC), and probabilistic density-based spatial clustering of applications with noise (PDBSCAN), in the Q model, will be employed to recommend areas with a potential for multi-metal exploration. For this purpose, the stream sediment geochemical data of six sheets at the 1:100000 scale (i.e. the Deh-Salm quadrangle) in the southern part of the South Khorasan province in the east of Iran is used. This region is one of the most

important areas of exploration in Iran due to the presence of volcanic and plutonic rocks belonging to the Tertiary geological era. There is a hope that several world-class deposits in this part of Iran will be introduced in the forthcoming years.

2. Probabilistic clustering methods

Clustering is a process through which a series of samples is divided into several clusters in such a way that the samples in each cluster are very similar to one another, while the cluster has the lowest possible similarity [21]. In other words, clustering is an unsupervised data mining method. The most important measure of similarity in the clustering methods is the distance between the samples in the spatial data cloud. The clustering methods are divided into four groups, namely the linkage-based, centroid-based, frequency function-based, and density-based techniques [22]. Each one of these techniques can cluster the dataset into either the algorithmic or the probabilistic method. In the algorithmic methods, the linkage measures are used, which are simpler and more effective in most cases. However, these methods suffer from three serious disadvantages: firstly, deselection of the distance measure; secondly, the inability to cluster the dataset of which some variables are not measured; and thirdly, local clustering of the data [22]. Nevertheless, application of the probabilistic clustering methods can compensate for some of these defects.

A generating model is used in the probabilistic clustering algorithms. Thus it is supposed that in these algorithms, the data of each cluster have been generated by a probability distribution function (for instance by a Gaussian, Bernoulli or other distribution functions), a part of which exists in a data cluster. Suppose the multi-dimensional dataset is defined as follows:

$$D = \{x_1, x_2, \dots, x_n\} \quad (1)$$

where x_i is a multi-dimensional data. If the dataset is generated by the multi-dimensional Gaussian distribution functions, the probability that a sample such as $x_i \in D$ is generated by these models is equal to [39]:

$$p(x_i | \mu_k, \Sigma_k) = p(Q_{ik}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp \left\{ -\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right\} \quad (2)$$

$$i = 1, 2, \dots, n \quad k = 1, 2, \dots, c$$

where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$, the parameters of the generating model, are the mean vector and

variance-covariance matrix of the k th cluster, respectively. $p(Q_{ik})$ is the probability of the i th sample belonging to the k th cluster and c is the number of clusters. Consequently, the value of the likelihood of this dataset being generated by these models is equal to:

$$L(N(\mu_k, \Sigma_k): D) = \prod_{i=1}^n p(x_i | \mu_k, \Sigma_k) \quad (3)$$

Therefore, in the probabilistic clustering algorithms, we look for the model parameters (i.e. μ_0 and Σ_0) whose likelihood value is maximized as follows [36]:

$$N(\mu_0, \Sigma_0) = \arg \max \{L(N(\mu_k, \Sigma_k): D)\} \quad (4)$$

2.1. PHC algorithm

The probabilistic hierarchical clustering (PHC) algorithm is one of the linkage-based techniques that use the connection method between the samples. The samples that are more similar to each other are linked by the agglomerative (bottom-up merging) or divisive (top-down merging) methods [17, 21]. The agglomerative probabilistic hierarchical clustering method is started by letting each sample in a cluster. Then the two Q_i and Q_j clusters are merged if the distance between them is not negative. This distance can be calculated using the following equation [22, 24]:

$$dist(Q_i, Q_j) = \arg \min_{i \neq j} - \log \frac{p(Q_i \cup Q_j)}{p(Q_i)p(Q_j)} \quad (5)$$

Merging of clusters continues as long as the value of $\log \frac{p(Q_i \cup Q_j)}{p(Q_i)p(Q_j)}$ is greater than zero. In other words, the merging of clusters continues as long as it improves the quality of clustering (i.e. as long as a new cluster derived from the merging of two clusters better matches a distribution function, like the Gaussian distribution function). The mean vector and variance-covariance matrix are recalculated in each iteration for the newly generated clusters.

2.2. PCMC algorithm

The probabilistic c-mean clustering (PCMC) algorithm is one of the centroid-based clustering techniques that was first introduced by Krishnapuram and Keller (1993) [25]. The PCMC algorithm is based on the fuzzy c-mean clustering method, and is proposed to overcome the difficulty of clustering the outliers [26]. In this algorithm, the membership degree that broadly represents the probability of each sample belonging to one cluster is used. In order to calculate the membership

degree, one should look for ways to minimize the following error objective function [26]:

$$J(U_{ki}, \mu_k; D) = \sum_{k=1}^c \sum_{i=1}^n u_{ki}^m d^2(x_i, \mu_k) + \quad (6)$$

$$\sum_{k=1}^c \eta_k \sum_{i=1}^n (1 - u_{ki})^m$$

where u_{ki} is the probability of the membership degree of the i th sample in the k th cluster, $d^2(x_i, \mu_k)$ is the square distance measure of the i th sample from the center of the k th cluster, m is the fuzzifier parameter (often considered as $m = 2$), and η_k is the resolution or scale parameter, which should be calculated for each cluster. Equation (6) has the following constraints:

$$0 \leq u_{ki} \leq 1 \quad k = 1, 2, \dots, c \quad i = 1, 2, \dots, n \quad (7)$$

$$\sum_{i=1}^n u_{ki} > 0 \quad k = 1, 2, \dots, c \quad (8)$$

The inequalities (7) and (8) represent, respectively, the removal of the normalization conditions of the membership degree and not being null (any cluster of the samples). Since it is not possible to minimize Eq. (6) directly, the iteration procedures are used to solve this problem. In this case, the degree of membership, the center of each cluster, and the scale parameter are updated using the following equations [26]:

$$u_{ki} = \left(1 + \left(\frac{d^2(x_i, \mu_k)}{\eta_k}\right)^{1/(m-1)}\right)^{-1} \quad (9)$$

$$k = 1, 2, \dots, c \quad i = 1, 2, \dots, n$$

$$\mu_k = \frac{\sum_{i=1}^n u_{ki}^m x_i}{\sum_{i=1}^n u_{ki}^m} \quad k = 1, 2, \dots, c \quad (10)$$

$$\eta_k = K \frac{\sum_{i=1}^n u_{ki}^m d^2(x_i, \mu_k)}{\sum_{i=1}^n u_{ki}^m} \quad k = 1, 2, \dots, c \quad (11)$$

where K is a positive constant factor usually considered to be close to one [26]. Applying the PCMC algorithm is considered as the first stage, and is done by selecting the number of clusters and a random selection of the cluster centers. This is followed by the iterative steps including calculating the distance measure of each sample from its cluster center and updating the membership degree, the center of each cluster, and also the scale parameter. If the following inequality is established, the iteration steps will stop, and the

membership degree matrix of the penultimate stage (i.e. U_{ki}^t) will be considered as the answer [27].

$$\max \|U_{ki}^{t+1} - U_{ki}^t\| < \varepsilon \quad (12)$$

where ε is the stopping criterion defined by the user.

2.3. PEMC algorithm

The probabilistic expectation maximization clustering (PEMC) algorithm, known as the random expectation maximization, was first introduced by Celeux and Diebolt (1985) [28]. The PEMC algorithm is one of the frequency-function-based techniques that seeks the maximum likelihood estimate by optimizing the estimates of the statistical parameters of the models. This process is achieved in two steps: (i) an expectation step (E-Step); and (ii) a maximization step (M-step) [29]. The PEMC has four advantages compared to the conventional EM: (a) it has a higher speed in each iteration; (b) it provides optimal solutions with missing data; (c) it does not become trapped in local optima; and (d) it provides better estimates of the statistical parameters [23]. The PEMC algorithm includes the following iterative steps [23, 29-30]:

1. The initial random selection of the probability model parameters for each cluster is as follows:

$$\theta_k = N(\mu_k, \Sigma_k) \quad k = 1, 2, \dots, c \quad (13)$$

2. E-step: Calculation of the probability of each sample belonging to each cluster with the following formula:

$$Z_{ki} = p(\theta_k | x_i) = \frac{p(x_i | \theta_k)}{\sum_{k=1}^c p(x_i | \theta_k)} = \frac{P(Q_{ki})}{\sum_{k=1}^c P(Q_{ki})} \quad (14)$$

$$i = 1, 2, \dots, n \quad k = 1, 2, \dots, c$$

The value of $P(Q_{ki})$ can be found by Eq. (2). Therefore, the samples belong to the cluster having the highest value of z_i .

3. M-Step: Calculation of the statistical parameters of each cluster according to the samples of each cluster to achieve the maximum likelihood using the following formulas:

$$\mu_k = \pi_k \frac{\sum_{i=1}^n x_i p(\theta_k | x_i)}{\sum_{i=1}^n p(\theta_k | x_i)} \quad k = 1, 2, \dots, c \quad (15)$$

$$\Sigma_k = \frac{\sum_{i=1}^n p(\theta_k | x_i) (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^n p(\theta_k | x_i)} \quad (16)$$

$$k = 1, 2, \dots, c$$

where π_k is the primary probability of each cluster, and can be computed by dividing the number of samples in each cluster into the total number of samples. If each cluster is supposed to contain the same number of samples, the primary probability will be $\pi_k = 1/c$; otherwise, this parameter will be updated in each iteration [31].

4. The iteration steps 2 and 3 until the convergence of clustering are obtained according to the following inequality or condition [23]:

$$\max \|Z_{ki}^{t+1} - Z_{ki}^t\| < \varepsilon$$

or (17)

$$\operatorname{argmax}\{L(\theta_k^t; D)\}$$

In this case, the matrix Z_{ki}^t will be the solution.

2.4. PDBSCAN algorithm

The probabilistic density-based spatial clustering of applications with noise (DBSCAN) algorithm, known as fuzzy logic, for uncertain data, was first introduced by Kriegel and Pfeifle (2005) as FDBSCAN [32]. It was later developed by Xu and Li (2008) as the probabilistic DBSCAN algorithm [33]. This algorithm is one of the density-based techniques possessing the potential advantages of the probabilistic clustering methods listed above, compared to the conventional DBSCAN method. In this method of clustering data, the samples with the density neighborhood are considered as a probabilistic object or sample core. Then by connecting the probabilistic core samples to the samples in the neighborhood, the density areas are obtained that are considered as a cluster. In this algorithm, ε , $MinPts$, and p parameters are used. The ε , $MinPts$, and p parameters are namely the neighborhood distance, neighborhood density, and probability thresholds, respectively. These parameters are defined by the user. It is necessary to provide the following definitions in order to employ this algorithm [33-34].

