Journal of Geochemical Exploration 270 (2025) 107662

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Delineation of geochemical anomalies through empirical cumulative distribution function for mineral exploration

Shahed Shahrestani, Ioan Sanislav *

College of Science and Engineering, Economic Geology Research Centre (EGRU), James Cook University, Townsville, Australia

ARTICLE INFO

Keywords:
ECOD
LOF
Geochemical anomaly
Mineral occurrences

ABSTRACT

In this paper, a statistical outlier detection technique based on empirical cumulative distribution functions (ECOD) is applied to a multivariate geochemical dataset from southeastern Iran, which is known for its porphyry and vein-type copper mineral occurrences. The ECOD method assumes that outlier samples are situated in both the left and right tails of the cumulative distribution functions, and it determines whether the outliers are located in the right or left tails using the concept of skewness. Anomaly maps produced by the ECOD method are compared with those generated by the local outlier factor (LOF) method. Both ECOD and LOF are applied to two subsets, including 4 and 12 trace elements. The anomaly maps are evaluated by comparing the number of delineated known mineral deposits and using ROC curves. The result revealed that LOF was outperformed by ECOD in the delineation of known Cu mineralization and in the identification of zones containing mineralized samples collected during the anomaly checking stage. The ECOD anomaly map is also compared with results from the k-means clustering method, and the superiority of ECOD over k-means clustering is demonstrated. The implementation of ECOD on clr-transformed multivariate geochemical data shows promise but assumes statistical independence among features, often unmet in geochemical exploration. To address this, we transformed clr data into new principal and independent feature spaces using principal component analysis (PCA) and independent component analysis (ICA), enhancing anomaly detection efficiency. ECOD_ICA outperformed ECOD_PCA, successfully classifying all mineralized samples and 15 of 18 Cu mineral occurrences in the highest score class (Q4), as confirmed by ROC analysis. However, the reliance of the ECOD method on univariate tail probabilities limits its ability to detect multivariate anomalies arising from complex inter-element relationships. Strong correlations in geochemical datasets can lead to false positives, necessitating dimension reduction techniques. While PCA and ICA help manage these correlations, they may obscure meaningful signals. The ECOD outlier detection method is also sensitive to the skewness of the dimensions, so a careful feature selection stage is recommended before applying it. The method is less sensitive to the number of dimensions, which enhances its robustness. Additionally, the absence of hyperparameter tuning makes ECOD a reliable and efficient outlier detection method.

1. Introduction

In outlier detection, geochemical anomalies are often analogous to outliers; however, unlike individual extreme values, they may also appear as clusters within the main data cloud, representing unusual or rare concentrations of elements that do not necessarily lie at the extremes. Background populations correspond to inliers, reflecting typical or expected geochemical patterns (e.g., Shahrestani and Carranza, 2024). During regional geochemical exploration, stream sediment or soil samples are analyzed for key elemental commodities and their pathfinders, creating a feature space suitable for unsupervised outlier detection methods (e.g., Ding et al., 2024; Shahrestani et al., 2024). A variety of outlier detection techniques have been developed, each utilizing different characteristics of multivariate datasets. These techniques include distance-based methods (e.g., nearest neighbor approaches), density-based methods (e.g., local outlier factor), statistical methods (e.g., kernel density estimation), learning-based approaches (e.g., one-class support vector machine), ensemble methods (e.g., bagging-based approaches), and neural network-based methods (e.g., autoencoders), to name but a few (e.g., Hinton and Zemel, 1993; Breunig et al., 2000; Yang

* Corresponding author.
E-mail address: ioan.sanislav@jcu.edu.au (I. Sanislav).

https://doi.org/10.1016/j.gexplo.2024.107662
Received 19 August 2024; Received in revised form 10 November 2024; Accepted 1 December 2024
Available online 2 December 2024

Table 1
Concentrations of Au, Cu, Pb, and Zn in mineralized samples collected in anomaly checking stage in the study area.

ID	Au (ppb)	Cu (ppm)	Pb (ppm)	Zn (ppm)
1	0.75	61	13	1108
2	0.75	2380	21	108
3	1	71	1129	968
4	1	74	1403	1126
5	3	86	13,440	11,622
6	5	65	1887	1600
7	13	134	17,370	13,350
8	18	1100	150	1840
9	20	941	1239	669
10	20	943	13	55
11	44	2824	230	797
12	71	5200	8	36
13	122	36	177	60
14	134	3450	107	405
15	184	74	121	61
16	201	186	194	93
17	476	3230	609	2140
18	561	6400	788	184

et al., 2023; Zhou et al., 2024).

