JME

Journal of Mining & Environment,
Vol.8, No.2, 2017, 179-189.
DOI: 10.22044/jme.2016.656

Local multivariate outliers as geochemical anomaly halos indicators, a case study: Hamich area, Southern Khorasan, Iran

H. Moeini[1] and A. Aryafar[2*]

1. Faculty of Mining and Metallurgy, University of Yazd, Yazd, Iran
2. Department of Mining, Faculty of Engineering, University of Birjand, Birjand, Iran

Received 2 April 2016; received in revised form 13 July 2016; accepted 18 July 2016
**Corresponding author: aaryafar@birjand.ac.ir (A. Aryafar).*

Abstract

Anomaly recognition has always been a prominent subject in preliminary geochemical explorations. Among the regional geochemical data processing, there are a range of statistical and data mining techniques as well as different mapping methods, which serve as presentations of the outputs. The outlier's values are of interest in the investigations where data are gathered under controlled conditions. These values in exploration geochemistry indicate the mineralization occurrences, and therefore, their identification is vital. Both the robust parametric (based on Mahalanobis distance) and non-parametric (based on depth functions) techniques have been developed for a multivariate outlier identification in geochemistry data. In this research work, we applied the local multivariate outlier identification approach to delineate the geochemical anomaly halos in the Hamich region, which is located in the SE of Birjand, South Khorasn province, East of Iran. For this purpose, 396 litho-geochemical samples that had been analyzed for 44 elements were used. The obtained results show a good agreement with the geological and mineral indices of Pb, Zn, and Cu in the southern part of the area. Such studies can be used by a project director to optimize the core drilling places in detailed exploration steps.

Keywords: *Geochemistry Data, Local Multivariate Outlier, Anomaly, Southern Khorasan, Hamich.*

1. Introduction

Anomaly identification has always been an important issue in geochemical regional explorations [1, 2]. Its output is essential in further detailed exploratory operational decisions. In fact, any uncertainty and risk reductions in the next phases depend on the results of the specified regional geochemical targets [3]. Especially for core drilling, one of the most referenced data would be the accurately processed geochemical data. Among the regional geochemical data processing, there are a variety of statistical and data mining approaches as well as different mapping techniques, which serve as presentations of the outputs [4]. They include convenient methods such as statistical distribution thresholds of gaussian distribution tails or extremes [5]. Extreme values are of interest in the investigations where data are gathered under controlled conditions. In contrast, geochemists are typically interested in outliers as indicators of rare geochemical processes. In such cases, these outliers are not part of one, and the same distribution. For example, in exploration geochemistry, samples indicating mineralizations are the outliers sought. In environmental geochemistry, the recognition of contamination is of interest [6]. Outliers are statistically defined as [7, 8] values belonging to a different population because they originate from another process or source, i.e. they are derived from a/some contaminating distribution/s [9]. In exploration geochemistry, the values within the range (mean± 2sdev) are often defined as the "geochemical background", recognizing that the background is a range, and not a single value [10]. The exact value for (mean±2sdev) is still used by some workers as the "threshold", differentiating background from anomalies (exploration geochemistry) or for defining "action levels" or "clean-up goals" (environmental geochemistry) [11]. In geochemistry, traditionally, low values, or lower outliers, have not been seen as important as high values; this is incorrect because low values can be important. In exploration geochemistry, they may indicate the alteration zones (depletion of certain elements) related to nearby mineral accumulations (occurrences). Thus starting geochemical data analysis with statistical tests based on assumptions of normality, independence, and identical distribution may not be warranted [11]. Thus methods that do not strongly build on statistical assumptions had been the first choice, i.e. robust methods like median and median absolute deviation (MAD) or using boxplot both proposed by Tukey (1977) or thresholds based on multi-fractal nature of geochemical data [12, 13]. It should be noted that in the literature on robust statistics, many other approaches have been proposed for outlier detection [14-17, 8]. Multivariate outlier detection belongs to the most important tasks for the statistical analysis of multivariate data. Multivariate outliers behave differently from the majority of observations that are assumed to follow some underlying models like a multivariate normal distribution [18]. They are divided into two major categories: (a) global outliers that deal with the whole shape of the population derived from the bulk of the data, and (b) local outliers that are a spatial concept of neighborhood around each observation that differs from the rest [19, 20]. The deviations of outlying observations from the majority of data points can also be understood in an exploratory context, e.g. by visualizing a measure describing outlyingness and inspecting possible deviations or gaps in the resulting plot [21]. In this work, we applied a method of local multivariate outlier identification developed by Filzmoser et al. (2013) to delineate the anomaly halos in geochemical exploration. To achieve this goal, the litho-geochemical samples taken from the Hamich area located in the SE of Birjand, Southern Khorasn, east of Iran, were studied. It was a part of a regional geochemical exploration project carried out by the Iranian Industry, Mine, and Trade Organization (IMTO).

2. Methodology

The outlying patterns, as stated, may be divided into two types, global and local outliers. In a general definition, a global outlier is an object that has a significantly large distance to its *k*-th nearest neighbor (usually greater than a global threshold), whereas a local outlier has a distance to its *k*-th neighbor that is large relative to the average distance of its neighbors to their own *k*-th nearest neighbors [9]. Haslett et al. [22] have stated that a global outlier is an observation that might have non-spatial attributes with significantly differing values with respect to the majority of the data points. A local outlier is an observation that might have non-spatial attributes with significantly differing values with respect to its neighbors. Usually, a global outlier is also a local one but not vice versa [23]. Studies on outlier detection can generally be divided into two categories, stemming from: (i) statistics and (ii) data mining. In the statistical approach, most methods assume that the observed data is governed by some statistical processes to which a standard probability distribution (e.g. Binomial, Gaussian, Poisson) with appropriate parameters can be fitted to. An object is identified as an outlier based on how unlikely it could have been generated by that distribution [8]. On the other hand, the data mining techniques attempt to avoid model assumptions, relying on the concepts of distance and density, as stated earlier [23]. For most distance-based methods [24, 25], two parameters, called distance d and data fraction α, are required. Following that, an outlier has at least fraction α of all instances farther than d from it [26]. As both d and α are the parameters defined over the entire data, methods based on distance can only find global outliers [23]. The most commonly used measure of outlyingness is the Mahalanobis distance [27]. This multivariate distance measure assigns each observation a distance to the center, taking account of the multivariate covariance structure. Thus for observations $Z_1, \ldots, Z_n$ in the *p*-dimensional space with center μ and covariance Σ, the Mahalanobis distance is defined as Eq. 1 [9].