- I. **Neighborhood of a sample:** The uncertain samples x_i and x_j are in the neighborhood if the probability of their distance is less than ε and greater than p (i.e. $P(d(x_i, x_j) \leq \varepsilon) \geq p$). Thus the number of samples that can be in the neighborhood of the i th sample is equal to:

$$N_i(\varepsilon, p) = \{x_i \in D | P(d(x_i, x_j) \leq \varepsilon) \geq p \quad x_i \& x_j \in D\} \quad (18)$$

- II. **Probabilistic core sample:** An uncertain sample x_i is called the probabilistic core sample if the

number of samples located in its neighborhood is equal to at least *MinPts* (i.e. $|N_i(\epsilon, p)| \geq MinPts$).

III. Probabilistic density-reachable: An uncertain x_j sample is probabilistic directly density-reachable from the uncertain sample x_i if x_i is a probabilistic core sample and x_j is in the neighborhood of x_i . Besides, the uncertain sample x_k is probabilistic indirectly density-reachable by uncertain sample x_i if x_i and x_j are the probabilistic core samples, x_k is in the neighborhood of x_j , and x_j is in the neighborhood of x_i .

IV. Probabilistic density-connected: The uncertain sample x_j is probabilistic density-connected to uncertain sample x_i if both of them are probabilistic density-reachable by another sample.

The PDBSCAN algorithm includes the following steps:

1. A sample is randomly selected and the algorithm checks whether it is a probabilistic core sample.
2. If the selected sample is a probabilistic core sample, the first cluster is formed and all of the samples that are probabilistic directly or indirectly density-reachable from and/or probabilistic density-connected to this probabilistic core sample are assigned to this cluster; otherwise they are moved one step back.
3. Steps 1 and 2 are repeated until all the samples have been assigned.

In this algorithm, instead of calculating the expected distance, the minimum and maximum probabilistic distances of the samples are used. For this purpose, firstly, the R*-Tree index is created for all samples of the dataset (for more details, refer to Beckmann *et al.*, 2012; [35]). Then for each sample (e.g. x_i), if the minimum distance between the sample and the minimum bounding rectangle (MBR: A rectangle, oriented to the X and Y axes, which bounds a geographic dataset) is larger than ϵ , then the samples of MBR will be pruned. The remaining samples will be considered to be in the neighborhood of x_i . For each one of these remaining samples (e.g. x_j), the maximum probabilistic distance ($d_{max}(x_i, x_j)$), the minimum probabilistic distance ($d_{min}(x_i, x_j)$), and thus the estimates of the probability of its neighborhood can be calculated [33, 36].

3. Clustering validation methods

There are different criteria for evaluating the clustering methods, which can be grouped into the three types of external, internal, and relative [37-38]. Clustering validation of the first two criteria necessitates time-consuming calculations and

statistical tests, whereas the third criterion does not require the use of statistical tests [37]. The main idea of the relative criteria is to choose the best clustering plans based on the pre-specified criteria [39]. The external and internal criteria can only be used to validate the hard-clustering methods. Over the past years, several indices have been presented as the relative criteria, of which three indices are used in this work. In general, there is no significant difference in terms of superiority among these indices. What is of paramount importance is the validity of one algorithm according to several indices simultaneously. The modified Huber index (MHI) is one of the indices used in this work, which can be found as follows [39]:

$$MHI = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^n P_{ij} Q_{ij} \tag{19}$$

where $M = n(n-1)/2$, P is the proximity matrix of the dataset, and Q is the distance matrix from the center point of the clusters. These matrices can also be calculated as follows:

$$P_{ij} = d(x_i, x_j) \tag{20}$$

$$Q_{ij} = d(\mu_i, \mu_j) \quad i, j = 1, 2, \dots, n$$

where μ_i and μ_j are the central points of the clusters that the samples i and j belong, respectively. The higher the index is, the farther the central points of the clusters are, and consequently, the more compact the clusters are. Therefore, the clustering algorithm will be more valid.

The Davies-Bouldin index (DBI) is another utilized index whose validity is not dependent on the number of clusters or the clustering algorithm. In order to estimate this index, firstly, the measure of dispersion for each cluster must be calculated as follows [39]:

$$S_k = \sqrt{\frac{1}{n_k} \sum_{i=1}^{n_k} d^2(x_i, \mu_k)} \quad k = 1, 2, \dots, c \tag{21}$$

where n_k is the number of samples and μ_k is the center of the k th cluster. DBI is [40]:

$$DBI = \frac{1}{c} \sum_{k=1}^c \max_{k \neq l} \left\{ \frac{S_k + S_l}{d(\mu_k, \mu_l)} \right\} \tag{22}$$

$$l = 1, 2, \dots, k-1, k+1, \dots, c$$

DBI indicates the average of the similarity measure among the clusters. Therefore, a clustering

algorithm that has the least amount of similarity will be valid (i.e. minimum DBI).

The R-squared index (RSI), which is the same as the coefficient of determination, is another index of cluster validity. This index is the ratio of the sum of squares between groups (SS_b) to the total sum of squares of the whole dataset (SS_t). RSI is computed by the following formula [39, 41]:

$$RSI = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t} \quad (23)$$

where,

$$SS_w = \sum_{k=1}^c \sum_{i=1}^{n_k} (x_i - \mu_k)^2 \quad (24)$$

$$SS_t = \sum_{i=1}^n (x_i - \mu)^2$$

where μ is the mean vector of the whole dataset. This index value varies from 0 to 1. The closer this index is to one, the smaller the distribution of the data within a cluster is and the larger the distance between the clusters is, which results in a higher validity of clustering.

4. Geology of studied area

The studied area includes the Deh-Salm (or Chah-Vak) quadrangle, which is located between the coordinates of $58^{\circ} 30'$ to 60° east longitude and 31° to 32° north latitude. The area is located in the South Khorasan province in the eastern part of Iran. In terms of the geological-structural sub-divisions of Iran, this area is situated in the central part of the Lut Block, eastern Iran (Figure 1-A). The Lut Block zone extends over ~ 900 Km in the north-south direction and ~ 200 Km in the east-west direction, and is one of the sub-continentals of central Iran. This zone was separated from the northern parts of the Gondwanaland supercontinent during the opening of the Neo-Tethys in the Permian period. Then it was connected to the Eurasian supercontinent due to the closing of the Paleo-Tethys in the late Jurassic [42]. The Lut Block zone is limited by Nehbandan fault and Sistan zone in the east, Nayband fault and Tabas subzone in the west, Sabzevar subsidence in the north, and Urumieh-Dokhtar magmatic arc and Makran subduction zone in the south (Figure 1-A). The southern parts of the Lut Block are covered by a large area of salt flats and Quaternary sediments. The visible igneous rocks are only located on its southern margin, and are themselves part of the magmatic arc of the Makran subduction zone [43].

However, the strike-slip fault activities on both sides of the Lut Block in addition to the subduction of the Afghan Block beneath it have formed calc-alkaline rocks in large parts of the central and northern parts of the Lut Block zone [44-46]. On the Precambrian and late Jurassic metamorphosed bedrocks, volcanic, volcanioclastic, and sub-volcanic rocks and intrusive masses from the Late Cretaceous to the Quaternary are visible [47-48].

Figure 1-B shows a simplified geological map of the studied area. Amphibolite schist and mica schist metamorphic rocks of Precambrian age, which have been formed by metamorphism of acidic igneous rocks, are the oldest rock units of the studied area. The other part of the metamorphic rocks comprises hornfels, schist, and gneiss of the Lower Triassic and Jurassic, which are more visible in the central parts of the studied area (Figure 1-B). Intermediate igneous rocks include three groups including the Precambrian diorite rocks, Jurassic andesite and dacite rocks, and Cenozoic rocks. The third group contains mostly Eocene dacitic and Oligocene-Miocene semi-deep andesitic lavas. Moreover, the first and second group rocks exist in the form of small intrusive masses in the eastern part of the studied area. However, the third group of rocks forms an andesite band from north to south in the central part of the studied area. Besides, great dacite masses are formed in the southwestern part of the region (Figure 1-B). Alkaline igneous rocks are also formed from two rock groups, namely Cretaceous ultrabasic rocks such as gabbro and serpentinite in the northeastern part of the studied area and Oligocene alkaline basalts in the northwestern part of the studied area (Figure 1-B). A small part of the alkaline igneous rocks is also characterized as Pliocene basaltic rocks, recognized as the youngest igneous rocks of the studied area. Acidic igneous rocks are visible as granitic intrusive masses with pegmatitic texture from the low to upper Jurassic period and Eocene-Oligocene granite, and rhyolite rocks located in the southeastern, and middle parts of the studied area, respectively (Figure 1-B).

Crystalline Cherty limestone rocks of Precambrian age are the oldest sedimentary rock unit. Jurassic and Cretaceous sedimentary rocks include conglomerate, red-to-brown sandstone, shale, marl, and marly and sandy limestone. However, Tertiary sedimentary rocks mostly contain nummulitic fossils. As a whole, the sedimentary rocks are younger from the west to the east. This means that the Mesozoic sedimentary rocks outcrop mostly in the west, while the Tertiary sedimentary rocks outcrop mostly in the east of the

studied area. The intermediate to alkaline pyroclastic and tuffite rocks with an age of Paleogene to Neogene are generally colored green to gray, and are mostly abundant in the margin of the intermediate to alkaline igneous rocks and in

the western part of the studied area (Figure 1-B). Quaternary sediments are terraces, gravel fans, salt flats, and recent alluviums, which constitute a large portion of the studied area (Figure 1-B).