Recently, there has been a notable trend in applying outlier detection methods to geochemical data to differentiate between anomalies originating from anomalous sources and those arising from background variations caused by lithological diversity (e.g., Chen and Wu, 2017; Shahrestani and Carranza, 2024; Mou et al., 2023; Esmaeiloghli et al., 2024; Hajihosseinlou et al., 2024). This trend reflects a growing recognition of the importance of accurate anomaly detection in guiding exploration efforts. Geochemical datasets often contain a mix of signals, including those that indicate mineralization and those that are merely the result of geological variations. Anomaly detection methods can help isolate the true signals of interest by identifying patterns that deviate significantly from the norm, thus distinguishing between anomalies and background levels (e.g., Shahrestani et al., 2019). As geochemical exploration becomes more sophisticated, the need for methods that can effectively handle complex datasets and reveal significant patterns has become crucial. Effective anomaly detection not only improves the efficiency of exploration but also enhances the accuracy of locating potential mineral deposits. This underscores the need to assess various outlier detection techniques to find those that best address the inherent complexities of geochemical datasets, ensuring that exploration efforts are both targeted and informed.

This paper evaluates the effectiveness of the empirical cumulative distribution functions for outlier detection (ECOD) method for detecting anomalies in a multivariate stream sediment geochemical dataset. The distribution-agnostic nature of ECOD and its robustness to correlations among elements, makes it a promising method for geochemical anomaly detection. The ability of ECOD method to handle diverse distributions, extreme values, and complex multi-element relationships, combined with its scalability for large, high-dimensional datasets and lack of need for hyperparameter tuning (Li et al., 2022), makes it a highly effective tool for identifying geochemical anomalies.

2. Geological setting and mineralization

The Baft district is situated within the Kerman Cenozoic Magmatic Arc (KCMA), part of the larger Urumieh-Dokhtar Magmatic Belt (UDMB) in Iran, covering approximately 2500 km^2 (e.g., Shafiei, 2010). This region features a diverse geological history with several distinct lithological units. The basement comprises Paleozoic metamorphic rocks, overlain by a Cretaceous mélange. Above this, Eocene volcanic and sedimentary rocks are present, which are further intruded by Miocene intrusive rocks. Oligocene-Miocene sedimentary and volcanic formations overlay the sequence (Srdic et al., 1972). The youngest units in the area are Quaternary deposits, including alluvial sediments, terraces, clays, and gravel fans (e.g., Ghasemzadeh et al., 2022).

In the northern and northeastern parts of the district, Eocene volcanic rocks dominate, including andesites, andesite-basalts, dacites, with minor basalts and rhyolites (e.g., Sepidbar et al., 2019). These formations are intruded by Upper Miocene quartz monzodiorite, granodiorite, monzonite, and Mio-Pliocene dioritic dikes (e.g., Jamali, 2017). These intrusions are surrounded by silicic and argillic hydrothermal alteration (e.g., Shafiei et al., 2009). The volcanic rocks contain Miocene intrusions known as the Lalezar granitoids or the Rabor-Lalehzar Magmatic Complex (RLMC) (Niktabar et al., 2015; Moghadam et al., 2018). The intrusions range from gabbrodiorites to granites, with the most felsic rocks exhibiting high-K calc-alkaline characteristics and slightly peraluminous compositions (e.g., Srdic et al., 1972). The mineral compositions of the granitoids include Na-plagioclase, quartz, alkali feldspar, biotite, and hornblende, with gabbro-diorite rocks featuring Ca-rich plagioclase, hornblende, biotite, and clinopyroxene (e.g., Dimitrijevic, 1973).

The Baft district hosts significant economic potential for porphyry copper deposits (e.g., Aghazadeh et al., 2015). Mineralization in the area includes porphyry, skarn, and vein systems. Notable porphyry deposits are found at Bid Khan, Ghaleh Asghar, Ghaleh Asghar_1, Harij (Haraan), Lalezar_2, Lalezar_4, Lalezar, and Lalezar_3, characterized by widespread copper and molybdenum mineralization. Vein-type mineralization is observed at Goghar, Ab Bahri, and Paynegin, associated with intrusive masses (Ghasemzadeh et al., 2019).

3. Materials and methods

3.1. Stream sediment data and preprocessing

A total of 911 stream sediment samples were collected across the Baft district, with an average sampling density of one sample per 1 km^2. The samples were analyzed for 49 elements using inductively coupled plasma - optical emission spectroscopy (ICP-OES). For Au, the fire assay method was used, followed by atomic absorption spectroscopy (AAS) for precise quantification. To assess analytical accuracy, duplicate specimens were analyzed using the method proposed by Howarth and Thompson (1976). This approach demonstrated that the precision of the results was better than 10 % for most elements (Ghasemzadeh et al., 2019). Along with stream sediment geochemical survey, a follow-up anomaly checking practice was also conducted in some zones and the main metal contents are reported in Table 1.