$$MD_{\mu,\Sigma}\left(Z_i\right) = \left[\left(Z_i - \mu\right)^t \Sigma^{-1} \left(Z_i - \mu\right)\right]^{1/2} \quad , \quad for\ i = 1, \ldots, n. \tag{1}$$

Practically, for obtaining a reliable distance measure for multivariate data, it is crucial that how center μ and covariance Σ are estimated using the data. Classical estimates (arithmetic mean and sample covariance matrix) can be influenced by outlying observations, and thus robust estimates have to be used instead [28,29]. A frequently used robust estimator of multivariate location and scatter is the minimum covariance determinant (MCD) estimator. MCD looks for a subset of observations with smallest determinant of the sample covariance matrix [9]. MCD estimator is defined as Eq. 2.

$$S_H = \frac{1}{|H|}\sum_{i\in H}\left(x_i - \bar{x}_H\right)\left(x_i - \bar{x}_H\right)^T \qquad (2)$$

for some specific subset H of $\{1,\ldots,n\}$ observations that minimize the determinant. The estimator is robust but not invertible if $|H| < p$ [30]. Rousseeuw and Van Driessen (1999) have introduced a fast algorithm for computing the MCD estimator. As a cut-off value for the robust Mahalanobis distance, $chisq\left(p;0.975\right) = \sqrt{\chi^2_{p;0.975}}$ was suggested, that is the square root of the 97.5% quantile of the non-central chi-square distribution with *p* degrees of freedom. Thus the Mahalanobis distance values larger than this cut-off value are considered as the potential multivariate outliers. The distance MD is limited to identify overall, “global” outliers, but not necessarily outliers in a local neighborhood [9]. Interestingly, spatial or “local” outliers are most often also outlying according to the spatial dependence. Usually, it turns out that spatial data sets contain positive spatial autocorrelation, which means that observations with high (respectively low) values for an attribute are surrounded by neighbors that are also associated with high (respectively low) values. Thus in a positive autocorrelation scheme, observations that differ from their neighbors do not follow the same process of spatial dependence as the main bulk of the data. If global outliers are present in the data set, they are usually also local outliers, and can completely mask other local outliers [9]. Suppose some observations in a dataset with two geographical coordinates in a square and two quantitative attributes as an example. The left plot in Figure 1 shows a 2D data, where the majority of the points come from a bivariate normal distribution. The ellipse corresponds to values of $chisq\left(2;0.975\right) = 2.72$ of the robust Mahalanobis distance based on MCD location and scatter estimates. Hence, all squares and the filled rhomb are outside the ellipse, and thus they are identified as global outliers. Hence, all squares and the filled rhomb are outside the ellipse, and thus they are identified as global outliers. Figure 1 (right) shows the spatial X- and Y-coordinates of the data. For four selected points (shown by the filled symbols), circles are drawn, which correspond to a Euclidean distance of 2 units from the points. All points within this distance are drawn with the corresponding open symbols, and they can be considered as the neighbors to the points in the center of the circles. Since the same symbols were used in the left plot of Figure 1, it is possible to see the relation of the points in the variable space and in the coordinate space. The filled square and all its neighbors (at a distance of 2 units) are multivariate outliers. The filled rhomb is a multivariate outlier too, but not the neighbors. The filled triangle is on the boundary of the cut-off value 2.72, and the neighbors (open triangles) are far away in the variable space. Finally, the filled circle is in the center of the data cloud, but its neighbors are very different in the variable space. The filled triangle and circle should thus be identified as local outliers because their neighboring points are very different. The filled rhomb and the filled square are already identified as global outliers, and their neighbors are different for the rhomb but similar for the square [9].

Many different methods have been proposed to deal with these four types of outliers. Graphics such as the variogram cloud [31] and the Moran scatterplot [32] are interesting tools for detecting local outliers in a univariate framework. Cerioli et al. (1999) have used the forward search approach to identify spatial outliers in the univariate context [33]. However, in a multivariate framework, there are a few research works in the literature. One of the most recent ones is by Filzmoser et al. (2013) that puts forward the use of the variogram cloud in a different way. For a pair $\left(c_i, c_j\right)$ of data locations, $(i, j = 1, \ldots, n, i \neq j)$, let us consider the geographical Euclidean distance as Eq. 3 [9].

$$ED\left(c_i, c_j\right) = \left[\left(c_i - c_j\right)^t \left(c_i - c_j\right)\right]^{1/2} \qquad (3)$$

The variogram cloud consists of plotting for all pairs $\left(c_i, c_j\right)$ and for a single variable Z, the values $1/2\left(Z\left(c_i\right) - Z\left(c_j\right)\right)^2$ versus $ED\left(c_i, c_j\right)$. Thus in this context, Filzmoser generalizes the variogram cloud for multivariate data by replacing the absolute differences with pairwise Mahalanobis distances defined as Eq. 4 [9].

$$MD_\Sigma\left(Z_i, Z_j\right) = \left[\left(Z_i - Z_j\right)^t \Sigma^{-1}\left(Z_i - Z_j\right)\right]^{1/2} \qquad (4)$$

This distance measure between all pairs of observations accounts for the overall covariance structure. The multivariate variogram cloud is a scatter plot of $MD_\Sigma\left(Z\left(c_i\right), Z\left(c_j\right)\right)$ versus the geographical Euclidean distances