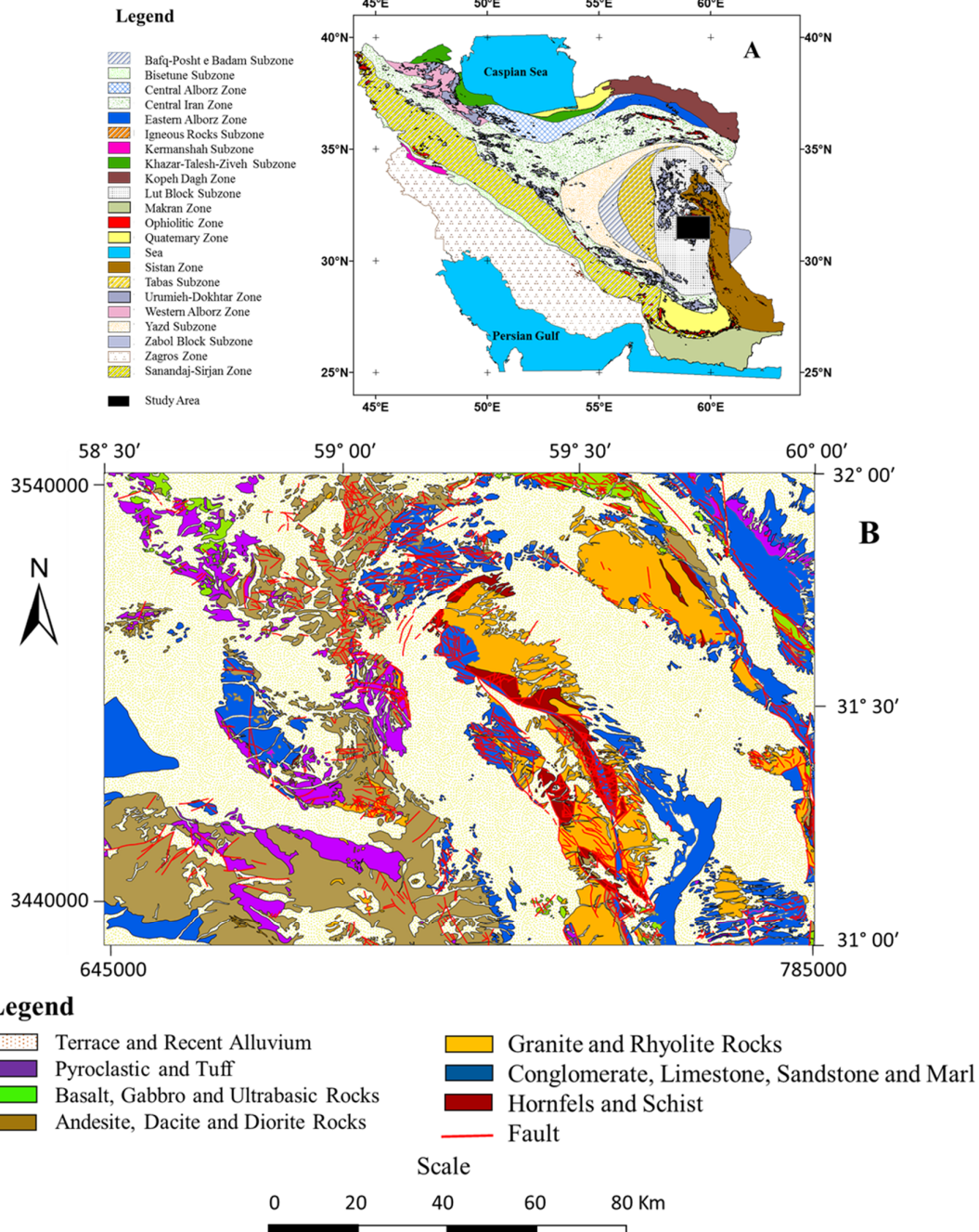


Figure 1. Location of the studied area with the geological-structural zones of Iran (A) (modified after Nabavi, 1976, and Stöcklin; 1968; [61-62]) and simplified geology map of the studied area (B) (modified after Deh-Salm quadrangle, Geological Survey of Iran, 1992).

The subduction of oceanic crust from both sides (under the Afghan Block on one side and under the Lut Block on the other) during the Paleocene to

Oligocene age caused magmatic and metamorphic activities, and thereby, formed metal and non-metal mineralization areas in the studied area [43, 47, 49-

50]. Qaleh-Zari copper mine is the most well-known mining index in this area. The genesis of this deposit is of IOCG type, possessing copper as well as economic gold and silver mineralization [51]. Other mines of the studied area include Mahor Cu deposit (Porphyry type), Kaviran Pb and Zn deposit, Hired Au deposit (associated with reduced granitoids), Chah-e-Zaghoo and Chah-Shalghami Cu-Au deposits (porphyry-epithermal type), and Chah-Kolub and Shah-Kuh Sn and W

deposits (associated with reduced granitoids and magmatic-skarn deposits) [51-54]. The genesis of iron ore deposits of the studied area is skarn such as Bisheh Fe deposits and placer. In addition to the above listed examples, the studied area includes about 125 mineralized areas with metal mineralization. Figure 2 depicts the locations of the mines with their names and the mineralized areas of the studied area.

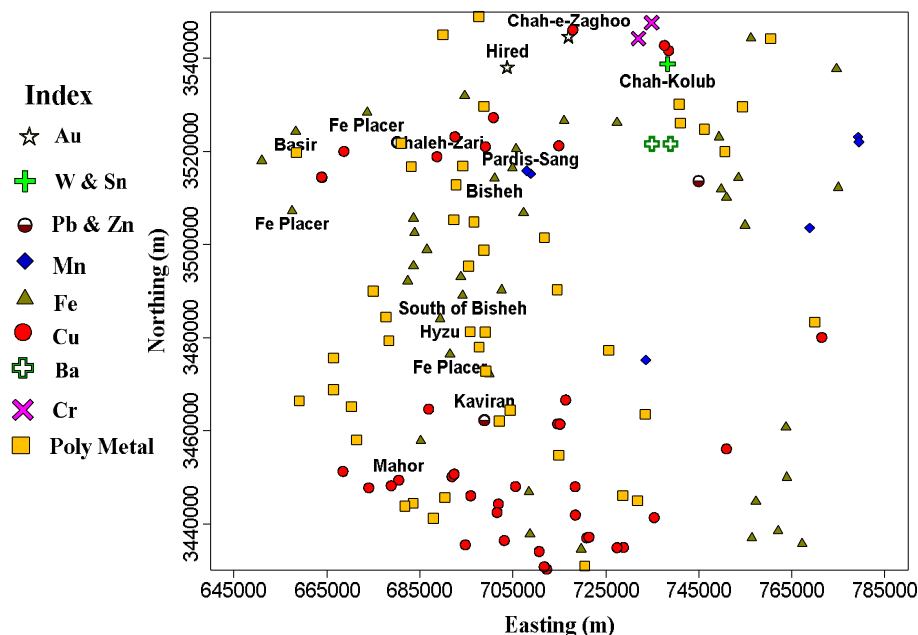


Figure 2. Location map of the mines and mineralized areas in the studied area.

5. Geochemical Dataset

The Deh-Salm quadrangle, with an area of 15,000 Km², is composed of six sheets, namely Chahar-Farsakh, Basiran, Kudegan, Chah-Dashi, Deh-Salm, and Bala-Zard, whose area is 2,500 Km². The geochemical exploration studies at the 1:100000 scale running throughout Iran and including 659 sheets are part of national projects conducted by the Geological Survey and Mineral Explorations of Iran (GSMEI). 4138 stream sediment samples were collected by GSMEI from the studied area, the sample number of each sheet is shown in Figure 3. The section of the samples smaller than mesh size 80 was selected for chemical analysis by ICP-OES. The samples were analyzed for 44 to 52 elements in the laboratory of GSMEI. Also 147 duplicate samples were collected in order to verify the sampling and analysis methods. The relative standard deviation (RSD) calculated by Thompson and Howarth's (1976) method [63] indicates that the value of this

parameter is less than 5% for the major elements and less than 10% for the others.

32° 00'	Kudegan N=843	Basiran N=585	Chahar-Farsakh N=586
31° 30'	Bala-Zard N=855	Deh-Salm N=635	Chah-Dashi N=634
31° 00'	58° 30'	59° 00'	59° 30' 60° 00'

Figure 3. Layout of the 1:100,000 scale sheets in the studied area with the numbers of their stream sediment geochemical samples.

The location and distribution of the samples of the studied area are shown in Figure 4. With respect to the aim of this work, which is to determine the metal mineralization area, 18 elements were selected in the form of two chalcophile and

siderophile element groups. Table 1 shows the descriptive statistical parameters of these selected elements. The concentration of the stream sediment samples is reduced due to their dilution. However, when the means of the element concentrations are compared with their Clarke numbers, the rich metal element studied area is obtained (except for the Cu, Hg, Co, and Ni elements, whose means are slightly smaller than their Clarke numbers). The skewness and kurtosis values also represent a non-normal distribution of the elements and multi-populations of the dataset. Therefore, the statistical parameters indicate the existence of anomalies for these elements in the studied area. In order to determine the extent of the composite geochemical anomalies by the clustering algorithms, the following three

steps were used as the per-processing techniques of the dataset.

- Firstly, the censored data is replaced by 3/4 detection limits of the analytical methods.
- Secondly, centered logratio (clr) transformation, as suggested by Templ *et al.*, (2008) and Zhou *et al.*, (2017) for clustering the geochemical data [18, 64], was used to convert the dataset from a closed-number system to an open-number system.
- Thirdly, the clr-transformed dataset was standardized to interval of zero to one range in order to eliminate the effect of the measurement unit.