$ED\left(c_i,c_j\right), for\ i,j=1,...,n$. Here, it is used as the same index for the observations in the variable space and in the coordinate space, i.e. $Z\left(c_i\right)=z_i, for\ i=1,...,n$ [9]. Since local outliers are supposed to be different from their neighbors, one could define a quantile of the non-central *chi-square* distribution and count the number of neighbors falling into this defined range. In case of independence and normal distribution, we would expect 10% of the values falling inside an ellipse of cut-off value drawn in the multivariate space (Figure 1 (left)). Consequently, the ellipses in the center of the data cloud are smaller than on the boundary, which is according to the non-centrality parameter of the chi-square distribution. The assumption of independence will not be valid, in particular, for spatially dependent data. Here, Σ will be estimated robustly using the MCD estimator [9]. Local outlyingness could also be defined differently by measuring the distance to the next neighbor. Let z_j be the next neighbor of z_i, i.e. the distance of z_j and z_i is the smallest among all the neighbors of observation z_i. The pairwise squared Mahalanobis distance $MD^2\left(z_i,z_j\right)$ is equal to a certain $\alpha\left(j\right)$-quantile $\chi^2_{p;\alpha(j)}\left(MD^2\left(z_i\right)\right)$ of the chi-square distribution. Since, just by chance, the next neighbor could be close, it can be more sensible to search for α-quantiles such that the corresponding ellipses include a pre-defined percentage, e.g. 10% of the next neighbors. Thus a characterization of local outliers requires a definition of the local neighborhood. For this purpose, two concepts are common, namely to fix a maximum distance d_{max} in the space of the spatial coordinates, and to define the neighbors of an observation z_i as all points $z_j\left(j=1,...,n; j\neq i\right)$, where the distance $d_{i,j}$ between z_i and z_j is not larger than d_{max}. As distance measure $d_{i,j}$, the Euclidean distance can be considered. A second concept is to define neighborhood by the nearest k observations. For finding the k nearest neighbors (kNN) of an observation z_i, we have to consider the sorted distances $d_{i,1}\leq d_{i,2}\leq ...\leq d_{i,k}\leq d_{i,n}$ to all other observations. kNN to z_i are all observations where $d_{i,j}\leq d_{i,k}$, $for\ j=1,...,n,(j\neq i)$ [11]. Let $MD^2\left(z_i,z_j\right)$ denote the sorted squared pairwise Mahalanobis distances of observation z_i to all neighbors z_j with $j\in N_i=\left\{i_1,...,i_{n(i)}\right\}$. The degree of isolation of an observation z_i from a fraction β of its neighbors can be characterized by the $\alpha\left(i\right)$-quantile by equation (5) [9]:

$$\chi^2_{p;\alpha(i)}\left(MD^2\left(z_i\right)\right)=MD^2\left(z_i,z_{(n(i).(1-\beta))}\right)$$
$$, for\ \ i=1,\dots,n \tag{5}$$

where $\alpha\left(i\right)$ measures the local outlyingness of an observation z_i. If, for example, the fraction β is fixed with 5%, then it means that for each observation, we are computing the degree of isolation from $\left(1-\beta\right)=95\%$ of its neighbors using equation 5. For a large number of neighbors, and in case of independence and normal distribution, $\alpha\left(i\right)$ should approximate β. However, if $\alpha\left(i\right)$ is substantially larger than β, observation z_i is considered as a potential local outlier. This characterization of local outliers depends on the size of the neighborhood (d_{max} or k), and on the fraction β [9]. Thus the local outlier detection in this method can be summarized as follows. For each observation $x_i\left(i=1,...,n\right)$:

(i) Compute the pairwise Mahalanobis distances between x_i and its k neighbors x_j using the global structure (Eq. 3) and MCD based on *ilr*-transformed data. The reason for transforming the data is that the compositional data (here, geochemical data) is intrinsically in a simplex space rather than euclidean. Thus due to their compositional nature, problematic interpretations would be expected when they are used untransformed. After isometric logratio transformation, their distance transfers to Euclidean, and all the related statistical relations would be valid and significant. This problem has recently been considered in detail in the literature, and many solutions have been presented [34-36].

(ii) Determine the ellipsoid containing a proportion β of its k neighbors [9].

(iii) If the tolerance level of this ellipsoid is too large, according to the *chisquare* distribution, then the observation is considered as a local outlier [30]. Two improvements on this method can be proposed: (1) Use a local structure estimated separately on each neighborhood instead of the general one. As size k of the neighborhood can be smaller than dimension p , the local structure has to be estimated by a robust and regularized estimator. (2) Instead of testing the local outlyingness of each observation, we suggest to focus only on the observations corresponding to a positively spatially auto-correlated neighborhood. The multivariate autocorrelation of a neighborhood is estimated by means of the determinant of the regularized MCD covariance estimator computed on the neighborhood, and only the neighborhoods yielding the smallest values are selected [30]. This is the parametric technique for the local multivariate outlier identification. However, there is also a non-parametric detection technique for local outliers based on depth functions (not discuss here) [37].

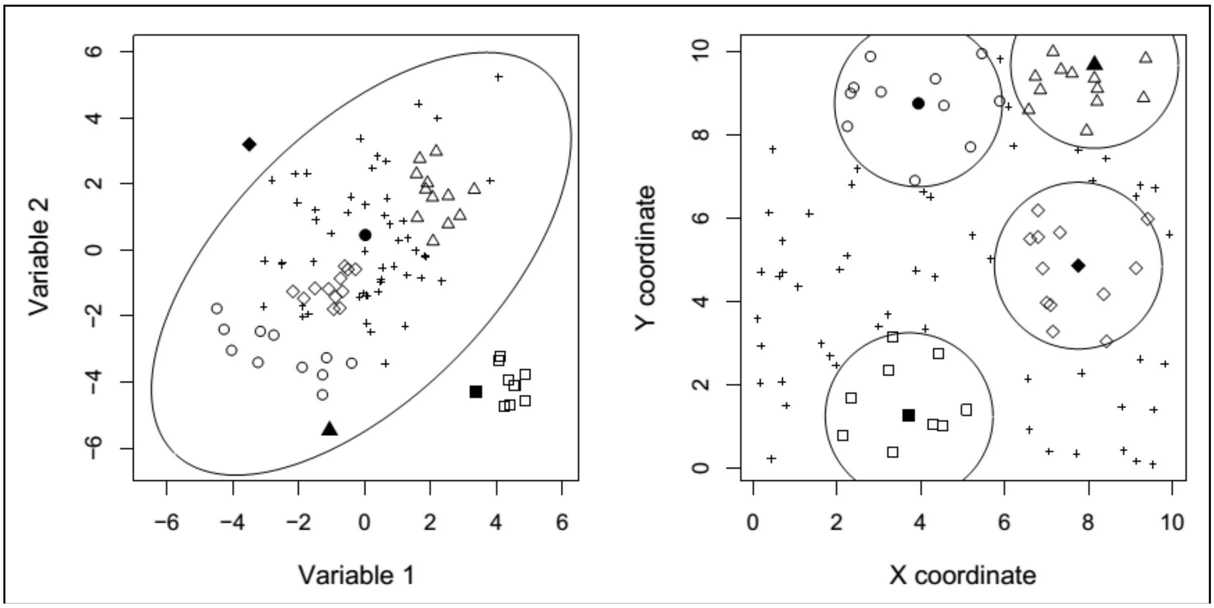

Figure 1. All types of outliers in a 2D dataset; (left) variable scatterplot space; (right) coordinate space [9].

3. Geological setting of studied area

The sampled area is situated within the eastern part of the so-called Lut block of eastern Iran. Eastern Iran, and particularly, the Lut block, has a great potential for different types of mineralization as a result of its past subduction zone tectonic setting, which lead to extensive magmatic activity forming igneous rocks of different geochemical compositions. The Lut block is characterized by extensive exposure tertiary volcanic and sub-volcanic rocks formed due to subduction prior to the collision of the Arabian and Asian plates [38-40]. Most of the studied area is covered by the upper Eocene-Oligocene altered volcanic rocks including andesite, dacite, tuff, and ignimbrite. These rocks are intruded by felsic to intermediate intrusive porphyritic rocks consisting of monzonite, diorite, and microgranodiorite porphyry stocks. Sedimentary rocks in this area consist of conglomerates, minor middle Eocene to upper Eocene tuffaceous marls in the southeastern to eastern area and Quaternary sediments [41]. The prospect area is similar to low-sulfidation epithermal systems. The rocks are dominantly altered andesite and dacite (Figure 2). Argillization, sericitization, and silicification are the common hydrothermal alterations in this area. Mineralization is not seen at surface [41].

The area comprises the moderately folded Tertiary volcanic zone of Kuh-e-Shah in the north, a zone of strongly tectonized and partly Eocene andesites and dacite-andesites. It is a zone of gently warped and tilted Upper Tertiary andesitic formations belonging to the Lut block. The tuff in the center of the area grades laterally and upwards into thick volcanic breccia, and locally, conglomerate with pebbles of alveolina and nummulitic limestone of Paleocene-early Eocene age. These clastic rocks are overlain by widespread dacites and dacite tuffs, which, due to the gradational contact relations with the underlying beds, are thought to be also of Paleocene age. The rocks tentatively attributed to Neogene are mainly various types of andesite. In the present area, the andesitic rock units seem to be genetically related to a characteristic formation of microdiorites, which protrude through the Paleogene volcanics. The volcanic breccias underlying the dacites extend eastward, where they are again not only associated with dacitic rocks but also with abundant, strongly altered, andesitic material. The Neogene andesites have been divided into several petrographic varieties, which seem to belong to different extrusive centers, and also differ slightly in age. Most widespread are pyroxene andesites. Some ancient workings for copper are found in the south of this area, apparently related to the aplitic intrusions in the area. Traces of malachite and chalcopyrite occur in several small ancient workings south and southeast of Hamich in dacitic and andesitic volcanic rocks of the Paleogene. Minor lead-zinc mineralisations with galena, cerussite, and smithsonite are found in the dacite and pyroxene andesite units south of this area [42].

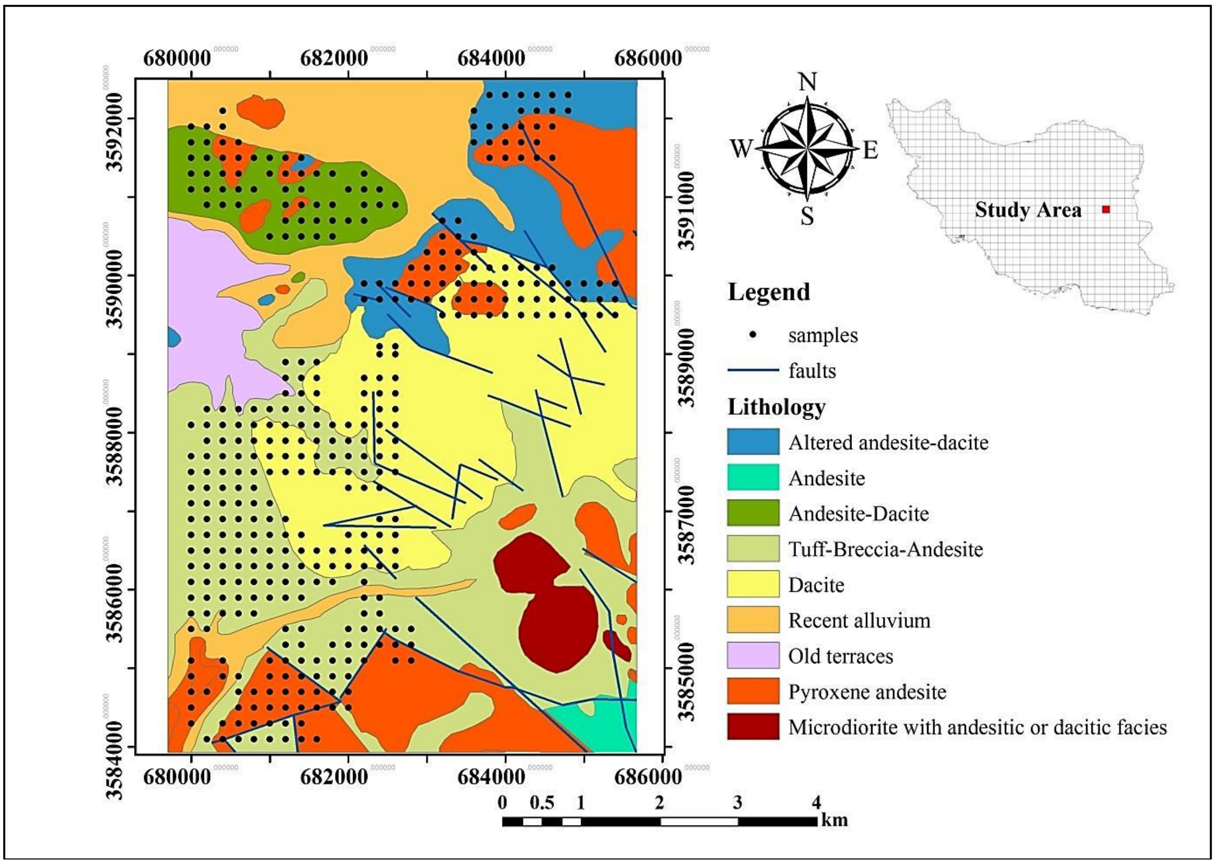

Figure 2. Lithological map of the study area and litho-geochemical sampling locations [42].