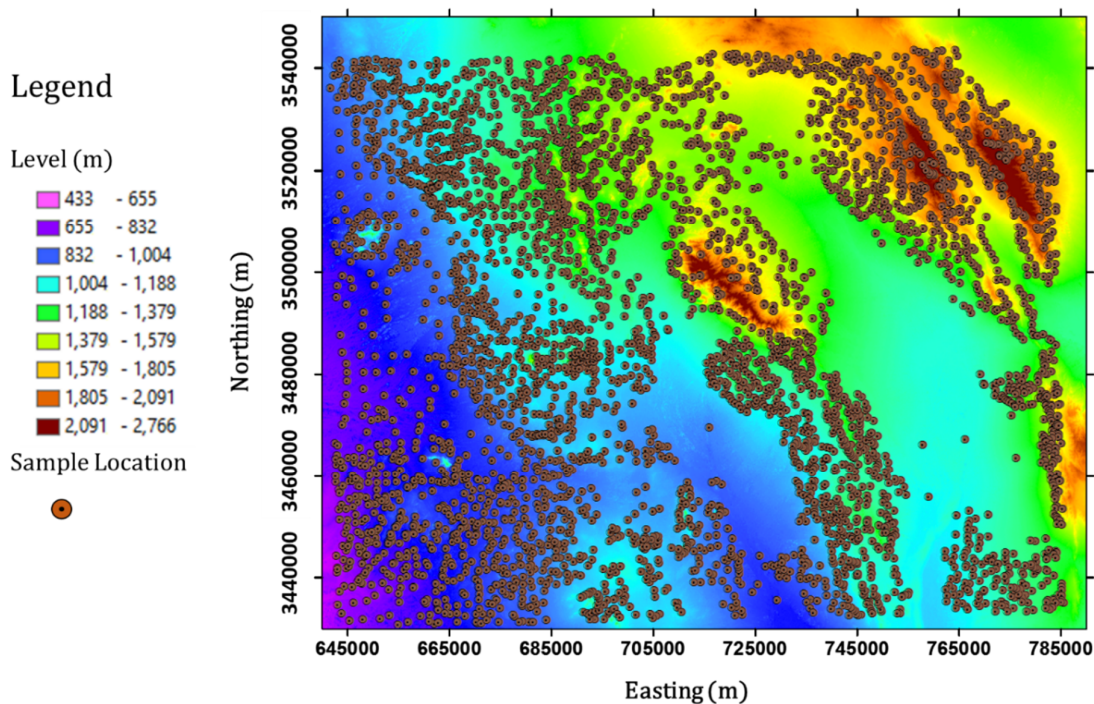


Figure 4. Location map of the geochemical samples on the topography map of the studied area.

6. Results and Discussion








The source of the calc-alkaline magma in the subduction zones can be attributed to the formed hydrated silicates derived from the metamorphism and serpentinization of the basaltic oceanic crust in the beginning of the process. With increases in the pressure and water output from magma, the melting point of rocks located in the upper part of the oceanic crust is reduced, and rhyolite and rhyodacite magmas are formed. This compound, in the vicinity of mantle rocks, produces a low-density garnet-pyroxenite composition rising like a

diapir to produce the calc-alkaline magma [55]. The magmatic activities in the studied area lead to a relatively full range of igneous rocks from acidic to alkaline and from intrusive to extrusive (Figure 1-B). A terrace and alluvial sediment unit covers most of the studied area, followed by an intermediate igneous rock unit (with an area of 2354 Km², or representing 14.9% of the studied area) and then the sedimentary and acidic igneous rock units. The alkaline igneous rock unit (with an area of 239 Km² or 5.1%) is the least frequent in the studied area (Table 2).

Table 1. Descriptive statistical parameters of the chalcophile and siderophile element.

	Variable (Unit)	Mean	StDev	Minimum	Median	Maximum	MAD	Skewness	Kurtosis
Chalcophile	Ag (ppm)	0.3747	0.4546	0.0110	0.2250	13.3813	0.17	10.77	275.66
	As (ppm)	13.067	11.743	0.302	12.000	518.000	1.490	30.55	1184.02
	Bi (ppm)	0.4755	2.6546	0.0940	0.2010	57.1288	0.039	14.01	224.55
	Cd (ppm)	0.33680	0.37658	0.01036	0.22500	5.60928	0.04	5.95	53.77
	Cu (ppm)	37.79	41.55	0.150	36.63	1738.43	9.25	33.98	1345.91
	Hg (ppm)	0.0578	1.791	0.0066	0.0220	102.639	0.007	57.28	3281.97
	Pb (ppm)	23.054	16.684	0.150	23.791	520.437	4.192	19.06	489.45
	Sb (ppm)	1.0580	0.8858	0.3000	0.9500	20.0140	0.25	9.49	121.98
	Sn (ppm)	4.6301	3.1574	0.1500	5.1590	28.1595	2.159	1.13	3.04
Zn (ppm)	92.364	33.615	0.150	84.208	589.044	14.792	3.05	27.04	
Siderophile	Au (ppb)	1.832	14.897	0.300	1.000	703.255	0.3	41.70	1841.59
	Co (ppm)	20.906	6.655	0.150	20.600	74.200	4.2	1.03	3.74
	Cr (ppm)	171.29	72.53	1.50	155.88	1504.80	31.87	4.06	42.33
	Fe (%)	6.7312	3.5275	0.0075	6.5110	26.6667	2.971	0.86	1.04
	Mn (ppm)	1112.1	450.3	1.50	1116.3	6183.7	395.6	1.09	5.81
	Mo (ppm)	1.01	1.52	0.05	0.86	79.00	0.25	41.89	2111.23
	Ni (ppm)	60.771	31.820	1.500	51.612	318.900	13.298	3.09	17.35
	W (ppm)	18.877	44.012	0.350	1.100	527.000	0.39	3.45	16.13

Table 2. Areas and covering percentage of the rock units in the studied area.

Rock Type							
Are (km ²)	8752	741	239	2354	1346	2101	273
Percent	55.4	4.7	1.5	14.9	8.5	13.3	1.7

Due to the abundance of the metal mines and mineralized areas in the studied area (Figure 2), the aim of this work is to identify the metal geochemical anomalies. Since based on the Goldschmidt rules, the geochemical characteristics of the elements are different, the selected metal elements have been divided into two categories: chalcophile and siderophile (Table 1). While the chalcophile elements are heavy metal and non-metal elements that have an affinity for covalent bonding with sulfur to form sulfide minerals, the siderophile elements are high-density transition metal elements with little reactivity, and have an affinity for creating metallic bonds and forming pure metal or oxide minerals [56]. Thus the composite geochemical anomalies of each one of these two groups are drawn individually by the four clustering algorithms.

The frequency of the main elements in the calc-alkaline magma depends on subtraction of the magma silicates. Increases in silica often results in a reduction of the concentration of these elements [55, 57]. The siderophile elements can be replaced by Mg and Fe in the crystal lattice of olivine and clinopyroxene minerals and aggregate in the rocks when the calc-alkaline magma subtracts. Meanwhile, when the amount of silica and the ratio of the FeO/MgO concentration in the magma are reduced, the concentration of the chalcophile elements can be increased [57]. The formation of

multi-element geochemical anomalies in the rocks of the studied area depends not only on the magma type but also on two other factors, namely the alteration and weathering processes. As the samples used in this work belong to the stream sediment geochemical samples, the weathering process contributes a lot to the formation of the anomalies. The determination and positioning of these anomalies is investigated below.

All of the clustering algorithms result in three clusters whose numbers are determinate by the location of their central points. The first cluster is the data population of the background, while the second cluster is the data population of the possible anomaly and the third cluster is the data population of the probable anomaly (theses anomalies have been defined by Hawkes and Webb, 1962; [58]). The dataset has been divided into two categories in order to determine the composite geochemical anomalies. The first category is related to the chalcophile elements, while the second one is associated with the siderophile elements. The dataset of the first category is 10-dimensional, while that of the second one is 8-dimensional.

Figure 5-A shows the results of the geochemical data clustering by the PHC algorithm for the chalcophile elements. From 4138 samples, 1912, 1568, and 658 are clustered into the background, possible anomaly, and probable anomaly populations, respectively, resulting in 5111 and

729 Km² of the possible and probable anomalies. Figure 6-A also shows the results of the clustering performed by this algorithm for the siderophile elements. The areas of the possible and probable anomalies were obtained 5039 and 463 Km², respectively. In this case, 2082, 1493, and 563 samples are in the background, possible anomaly, and probable anomaly populations, respectively. In order to calculate the percentage overlap of the

composite geochemical anomalies (in Figures 5 and 6) with lithology units (in Figure 1-B), the image processing technique is used. The results of this process are presented in Table 3. The highest percentage overlap obtained by this clustering algorithm belongs to the terraces and alluvial sediment unit, while the lowest one belongs to the sedimentary rock unit.

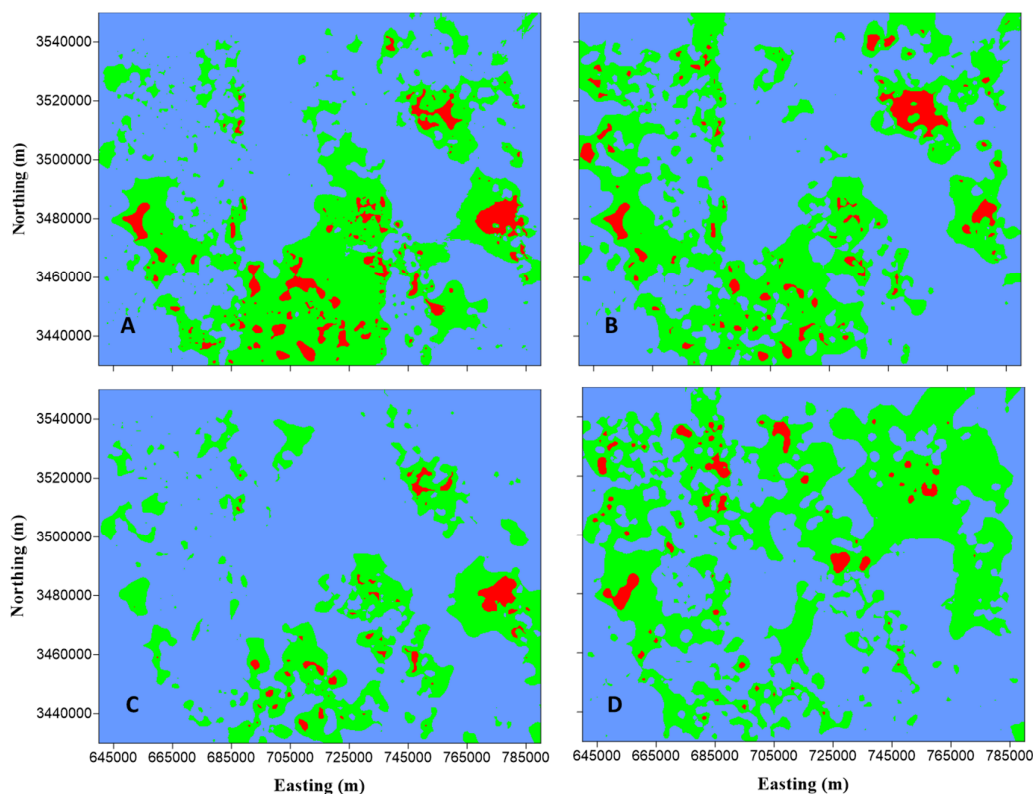


Figure 5. Predicting areas of background (blue), possibility anomaly (green), and probability anomaly (red) populations for the chalcophile element obtained by the PHC (A), PCMC (B), PEMC (C) and PDBSCAN (D) algorithms in the studied area.

In order to calculate the optimal membership degree in the PCMC algorithm, the value of ε is regarded as 10^{-3} . Thus the membership degree of each sample is calculated to an accuracy of three decimal places. When this algorithm is used to cluster the chalcophile elements data from 4138 samples, 2087, 1250, and 801 samples are placed in the background, possible anomaly, and probable anomaly populations, respectively. The locations and extents of the obtained populations are shown in Figure 5-B. The areas of the possible and probable anomalies have been estimated as 5197 Km² and 748 Km², respectively. The possible and probable anomalies of the chalcophile elements are more abundant in the terrace and alluvial sediment unit followed by the intermediate and acidic

igneous rock units. The percentage overlap of the composite anomalies with rock units in this studied area is shown in Table 3. When clustering the siderophile element data using the PCMC algorithm, 1229, 1360, and 549 samples are clustered into the background, possible anomaly, and probable anomaly populations, respectively. The shapes and locations of these populations are also shown in Figure 6-B. The possible anomaly area of the siderophile elements is 4928 Km² and mostly overlap with the terrace and alluvial sediment and intermediate igneous rock units, while the probable anomaly area of the siderophile elements is 761 Km² and is more abundant in the terrace and alluvial sediment and sedimentary rock units (Table 3).

In the PEMC algorithm, the nearest mean classifier is used for the initial selection of the probability parameters of each cluster. Then these model parameters are optimized in the next iteration steps. Finally, a value of ϵ smaller than 10^{-3} is used as the stopping condition of the iteration steps. Out of the 4138 geochemical samples clustered by this algorithm, 2730 chalcophile samples are placed in the background population, and 985 and 423 chalcophile samples are put in the possible anomaly and probable anomaly

populations, respectively. However, regarding the siderophile elements, the numbers of samples are 2330, 1297, and 511, respectively. Figures 5-C and 6-C show the obtained composite chalcophile and siderophile element anomalies, respectively. The areas of these anomalies and their percentage overlap with the rock units of the studied area are also shown in Table 3. In this clustering algorithm, the anomalies mostly overlap with the terrace and alluvial sediment unit followed by intermediate and acidic igneous rock units.

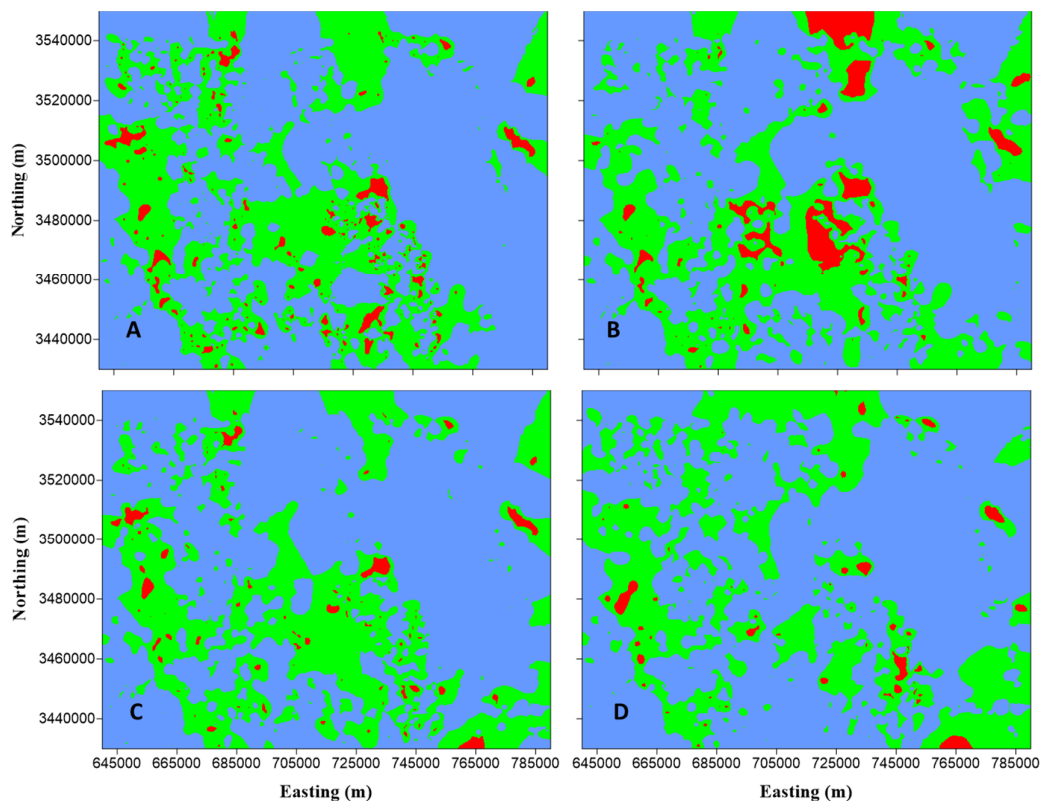


Figure 6. Predicting areas of background (blue), possibility anomaly (green), and probability anomaly (red) populations for the siderophile element obtained by the PHC (A), PCMC (B), PEMC (C) and PDBSCAN (D) algorithms in the studied area.

The trial-and-error method is used to estimate the neighborhood distance, the neighborhood density, and the probability thresholds in the PDBSCAN algorithm. The values of these parameters are estimated as $\epsilon = 0.23$, $MinPts = 6$, and $p = 0.5$. After clustering the dataset, the numbers of samples of the background, possible anomaly, and probable anomaly populations have been calculated as 1652, 2333, and 153 for the chalcophile element data, and 2180, 1834, and 124 for the siderophile element data. The locations of the composite anomalies are shown for the chalcophile elements in Figure 5-D and for the siderophile elements in Figure 6-D. The areas of these anomalies are also listed in Table 3.

The obtained possible and probable anomalies mostly overlap with the terrace and alluvial sediment unit followed by the intermediate and acidic igneous rock units (Table 3).

Figures 5 and 6 show that the composite anomalies obtained by both PHC and PCMC are larger than those obtained by the other two algorithms. Although the shapes of the composite geochemical anomalies are different in the four clustering algorithms, their extents and their locations are almost identical. This can represent the validity of the probabilistic clustering methods to determine the geochemical anomalies.

Figure 7 shows uni-element geochemical anomalies obtained by the concentration-area fractal method. The uni-element geochemical anomalies of the chalcophile elements, in Figure 7-A, are concentrated in two areas, one in the northwestern part and the other in the southern part of the studied area. Moreover, comparison of Figure 7-A with Figure 5 shows that the uni-element geochemical anomalies of the southern part are more similar to the map of Figure 5-A, while the anomalies of the northeastern part are consistent with the map of Figure 5-C. Therefore, it can be said that the PHC and PDBSCAN algorithms and in the next stage the PCMC algorithm have been able to determine the composite geochemical anomalies of the chalcophile elements well. In contrast, the uni-

element geochemical anomalies of the siderophile elements, in Figure 7-B, are scattered throughout the studied area. The results obtained from the PHC algorithm and then the PCMC algorithm have the strongest similarity to Figure 7-B. Therefore, these algorithms have been able to determine the composite geochemical anomalies of the siderophile elements better than the two another algorithm (i.e. PEMC and PDBSCAN algorithms). However, as a general result of comparing between the uni-element geochemical anomalies (Figure 7) with the composite geochemical anomalies (Figures 5 and 6), we can express the acceptable performance of the probabilistic clustering methods in identifying the mineralization areas in the studied area.

Table 3. Areas of the composite geochemical anomalies with their overlapping percentage with the rock units of the studied area.

Clustering algorithm	Type of elements	Type of anomaly	Area (Km ²)	Percent of overlapping with different rock units						
				1	2	3	4	5	6	7
PHC	Chalcophile	Possibility	5111	52.7	4.3	1.1	14.6	11.8	12.4	3.1
		Probability	729	48.1	2.7	1.2	18.1	18.8	7.7	3.4
	Siderophile	Possibility	5039	59.5	5.5	2.3	14.4	7.0	8.9	2.4
		Probability	463	54.1	4.2	5.4	14.7	9.3	7.1	5.2
PCMC	Chalcophile	Possibility	5197	55.8	5.5	1.4	16.4	7.9	10.8	2.2
		Probability	748	43.6	3.6	3.6	12.8	25.5	7.1	3.8
	Siderophile	Possibility	4928	57.9	6.1	2	15.2	5.8	11.1	1.9
		Probability	761	52.8	3.6	5.7	10.8	7.2	14.1	5.8
PEMC	Chalcophile	Possibility	3026	52.7	2.9	1.2	15.2	14.3	10.1	3.6
		Probability	324	49.2	0.6	0.3	12.5	27.6	6	3.8
	Siderophile	Possibility	4824	59.3	5.7	2	15.2	6.1	9.6	2.1
		Probability	369	47.7	4.3	6.3	14.5	9.8	9.9	7.5
PDBSCAN	Chalcophile	Possibility	6323	55.9	4.5	2	11.3	11	13.3	2
		Probability	385	49.9	3.7	0.7	17.1	14.3	8.1	6.2
	Siderophile	Possibility	4670	61.9	4.9	2.4	12.8	6.1	9.6	2.3
		Probability	236	43.2	1.6	8.9	9	16.3	12.2	8.8