4. Data preparation, discussion and results

The dataset used in this research work was taken from a part of an exploration project carried out in southwestern of Birjand, South Khorasan, by IMTO[1]. The data consists of 396 litho-geochemical samples analyzed for 44 trace elements in Amdel lab in Australia. The study area covers a 40 km^2 rectangle with Hamich, a village and the only populated area, at the west. Geographically, it is an arid region with an almost hill and creek topography.

The sample locations are shown in Figure (2). Out of all the variables, Ag, B, Bi, Cd, Hg, Sn, and Te were removed because they had more than about 60% of missing values. The 37 remained variables were used in the imputation process. 53 samples of Au had zero values, and 1 of As, 6 of Co, 185 of Cr, 20 of Mo, 28 of Sb, 33 of Tl, and 15 of W were below the detection limit (BDL). These missing values were replaced using the recent technique of ilr-Em imputation in *zCompositions* package in R [43, 44]. Ilr-EM (that implements model-based ordinary and robust expectation-maximization algorithms) and ilr-DA (that implements a simulation-based data augmentation algorithm) to impute left-censored values are the best introduced methods that deal with the multivariate compositional structure of the data using an array of their analytical BDL values.

[1] Industry, Mine & Trade Organization (IMTO).

Then a sub-composition of variables was selected based on their geological relations and paragenetical properties. It included Au, As, Cu, Mo, Fe, Mg, Pb, Zn, Ni, Co, Cr, and W. First, using *mvoutlier* package in R [45], the matrix of the selected variables was ilr-transformed and the global multivariate and univariate outliers were determined and plotted (Figures 3 and 4). The map shows only the outliers that are out of the 0.975 quantile of the multivariate chi-square distributed MD based on the MCD estimator. Comparing Figures 3 and 4 gives us some clue that indices 1 to 13 and 74 in the southern part show anomalies of Cu, Fe, Pb, Zn, Au and indices 29, 33, 35, 40, 43, 44 show anomalies of Ni, Co, Cr, Pb, Zn. Index 56 shows a strong anomalies of As and Ni.

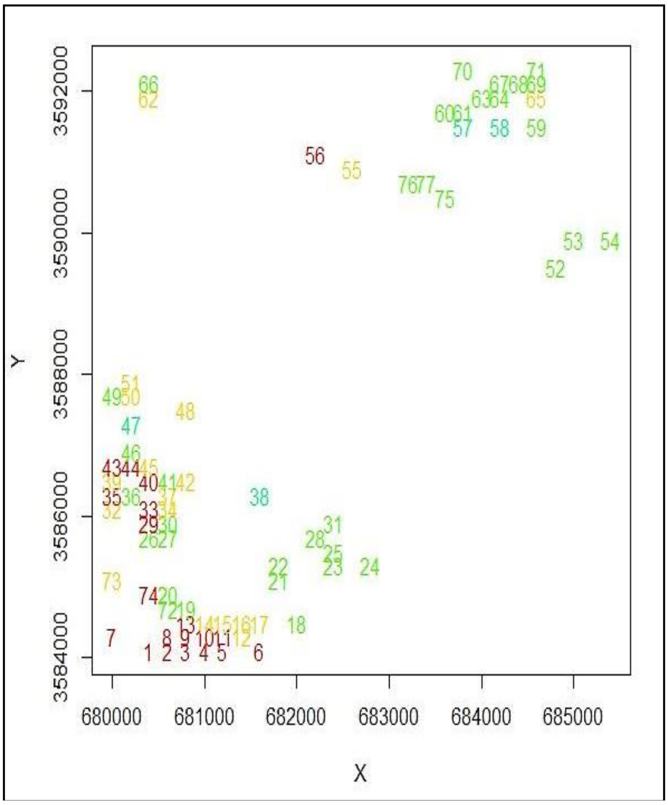

Figure 3. Global outliers map of the study area, indicating locations of outlying samples.

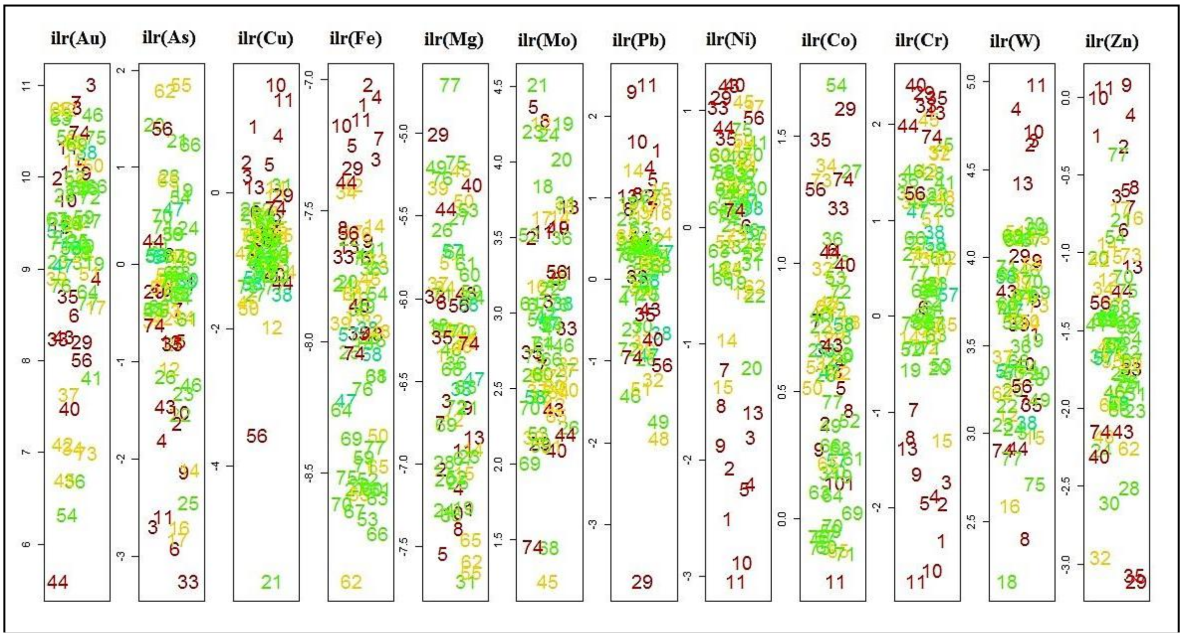

Figure 4. Univariate ilr-transformed data showing outliers.