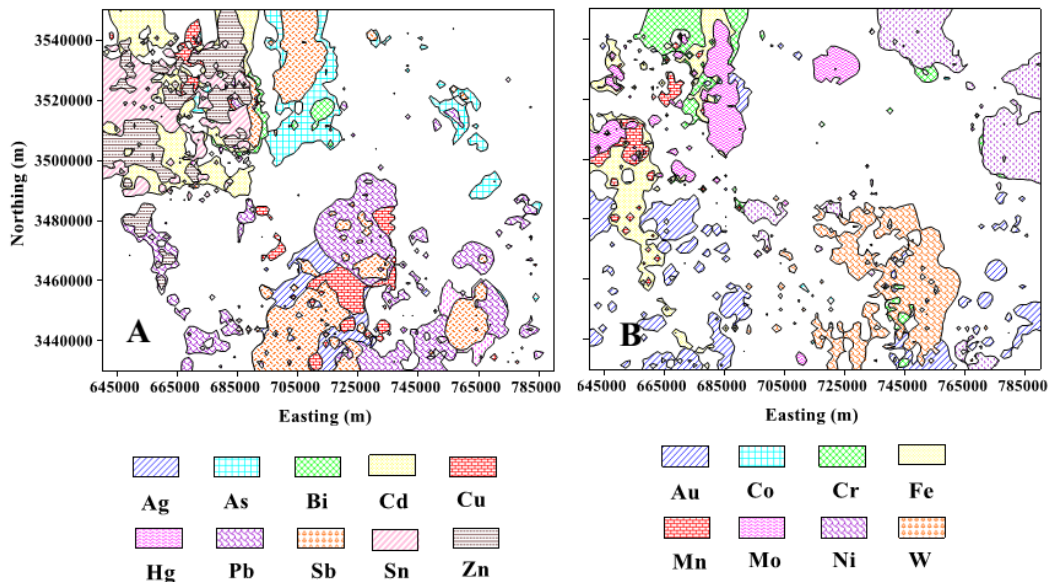


Figure 7. Maps of uni-element geochemical anomalies in the studied area for (A) chalcophile elements and (B) siderophile elements.

Table 3 shows that the composite geochemical anomalies mostly happen in the terrace and alluvial sediment unit, which could be attributed to (1) the types of geochemical samples that are stream sediment samples; and (2) the fact that most of the studied area is covered with this rock unit (Table 2). The smallest percentage overlap of the composite geochemical anomalies with rock units also belongs to the sedimentary rock unit, which can be reasonably justified by (1) the unexpected existence of the metal geochemical anomalies in the sedimentary rock unit; and (2) the fact that this rock unit has the least coverage in the studied area (Table 2). Therefore, Tables 2 and 3 should be taken into consideration simultaneously in order to identify the relationship between the mineralization area and/or the composite geochemical anomalies and that of the rock units in the studied area. If the area or percentage of the composite geochemical anomalies overlap by the rock units (the data of Table 2) is divided into the area or the proportion of the coverage of the rock units (the data of Table 3), it can be concluded that the chalcophile element anomalies are mostly related to the metamorphic rocks followed by the acidic igneous rocks in the studied area. However,

the siderophile element anomalies are mostly related to the alkaline igneous rocks followed by the metamorphic rocks. The relation between the composite geochemical anomalies and the intermediate igneous rocks is in the next rank for comparison.

In order to choose the best clustering algorithms, three methods are used in this work. The first method is based on the number of the mines and mineralized areas that is located in the composite geochemical anomalies obtained by the different clustering algorithms. For this purpose, it is assumed that each index shown in Figure 2 has an area of at least 2 Km². The results of the compliance of Figure 2 with Figures 5 and 6 are shown in Table 4. The anomalies of the chalcophile and siderophile elements obtained by the PDBSCAN and PHC approaches, respectively, cover most of the indices. Most remarkable is the fact that all the mines are placed in the probable anomaly obtained by the PHC and PCMC approaches. The results of Table 4 also show that the differences among the clustering algorithms used are not significant, and on average, more than 65% of the indices are placed in the composite geochemical anomalies.

Table 4. Number of the mines and mineralized areas placed in the statistical populations obtained by four clustering algorithms.

Element	Clustering algorithm	Population		
		Background	Possibility anomaly	Probability anomaly
Chalcophile	PHC	44	84	11
	PCMC	43	85	11
	PEMC	71	65	3
	PDBSCAN	47	88	4
Siderophile	PHC	52	83	4
	PCMC	51	78	10
	PEMC	59	76	4
	PDBSCAN	64	73	2

The second method for selecting the best clustering algorithms is the percentage overlap of the alteration map with the statistical population maps shown in Figures 5 and 6. A hydrothermal alteration map, which is related to the metal mineralization areas, has been provided in Figure 8. The Spectral Angle Mapper (SAM) method has been used to produce this map. In the alteration mapping of the studied area, sericite mineral represents sericitic alteration, kaolinite and montmorillonite minerals represent argillic alteration, chlorite and epidote minerals represent propylitic alteration, quartz mineral represents silicification, alunite, and jarosite minerals

represent alunatic alteration, and finally, hematite and goethite minerals represent iron oxide alteration. The iron oxide alteration map has been prepared from the Landsat images of the studied area, while the other alteration maps used are the Aster images. After the preliminary corrections, a simple false color composite (i.e. RGB) is used for initial processing. The images obtained are processed with the help of the digital spectral library of the minerals (refer to Clark *et al.*, 1993; [59]) and the SAM method (for more details, refer to Kruse *et al.*, 1993; [60]). Finally, Hydrothermal alteration map is obtained by integration of the above-mentioned alteration maps. This map has

also been overlapped with the mines and mineralized areas map (Figure 2) in order to evaluate its validity (Figure 8). As shown in Figure 8, 76 (or 57%) of the 134 mines and mineralized areas are located in the hydrothermal alterations. This implies that hydrothermal alteration map has an acceptable validity.

Table 5 represents the percentage overlap of the alteration map (Figure 8) with the maps obtained from the results of the clustering algorithms (Figures 5 and 6). The results of this table are obtained by the image analysis method. The results show that more than 60% of the alteration areas exist in the areas of the composite geochemical anomalies. Although there is no significant difference among the clustering algorithms, the PDBSCAN and PHC approaches show a better performance for the chalcophile and siderophile elements, respectively. The highest percentage

overlap of the probable anomaly with the alteration map is given by the PHC and PCMC approaches for the chalcophile and siderophile elements, respectively. The conformity of these results with the anomaly areas (Table 3) best justifies this.

The third method of choosing the best clustering algorithm is based on the values of the validity indices. These values are listed separately in Table 6 for the four clustering algorithms for the chalcophile and siderophile elements. The results obtained show that all the four approaches have high validities. Moreover, the results show that PDBSCAN has a higher validity for clustering the chalcophile elements, while PHC has a higher validity for the siderophile elements (the highest values for the MHI and RSI indices and the lowest value for the DBI index) followed by the PCMC and PEMC approaches for the chalcophile and siderophile elements, respectively.

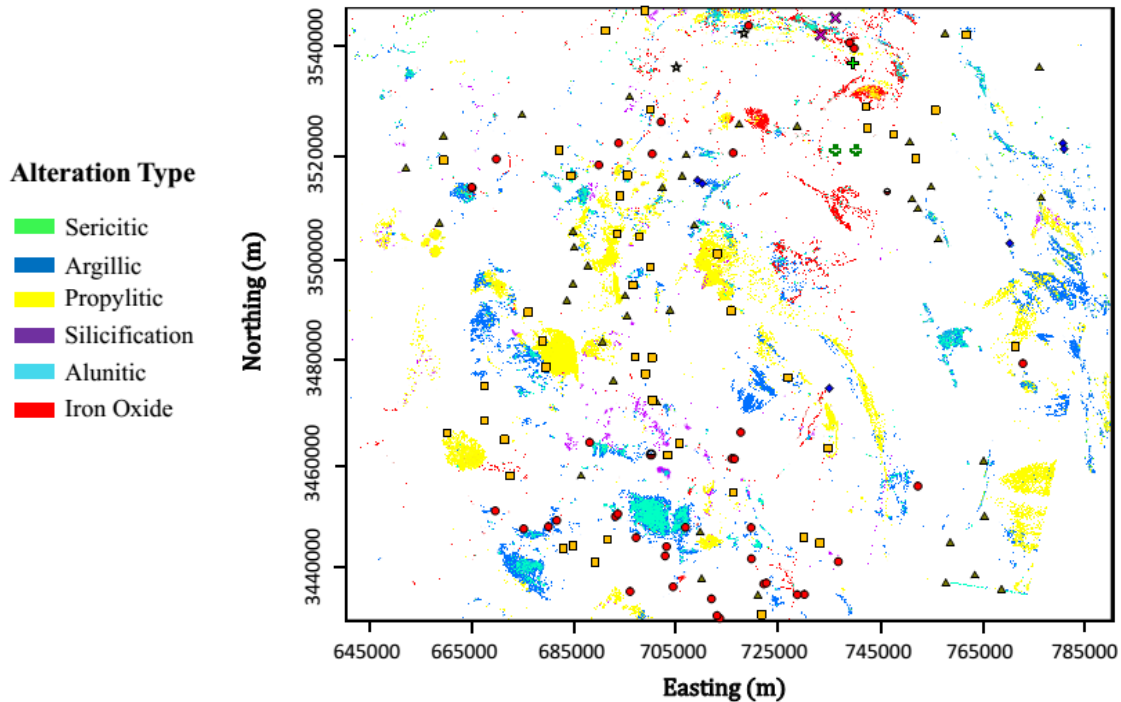


Figure 8. Hydrothermal alteration map with the mines and mineralized areas in the studied area. Legend for mines and mineralized areas are the same as in Figure 2.

Table 5. Overlapping percentage of the alteration areas with the statistical populations.

Element	Clustering algorithm	Population		
		Background	Possibility anomaly	Probability anomaly
Chalcophile	PHC	39.1	56.3	4.6
	PCMC	39.8	56.2	3.9
	PEMC	50.7	47.3	2.0
	PDBSCAN	38.2	59.5	2.3
Siderophile	PHC	35.6	60.6	3.8
	PCMC	37.6	55.5	6.9
	PEMC	45.5	51.1	3.4
	PDBSCAN	47.8	49.9	2.3

Table 6. Values of the cluster validation indices for the chalcophile and siderophile elements.

Element	Clustering algorithm	Relative criteria		
		MHI	DBI	RSI
Chalcophile	PHC	0.0680	6.7798	0.5620
	PCMC	0.0810	5.6629	0.5807
	PEMC	0.0740	6.5370	0.5732
	PDBSCAN	0.1050	3.8233	0.7222
Siderophile	PHC	0.2770	3.0834	0.7802
	PCMC	0.2450	4.5653	0.6510
	PEMC	0.2760	3.4016	0.6866
	PDBSCAN	0.2520	2.8878	0.6618

7. Conclusions

On the regional scale, using the composite geochemical anomalies for the reconnaissance exploration phase is superior to using the uni-element geochemical anomalies because they are able to represent the multi-element mineralization areas simultaneously. In order to determine these anomalies, special statistical methods should be used that are capable of determining the threshold of several elements simultaneously. The clustering methods are the most important statistical tools in this regard. The probabilistic clustering methods are grouped as the model-based clustering techniques. These methods are superior to the non-probabilistic clustering methods due to the use of better measures, high applicability when some variables are not measured, lack of trapping in local optima, and ultimately, the production of better results in the clustering of the dataset.

The four probabilistic clustering approaches, namely the PHC, PCMC, PEMC, and PDBSCAN algorithms, and the 4138-stream sediment geochemical samples taken from an area of 15,000 Km² show that, as a whole, the possible anomalies have a potential area of 5000 Km² for multi-metal mineralization in the studied area, while the potentiality area is 500 Km² for the probable anomalies. Although it seems that the shapes and areas of the obtained geochemical anomalies are slightly different, their extents and locations are almost the same. This shows the validity of the probabilistic clustering approaches for determining the composite geochemical anomalies. The most important results can be listed as follow:

1. The biggest and smallest areas of the anomalies are obtained by the PHC and PEMC algorithms, respectively. The algorithms have brought about these points, which means the agglomeration of the PHC algorithm and optimization of the parameters of each cluster in each step of the PEMC algorithm.
2. The chalcophile element anomalies are mostly widespread in the metamorphism-acidic-

intermediate rock units of the studied area. However, the siderophile element anomalies are more abundant in the alkaline-metamorphism-intermediate rock units.

3. The obtained geochemical anomalies could cover about 65% of the mineralized areas and almost all of the mines of the studied area. Some algorithms such as PHC and PDBSCAN with larger anomaly areas have a superior performance in this respect.
4. The alteration map of the studied area shows almost 60% conformity with the anomaly maps. As mentioned earlier, the PDBSCAN and PHC algorithms have a better performance for the chalcophile and siderophile elements, respectively.
5. The validation indices of the clustering methods also indicate more than 70% validity for the separation of the statistical populations or the anomalies. Moreover, the approaches with bigger anomalies (i.e. the PHC and PDBSCAN algorithms) have higher validity indices.

As a final result of this work, it can be stated that the probabilistic clustering methods can be used as direct methods for identifying the mineralization areas, especially in the regional-scale of geochemical exploration.

Acknowledgments

Special thanks go to the Geological Survey and Mineral Explorations of Iran for providing us with the geochemical data and geology map of the studied area. We also highly appreciate the Department of Industry and Trade and Mines of South Khorasan great contribution in providing the researchers with the exploration areas and mines of the studied area. Besides, the researchers would like to express their gratitude to Prof. Jiancong Fan for kindly sending us the source codes of the OPE-HCA program.

References

- [1] Haldar, S.K. (2013). Mineral Exploration: Principles and Applications, Elsevier, 372 p.
- [2] Galuszka, A. (2007). A review of geochemical background concepts and an example using data from Poland. *Environmental Geology* 52(5): 861-870.
- [3] Wellmer, F.W. (1998). Statistical Evaluations in Exploration for Mineral Deposits, Springer-Verlag Berlin Heidelberg, 379 p.
- [4] Chork, C.Y. (1990). Unmasking multivariate anomalous observations in exploration geochemical data from sheeted-vein tin mineralization near Emmaville, N.S.W., *Journal of Geochemical Exploration* 37 (2): 205-223.
- [5] Geranian, H., Mokhtari, A.R. and Cohen, D.R. (2013). A comparison of fractal methods and probability plots in identifying and mapping soil metal contamination near an active mining area. *Iran, Science of the Total Environment* 463-464: 845-854.
- [6] Wang, J. and Zuo, R. (2016). An extended local gap statistic for identifying geochemical anomalies, *Journal of Geochemical Exploration* 164: 86-93.
- [7] Ghavami-Riabria, R., Seyedrahimi-Niaraqa, M.M., Khalokakaiea, R. and Hazarehb, M.R. (2010). U-spatial statistic data modeled on a probability diagram for investigation of mineralization phases and exploration of shear zone gold deposits. *Journal of Geochemical Exploration* 104 (1-2): 27-33.
- [8] Cheng, Q., Xu, Y. and Grunsky, E. (2000). Integrated Spatial and Spectrum Method for Geochemical Anomaly Separation. *Natural Resources Research* 9: 43-52.
- [9] Cheng, Q., Agterberg, F.P. and Bonham-Carter, G.F. (1996). A spatial analysis method for geochemical anomaly separation. *Journal of Geochemical Exploration* 56 (3): 183-195.
- [10] Daya, A.A. (2015). Comparative study of C-A, C-P, and N-S fractal methods for separating geochemical anomalies from background: A case study of Kamoshgaran region, northwest of Iran. *Journal of Geochemical Exploration* 150: 52-63.
- [11] Jimenez-Espinosa, R., Sousa, A.J. and Chica-Olmo, M. (1993). Identification of geochemical anomalies using principal component analysis and factorial kriging analysis. *Journal of Geochemical Exploration* 46: 245-256.
- [12] Cao, M., and Lu, L. (2015). Application of the multivariate canonical trend surface method to the identification of geochemical combination anomalies. *Journal of Geochemical Exploration* 153 (1): 1-10.
- [13] Meng, H.D., Song, Y.C., Son, F.Y. and Shen, H.T. (2011). Research and application of cluster and association analysis in geochemical data processing. *Computational Geosciences* 15: 87-98.
- [14] Zaremotlagh, S., Hezarkhani, A. and Sadeghi, M. (2016). Detecting homogenous clusters using whole-rock chemical compositions and REE patterns: A graph-based geochemical approach. *Journal of Geochemical Exploration* 170: 94-106.
- [15] Collyer, P.L. and Merriam, D.F. (1973). An application of cluster analysis in mineral exploration. *Mathematical Geosciences* 5 (3): 213-223.
- [16] Roy, A. (1981). Application of cluster analysis in the interpretation of geochemical data from the Sargipalli lead-zinc mine area, Sundergarh district, Orissa (India). *Journal of Geochemical Exploration* 14: 245-264.
- [17] Ellefsen, K.J. and Smith, D.B. (2016). Manual hierarchical clustering of regional geochemical data using a Bayesian finite mixture model. *Applied Geochemistry* 75: 200-210.
- [18] Morrison, J.M., Goldhaber, M.B., Ellefsen, K.J. and Mills, C.T. (2011). Cluster analysis of a regional-scale soil geochemical dataset in northern California. *Applied Geochemistry* 26: S105-S107.
- [19] Fatehi, M. and Asadi, H.H. (2017). Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data, a case study from the Dalli Cu-Au porphyry deposit in central Iran. *Ore Geology Reviews* 81: 245-255.
- [20] Ellefsen, K.J., Smith, D.B. and Horton, J.D. (2014). A modified procedure for mixture-model clustering of regional geochemical data. *Applied Geochemistry* 51: 315-326.
- [21] Aggarwal, C.C. and Reddy, C.K. (2013). *Data Clustering: Algorithms and Applications*. CRC Press, 652 p.
- [22] Han, J., Kamber, M. and Pei, J. (2011). *Data mining: concepts and techniques*, 3rd Edition. Morgan Kaufmann, 744 p.
- [23] Brauer, S. (2014). *A Probabilistic Expectation Maximization Algorithm for Multivariate Laplacian Mixtures*. MS Thesis of Paderborn University, 78 p.
- [24] Fan, J. (2019). OPE-HCA: an optimal probabilistic estimation approach for hierarchical clustering algorithm. *Neural Computing and Applications* 31: 2095-2105.
- [25] Krishnapuram, R. and Keller, J.M. (1993). A Possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems* 1 (2): 98-110.
- [26] Xie, Z., Wang, S. and Chung, F.L. (2008). An enhanced possibilistic C-Means clustering algorithm EPCM. *Soft Computing* 12: 593-611.
- [27] Salgado, P. and Igrejas, G. (2007). Probabilistic Clustering Algorithms for Fuzzy Rules Decomposition. *IFAC Proceedings Volumes* 40 (21): 115-120.