Then, to determine the local outliers, we needed to define parameters (k, β). One way is to fix one item and change the other one to get the best value out of the sensitivity analysis. Thus using the first tool in *mvoutlier* package, we varied the number of neighbors of each observation using kNN with $k = 1, \ldots, 395$. Thus if $k = 395$, this means that all observations are neighbors of any observation except itself. The fraction β is arbitrarily fixed with 10% (though it is the best starting value). For each observation, the degree of isolation from $(1-\beta) = 90\%$ of its neighbors was computed using equation 5. Figure 5 shows the results in two separate plots for the regular observations and for global outlying observations. Each line in the plots belongs to one specific observation [9]. Because, here, we needed to study the anomalies, the considerable area in the plots is outlying observations (the right plot). The regular observations or inliers are like those inside the ellipsoid (cut-off value for chisquare of robust MDs) of Figure 1. Looking at the plot, for very small values of k, we observe some instability. The reason is that, just by chance, two observations could be close in the spatial sense but very different in the variable space. For a larger neighborhood, the local outlier measure becomes more reliable [9]. Thus in this plot, some observations for a neighborhood size of $k = 15$ show better exceptional behaviors than the others.

In the next step, fixing $k = 15$ using another tool in the package, the fraction β was varied to compute the best degree of isolation. Figure 6 shows the resulting plots for the inliers (left) and the outliers (right). The horizontal axes represent β and the vertical axes are degrees of isolation. It shows that isolatedness of observation from a varying percentage of the nearest 15 neighbors for some of them is significantly high. They are those that show the same isolatedness in Figure 5 for $k = 15$.

As it can be seen in Figures 5 and 6, the best isolatedness might be derived for $k = 15$ and $\beta = 0.1$. Then, at the final step, using these parameters, various local outliers were determined and mapped just for a box of the last 10 indices that were chosen from global outliers, as it can be seen in the left plot of Figure 7. Its x-axis shows the sorted observations according to the computed degree of isolation from a fraction of $(1-\beta)$ of their neighbors. This plot is also split into regular (left) and global (right) outliers.

The right plot in Figure 7 shows the spatial map of the potential local outliers of those selected by the box in the left plot together with their neighbors. Comparing the map in Figure 7 with that in Figure 3, it is evident that the local outliers may represent the geochemical anomaly halos around the global outlier samples 62, 66, 55, 56, 75, 76, 77 in the northern part, and 74, 6, 11, 18, 21, 22, 23, 25 in the southern part. If we wanted to study the halos around other global outliers or all of them, we could choose a wider box of indices.

Locations of some of the proposed borehole targets of identified anomalies in the detailed exploration report and the indices of mineralized Pb, Zn, and Cu are shown in Table 1. Comparing this data with the position of the identified local multivariate outliers confirms the accuracy of anomalies identified in Figure (3). Therefore, it could be verified that the geochemical halos of the anomalies defined by this method would be suitable prospecting targets for the detailed explorations of the next stage.

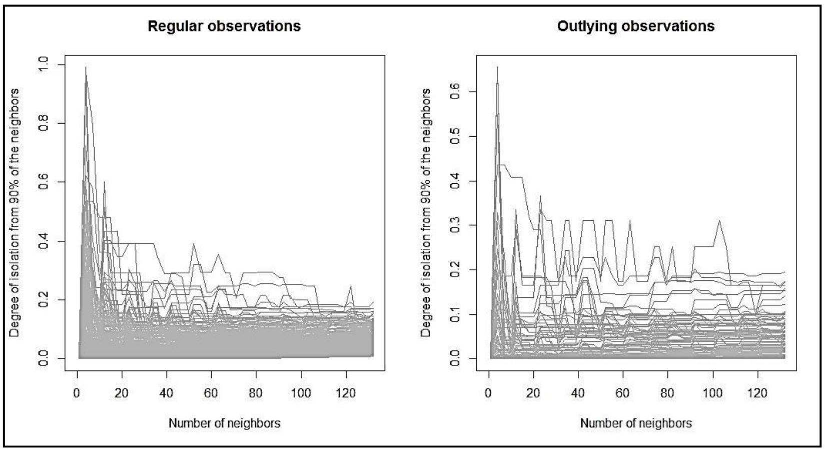

Figure 5. Degree of isolation of each observation (lines) from 90% of neighbors. Size of neighborhood is changed (horizontal axes).

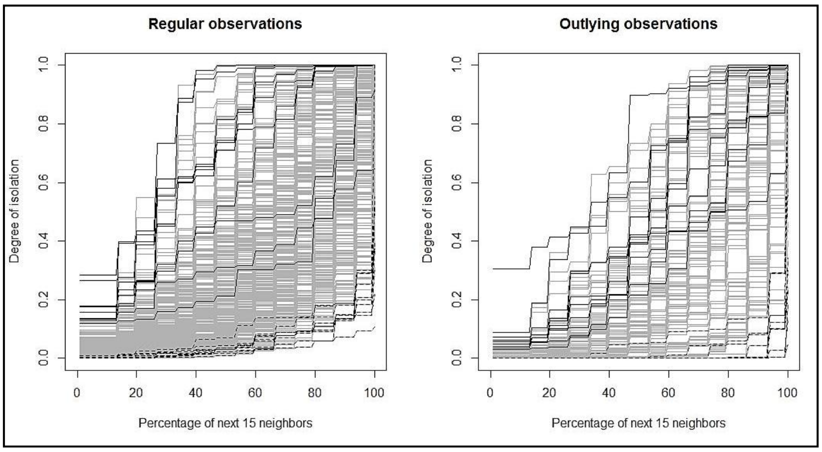

Figure 6. Degree of isolation of each observation (lines) from a varying percentage β of next 15 neighbors.

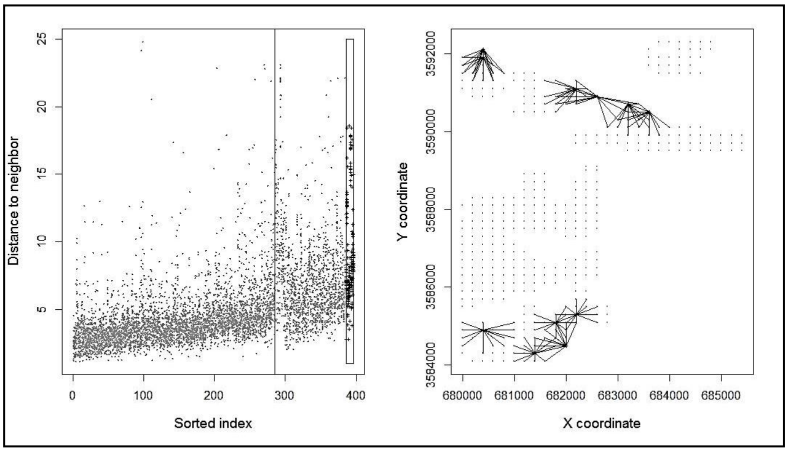

Figure 7. Local multivariate outliers of selected box of indices in the study area with $k = 15, \beta = 0.1$.

Table 1. Locations of proposed boreholes for core drillings as well as Pb, Zn, and Cu indices.

ID	X	Y
BH1	680274	3586595
BH3	680432	3586070
BH5	680558	3586108
BH9	679890	3586021
Cu index	684604	3584072
Pb-Zn index	681905	3583999
Cu index	680125	3584029

5. Conclusions

The reasons for local outlyingness of the marked points could be numerous. Some reasons are different data structures caused by local alterations of the rocks, exchanged samples, errors due to incorrect sample preparation, wrong laboratory analyses, and contaminations with different sources. Thus a much more detailed study of the area is required to recognize which source is more responsible for the causes of outlyingness. The first question that may rise is that how we can be so sure about the results. The answer, just as for all the anomaly identification methods, is that the most reliable and also simplest verification would be through field checking. In this research work, we compared the obtained locations of local outliers to the core drilling targets proposed in the detailed exploratory report of the area. Three out of four of them were just in the place of these local outliers. The analysis methods of anomaly separation that were used in the report to identify targets were the convenient ones that were methodologically completely different. Although this recent method is more reliable due to opening the closed system of compositional data, which nowadays is proven that if used raw (like in convenient classical methods of analysis), will deviate the related interpretations. On the other hand, the halos of local outliers also cover the mineral indices of Pb, Zn, and Cu marked in the geological 1:100,000 sheet and located in the southern part of the studied area. The principle that has to be noted is that the statistical dealing with this problem is not expected to yield fully complying solutions with the reality on the field of exploration due to the plenty of factors governing the geochemical transactions. However, it enlightens the way to further detailed explorations and important guidelines in making decisions to determine drilling points.

Acknowledgments

The Authors would like to express their special thanks to the South Khorasan Industry and Mine & Trade Organization (IMTO) for providing the required data. The financial support by the University of Birjand is also acknowledged.

References

[1]. Reimann, C. and Garrett, R.G. (2005). Geochemical background- concept and reality. Sci. Total Environ. 350 (1): 12-27.

[2]. Parslow, G.R. (1974). Determination of background and threshold in exploration geochemistry. J. Geochem. Explor. 3 (4): 319-336.

[3]. Botbol, J.M., Sinding-Larsen, R., McCAMMON, R.B. and Gott, G.B. (1978). A regionalized multivariate approach to target selection in geochemical exploration. Econ. Geol. 73 (4): 534-546.

[4]. Fletcher, W.K. (2013). Analytical methods in geochemical prospecting. Vol. 1.

[5]. Reimann, C., Filzmoser, P. and Garrett, R.G. (2005). Background and threshold: critical comparison of methods of determination. Sci. Total Environ. 346 (1): 1-16.

[6]. Parslow, G. (1974). Determination of background and threshold in exploration geochemistry. J. Geochem. Explor. 3 (4): 319-336.

[7]. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (2011). Robust statistics: the approach based on influence functions. Vol. 114. John Wiley & Sons.

[8]. Bamnett, V. and Lewis, T. (1994). Outliers in statistical data.

[9]. Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C. (2014). Identification of local multivariate outliers. Stat. Pap. 55 (1): 29-47.

[10]. Hawkes, H.E. and Webb, J.S. (1962). Geochemistry in Mineral ExplorationHarper and Row.

[11]. Reimann, C., Filzmoser, P. and Garrett, R.G. (2005). Background and threshold: Critical comparison of methods of determination. Sci. Total Environ. 346 (1-3): 1-16.

[12]. Tukey, J.W. (1977). Exploratory data analysis.

[13]. Cheng, Q., Agterberg, F.P. and Ballantyne, S.B. (1994). The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51 (2): 109-130.

[14]. Huber, P.J. (1981). Robust Statistics. Wiley. New york.

[15]. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (2011). Robust statistics: the approach based on influence functions. Vol. 114. John Wiley & Sons.

[16]. Leroy, A.M. and Rousseeuw, P.J. (1987). Robust regression and outlier detection. Wiley Ser. Probab. Math. Stat. N. Y. Wiley. Vol. 1.

[17]. Dutter, R., Filzmoser, P., Gather, U. and Rousseeuw, P. (2012). Developments in Robust Statistics: International Conference on Robust Statistics. Springer Science & Business Media.

[18]. Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C. (2014). Identification of local multivariate outliers. Stat. 55 (1): 29-47.

[19]. Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. Wiley-Intersci. N. Y. Vol. 15.

[20]. Atkinson, A. and Mulira, H-M. (1993). The stalactite plot for the detection of multivariate outliers. Stat. Comput. 3 (1): 27-35.

[21]. Chakraborty, A., Atkinson, A.C., Riani, M. and Cerioli, A. (2004). Exploring Multivariate Data with the Forward Search. JSTOR.

[22]. Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. Am. Stat. 45 (3): 234-242.

[23]. Dang, X.H., Micenková, B., Assent, I. and Ng, R.T. (2013). Local outlier detection with interpretation. in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 304-320.