- [28] Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* 2: 73–82.
- [29] Quost, B. and Dencoux, T. (2016). Clustering and classification of fuzzy data using the fuzzy EM algorithm. *Fuzzy Sets and System* 286 (1): 134-156.
- [30] González, M., Minuesa, C. and Puerto, I. (2016). Maximum likelihood estimation and expectation–maximization algorithm for controlled branching processes. *Computational Statistics & Data Analysis* 93: 209-227.
- [31] Hu, T. and Sung, S.Y. (2006). A hybrid EM approach to spatial clustering. *Computational Statistics & Data Analysis* 50: 1188–1205.
- [32] Kriegel, H.P. and Pfeifle, M. (2005). Density-based clustering of uncertain data. In *Proc. of KDD2005*, New York, NY, USA, 672–677.
- [33] Xu, H. and Li, G. (2008). Density-Based Probabilistic Clustering of Uncertain Data. *International Conference on Computer Science and Software Engineering (CSSE 2008)*, Wuhan, China, 474-477.
- [34] Zhang, X., Liu, H., Zhang, X. and Liu, X. (2014). Novel Density-Based Clustering Algorithms for Uncertain Data. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec, Canada, 2191- 2197.
- [35] Beckmann, N., Kriegel, H.P., Schneider, R. and Seeger, B. (1990). The R*-tree: an efficient and robust access method for points and rectangles. *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, 322-331.
- [36] Erdem, A. and Gündem, T.I. (2014). M-FDBSCAN: A multicore density-based uncertain data clustering algorithm. *Turkish Journal of Electrical Engineering & Computer Sciences* 22: 143 – 154.
- [37] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002). Clustering validity methods: Part I. *ACM SIGMOD Record* 31(2): 40-45.
- [38] Rendón, E., Abundez, I., Arizmendi, A. and Quiroz, E.M. (2011). Internal versus External cluster validation indexes. *International Journal of Computers and Communications* 5 (1): 27-34.
- [39] Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002). Clustering validity checking methods: Part II. *ACM SIGMOD Record* 31 (3).
- [40] Gurrutxaga, I., Albisua, I., Arbelaitz, O., Martín, J.I., Mugerza, J., Pérez, J.M. and Perona, I. (2010). SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition* 43: 3364–3373.
- [41] Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *IEEE International Conference on Data Mining*, 911-916.
- [42] Bröcker, M., Fotoohi Rad, G., Abbaslu, F. and Rodionov, N. (2014). Geochronology of high-grade metamorphic rocks from the Anjul area Lut block, eastern Iran. *Journal of Asian Earth Sciences* 82: 151–162.
- [43] Mirnejad, H., Blourian, G.H., Kheirkhah, M., Akrami, M.A. and Tutti, F. (2008). Garnet bearing rhyolite from Deh-Salm area, Lut block, Eastern Iran: anatexis of deep crustal rocks. *Mineral. Petrol.* 94: 259–269.
- [44] Asadi, S. and Kolahdani, S. (2014). Tectonomagmatic evolution of the Lut block, eastern Iran: A model for spatial localization of porphyry Cu mineralization. *Journal of Novel Applied Sciences* 3: 1058-1069.
- [45] Mazhari, S.A. and Safari, M. (2013). High-K Calc-alkaline Plutonism in Zouzan, NE of Lut Block, Eastern Iran: An Evidence for Arc Related Magmatism in Cenozoic. *Journal Geological Society of India* 81: 698-708.
- [46] Pang, K.N., Chung, S.L., Zarrinkoub, M.H., Mohammadi, S.S., Yang, H.M., Chu, C. H., Lee, H.Y. and Lo, C.H. (2012). Age, geochemical characteristics and petrogenesis of Late Cenozoic intraplate alkali basalts in the Lut-Sistan region, eastern Iran. *Chemical Geology* 306–307: 40–53.
- [47] Mahmoudi, S., Masoudi, F., Corfu, F. and Mehrabi, B. (2010). Magmatic and metamorphic history of the Deh-Salm metamorphic Complex, Eastern Lut block, (Eastern Iran), from U–Pb geochronology. *Int. J. Earth Sci.* 99: 1153–1165.
- [48] Malekzadeh Shafaroudi, A. and Karimpour, M.H. (2015). Mineralogic, fluid inclusion, and sulfur isotope evidence for the genesis of Sechangi lead–zinc (–copper) deposit, Eastern Iran. *Journal of African Earth Sciences* 107: 1–14.
- [49] Arjmandzadeh, R., Karimpour, M.H., Mazaheri, S.A., Santos, J.F., Medina, J.M. and Homan, S.M. (2011). Two-sided asymmetric subduction; implications for ectonomagmatic and metallogenic evolution of the Lut Block, eastern Iran. *Journal of Economic Geology* 3 (1): 1-14.
- [50] Wilmsen, M., Fürsich, F.T. and Majidifard, M.R. (2013). The Shah Kuh Formation, a latest Barremian e Early Aptian carbonate platform of Central Iran (Khur area, Yazd Block). *Cretaceous Research* 39: 183-194.
- [51] Arjmandzadeh, R. and Santos, J.F. (2014). Sr-Nd isotope geochemistry and tectonomagmatic setting of the Dehsalm Cu-Mo porphyry mineralizing intrusive from Lut Block, eastern Iran. *Int J Earth Sci (Geo Rundsch)* 103: 123-140.
- [52] Arjmandzadeh, R., Karimpour, M.H., Mazaheri, S.A., Santos, J.F., Medina, J.M. and Homam, S.M.

- (2011b). Sr-Nd isotope geochemistry and petrogenesis of Chah-Shaljami granitoids (Lut Block, Eastern Iran). *Journal of Asian Earth Science* 41: 283-296.
- [53] Eshraghi, H., Rastad, E. and Motevali, K. (2010). Auriferous sulfides from Hired gold mineralization, South Birjand, Lut Block, Iran. *J Miner Petrol Sci* 105: 167-174.
- [54] Ghorban, M. (2013). *The economic geology of Iran: Mineral Deposits and Natural Resources*, Springer Publication, Netherlands, 569p.
- [55] Pirajno, F. (2009). *Hydrothermal Processes and Mineral Systems*, Springer Publication, Australia, 1273 p.
- [56] White, W.M. (2013). *Geochemistry*, Wiley-Blackwell Publications, 668 p.
- [57] Santoa, A.P., Jacobsenb, S.B. and Baker, J. (2004). Evolution and genesis of calc-alkaline magmas at Filicudi Volcano, Aeolian Arc (Southern Tyrrhenian Sea, Italy). *Lithos* 72: 73–96.
- [58] Hawkes, H.E. and Webb, J.S. (1962). *Geochemistry in Mineral Exploration*. New York: Harper & Row, 415p.
- [59] Clark, R.N., Swayze, G.A., Gallagher, A.J., King, T.V.V. and Calvin, W.M. (1993). The U. S. Geological Survey, Digital Spectral Library Version 1: 0.2 to 3.0 μm . U.S. Geological Survey, Open File Report 93-592.
- [60] Kruse, F., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P. and Goetz, A. (1993). The spectralimage processing system (SIPS) - interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*,44: 145-163.
- [61] Nabavi, M.H. (1976). *An introduction to geology of Iran*. Geological Survey of Iran Publication, Tehran, Iran, 110 p. (in Persian).
- [62] Stöcklin, J. (1968). Structural history and tectonics of Iran; a review. *The American Association of Petroleum Geologists, Bulletin* 52 (7): 1229-1258.
- [63] Thompson, M. and Howarth, R.J. (1976). Duplicate analysis in geochemical practice. Part 1: Theoretical approach and estimation of analytical reproducibility. *Analyst* 101: 690–698.
- [64] Zhou, S., Zhou, K., Wang, J., Yang, G. and Wang, S. (2017). Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. *Frontiers of Earth Science* 12 (3): 491–505.

کاربرد روش‌های خوشه‌بندی احتمالی در تعیین مناطق کانی‌سازی در مطالعات اکتشافی با مقیاس ناحیه‌ای

حمید گرانیان^{۱*} و زهرا خواجه‌میری^۲

۱- گروه مهندسی معدن، دانشگاه صنعتی بیرجند، بیرجند، ایران

۲- سازمان صنعت، معدن و تجارت استان خراسان جنوبی، بیرجند، ایران

ارسال ۲۰۲۰/۰۷/۰۶، پذیرش ۲۰۲۰/۱۰/۱۳

* نویسنده مسئول مکاتبات: h.geranian@birjandut.ac.ir

چکیده:

نمونه‌برداری از رسوبات آبراهه‌ای در مقیاس ۱:۱۰۰۰۰۰ و آنالیز چند عنصره این نمونه‌ها، یکی از ابزارهای مهم اکتشافی در فاز شناسایی محسوب می‌شود. هدف از این مطالعات تعیین نواحی امیدبخش معدنی برای فازهای بعدی اکتشاف است. بنابراین روش‌های خوشه‌بندی که قادر به تعیین آنومالی‌های ژئوشیمیایی چند عنصره هستند، بر روش‌های تعیین آنومالی تک عنصره می‌تواند برتری داشته باشد. روش‌های خوشه‌بندی احتمالی بدلیل استفاده از سنجه مناسبتر، امکان کار با مجموعه‌ی دارای داده‌ها گمشده و عدم گیرکردن در بهینه محلی از روش‌های خوشه‌بندی الگوریتمیک عملکرد بهتری دارند. چهار الگوریتم خوشه‌بندی احتمالی PHC، PCM، PEM و PDBSCAN بر روی ۴۱۳۸ نمونه رسوبات آبراهه‌ای بکار رفته تا نمونه‌ها را به سه خوشه‌ی جامعه زمینه، جامعه آنومالی ممکن و جامعه آنومالی احتمالی تفکیک کنند. ۱۰ عنصر فلزی بعنوان عناصر کالکوفیل و ۸ عنصر فلزی بعنوان عناصر سیدروفیل برای تعیین این آنومالی‌ها انتخاب شده‌اند. نتایج نشان دهنده‌ی مناطقی با وسعت تقریبی ۵۰۰ و ۵۰۰۰ کیلومتر مربع به ترتیب بعنوان مناطق آنومالی ممکن و احتمالی است. با وجود شکل‌های متفاوت، موقعیت و محدوده آنومالی‌ها در هر چهار روش خوشه‌بندی یکسان است. آنومالی‌های ژئوشیمیایی عناصر کالکوفیل بیشترین ارتباط را با سنگ‌های دگرگونی-اسیدی-حد واسط و آنومالی عناصر سیدروفیل با سنگ‌های بازی-دگرگونی-حد واسط در منطقه مطالعاتی دارند. به لحاظ همپوشانی آنومالی‌ها با اندیس‌های اکتشافی و معدنی و نقشه آلتراسیون‌های هیدروترمالی روش‌های PHC و PDBSCAN بهتر عمل کرده‌اند. همچنین آنومالی‌های بدست آمده حدود ۶۵ درصد اندیس‌ها اکتشافی، تقریباً کلیه‌ی اندیس‌های معدنی و ۶۰ درصد آلتراسیون‌ها را پوشش داده‌اند. شاخص اعتبارسنجی روش‌های خوشه‌بندی اعتبار بالغ بر ۷۰ درصد را برای آنومالی‌های بدست آمده برآورد کرده‌اند. این نتایج نشان می‌دهند که روش‌های خوشه‌بندی احتمالی می‌توانند بعنوان یک ابزار آماری مناسب در اکتشافات ژئوشیمیایی ناحیه‌ای بکار روند.

کلمات کلیدی: روش‌های خوشه‌بندی احتمالی، آنومالی ژئوشیمیایی مرکب، پتانسیل‌یابی ژئوشیمیایی، نقشه چهار گوشه ده‌سلم.