[24]. Knorr, E.M., Ng, R.T. and Tucakov, V. (2000). Distance-based outliers: algorithms and applications. VLDB Journal- Int. J. Very Large Data Bases. 8 (3-4): 237-253.

[25]. Tao, Y., Xiao, X. and Zhou, S. (2006). Mining distance-based outliers from large databases in any metric space. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 394-403.

[26]. Knox, E.M. and Ng, R.T. (1998). Algorithms for mining distancebased outliers in large datasets. in Proceedings of the International Conference on Very Large Data Bases. pp. 392-403.

[27]. Mahalanobis, P.C. (1936). On the generalized distance in statistics. Proc. Natl. Inst. Sci. Calcutta. Vol. 2. pp. 49-55.

[28]. Rosseeuw, P. and Van Zomeren, B. (1990). Unmasking multivariate outliers and leverate points. J. Am. Stat. Assoc. Vol. 85. pp. 633-639.

[29]. Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). Robust Statistics: Theory and Methods. J Wiley.

[30]. Ernst, M. and Haesbroeck, G. (2013). Robust detection techniques for multivariate spatial data.

[31]. Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. Wiley-Intersci. N. Y. Vol. 15. pp. 105-209.

[32]. Anselin, L. (1995). Local indicators of spatial association-LISA. Geogr. Anal. 27 (2) 93-115.

[33]. Cerioli, A. and Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers. J. Comput. Graph. Stat. 8 (2): 239-258.

[34]. Aitchison, J. (1986). A Concise Guide to Compositional Data Analysis.

[35]. Pawlowsky-Glahn, V. and Buccianti, A. (2011). Compositional data analysis: Theory and applications. John Wiley & Sons.

[36]. Pawlowsky-Glahn, V. and Egozcue, J. (2006). Compositional data analysis in the geosciences: from theory to practice. Geol. Soc. Chapter Compos. Data Their Anal. Introd. pp. 1-10.

[37]. Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. Ann. Stat. pp. 461-482.

[38]. Camp, V. and Griffis, R. (1982). Character, genesis and tectonic setting of igneous rocks in the Sistan suture zone. eastern Iran. Lithos 15 (3): 221-239.

[39]. Berberian, M., Jackson, J.A., Qorashi, M., Khatib, M.M., Priestley, K., Talebian, M. and Ghafuri-Ashtiani, M. (1999). The 1997 May 10 Zirkuh (Qa'enat) earthquake (M(w) 7.2): Faulting along the Sistan suture zone of eastern Iran. Geophys. J. Int. 136 (3): 671-694.

[40]. Tirrul, R., Bell, I.R., Griffis, R.J. and Camp, V.E. (1983). The Sistan suture zone of eastern Iran. Geological Society of America Bulletin. 94 (1): 134-150.

[41]. Karimpour, M.H. and Mazaheri, S.A. (2009). Hydrothermal Alteration Mapping in SW Birjand, Iran, Using the Advanced Space borne Thermal Emission and Reflection Radiometer (ASTER) Image Processing. J Appl Sci. Vol. 9.

[42]. Vassighi, H., Soheili, M., Eftekharnejad, J. and Stoecklin, J. (1975). Geological 1:100000 sheet of Sar-e-chah-e-shur. 7754.

[43]. Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol. 35 (3): 253-278.

[44]. Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2015). zCompositions-R package for multivariate imputation of left-censored data under a compositional approach. Chemom Intell Lab Syst. Vol. 143. pp. 85-96.

[45]. Filzmoser, P. and Geschwandtner, M. (2011). mvoutlier: Multivariate outlier detection based on robust methods. Manual and package. version 1.9.1.

مقادیر خارج از ردیف چند متغیره محلی به عنوان تعیین‌کننده هاله‌های آنومالی ژئوشیمیایی مطالعه موردی: منطقه همیچ، خراسان جنوبی، ایران

حمید معینی[1] و احمد آریافر[2*]

۱- گروه معدن، دانشکده مهندسی معدن و متالورژی، دانشگاه یزد، ایران
۲- گروه مهندسی معدن، دانشکده مهندسی، دانشگاه بیرجند، ایران

ارسال ۲۰۱۶/۴/۲، پذیرش ۲۰۱۶/۷/۱۸

* نویسنده مسئول مکاتبات: aaryafar@birjand.ac.ir

چکیده:

شناسایی آنومالی همیشه یک موضوع بسیار مهم در اکتشافات ژئوشیمیایی مقدماتی بوده است. در طی پردازش داده‌های ژئوشیمیایی ناحیه‌ای، روش‌های مختلف آماری و داده‌کاوی معدنی وجود دارند که قادر به ارائه یک نتیجه به عنوان خروجی می‌باشند. مقادیر خارج از ردیف در بررسی‌های ژئوشیمیایی که داده‌ها تحت شرایط کنترل شده جمع‌آوری می‌گردند بسیار مورد توجه می‌باشند. این مقادیر در ژئوشیمی اکتشافی بیانگر رخدادهای کانی‌سازی می‌باشند بنابراین شناسایی آن‌ها بسیار مهم است. روش‌های پارامتری robust (بر اساس فواصل ماهالانوبیوس) و غیر پارامتری (بر اساس توابع عمق) برای شناسایی مقادیر خارج از ردیف چند متغیره در میان داده‌های ژئوشیمیایی، توسعه یافته‌اند. در این تحقیق روش شناسایی مقادیر خارج از ردیف چند متغیره محلی به منظور مشخص نمودن هاله‌های آنومالی ژئوشیمیایی در منطقه همیچ که در جنوب شرقی بیرجند، خراسان جنوبی، شرق ایران واقع شده است، بکار گرفته شد. برای این منظور تعداد ۳۹۶ نمونه لیتو ژئوشیمیایی که برای ۴۴ عنصر آنالیز گردیده بود، استفاده شد. نتایج به دست آمده نشان می‌دهد که انطباق خوبی بین زمین‌شناسی و اندیس‌های کانی‌سازی سرب، روی و مس در بخش جنوبی منطقه وجود دارد. چنین مطالعاتی می‌تواند توسط مجریان پروژه‌ها به منظور بهینه کردن محل‌های حفاری در مراحل اکتشاف تفضیلی مورد استفاده قرار گیرد.

کلمات کلیدی: داده ژئوشیمی، مقادیر خارج از ردیف چند متغیره محلی، آنومالی، خراسان جنوبی، همیچ.