3.2. Anomaly detection method

Given a multivariate matrix X with n samples and d independent variables, anomalies are typically found in the low-density regions or tails of the probability distribution (Lazarevic and Kumar, 2005; Pokrajac et al., 2007). In this context, the empirical cumulative distribution (ECOD) method for outlier detection calculates the probability of a sample being at least as "extreme" as the observed data, based on tail probabilities (Li et al., 2022). In the multivariate space, let F: $R^d \rightarrow [0,1]$ denotes the joint cumulative distribution function (CDF). For any sample x in R^d, this CDF represents the probability distribution from which x is drawn. By defining the joint CDF for each x, the method assesses the extremity of each sample in the multivariate space.

$$F(x) = \mathbb{P}(X(1) \leq x(1), \ldots, X(d) \leq x(d)) \quad (1)$$

The probability derived from $F(X_i)$ indicates the extremeness of each X_i in the lower tails of the distribution. A smaller value of $F(X_i)$ suggests that X_i is less likely to be sampled from the same distribution as the rest of the data in X, especially when considering the condition $X \leq X_i$. Similarly, for the upper tail, $1 - F(X_i)$ can be used to measure extremeness. If either $F(X_i)$ or $1 - F(X_i)$ is small, it implies that X_i represents a rare or unusual realization and can be identified as an anomaly (Li et al., 2022).

Table 2
Spearman correlation values between anomaly score resulting from ECOD and LOF on two elemental subgroups.

	Spearman correlation		
	ECOD_Group_2	ECOD_Group_1	LOF_Group_2
ECOD_Group_1	0.76		
LOF_Group_2	0.54	0.50	
LOF_Group_1	0.58	0.68	0.51

In practice, estimating the true joint cumulative distribution function (CDF) for a dataset is achieved through empirical CDFs (ECDFs). However, the reliability or convergence of ECDF estimation decreases as the number of dimensions increases (Naaman, 2021). To simplify this challenge, it is often assumed that the different variables (dimensions) are independent. Thus,

$$F(x) = \prod_{j=1}^{d} F^{(j)}\left(x^{(j)}\right) \tag{2}$$

where $F^{(j)}$ represents the univariate CDF of the j-th dimension, satisfying $F^{(j)}(z) = P(X^{(j)} \leq z)$ for z in R^d. Accordingly, the univariate CDF for the left tail can be calculated through the empirical CDF (ECDF) as:

$$\widehat{F}_L^{(j)}(z) := \frac{1}{n}\sum_{i=1}^{n} I\left\{X_i^{(j)} \leq z\right\} \text{ for } z \text{ in } R \tag{3}$$

where I{·} denotes a binary function that is 1 when its argument is true and 0 otherwise. For the right tail, we can compute the univariate CDF as:

$$1 - F^{(j)}(z) = 1 - \mathbb{P}\left(X^{(j)} \leq z\right) = \mathbb{P}\left(X^{(j)} > z\right) \tag{4}$$

In ECOD approach the "right-tail" ECDF $\left(\widehat{F}_R^{(j)}\right)$ is also estimated as:

$$\widehat{F}_R^{(j)}(z) := \frac{1}{n}\sum_{i=1}^{n} I\left\{X_i^{(j)} \geq z\right\} \text{ for } z \text{ in } R \tag{5}$$

In the next step, considering the independence of the variables (dimensions), the joint left-tail and right-tail empirical cumulative distribution functions (ECDFs) ECDF ($\widehat{F}_L(x)$ and $\widehat{F}_R(x)$) for all dimensions d are estimated as:

$$\widehat{F}_l(x) = \prod_{j=1}^{d} \widehat{F}_l^{(j)}\left(x^{(j)}\right) \text{ and } \widehat{F}_R(x) = \prod_{j=1}^{d} \widehat{F}_R^{(j)}\left(x^{(j)}\right) \text{ for } x \text{ in } R \tag{6}$$

In ECOD, the first step involves estimating the joint left-tail $\widehat{F}_L^{(j)}\left(X_i^{(j)}\right)$ and right-tail $\widehat{F}_R^{(j)}\left(X_i^{(j)}\right)$ ECDFs for each dimension (Eqs. (3) and (5)). In the subsequent stage, these tail probabilities are integrated to compute the final outlier score Oi. These outlier scores are then used for sample comparison. An important aspect of aggregating tail probabilities is recognizing that focusing solely on either the left tail or the right tail for every dimension might not always be appropriate. Especially in high-dimensional data, where evaluating all 2^d possible combinations of left and right tail probabilities is infeasible, relying exclusively on one tail can limit outlier detection. For instance, being in the left tail of one dimension might indicate an outlier, while the right tail might be more relevant for another dimension. Thus, a flexible approach that adapts to the specific characteristics of each dimension is essential (Li et al., 2022).

In ECOD, the decision to use left or right tail probabilities is informed by skewness of the dimension. For dimensions with a negatively skewed distribution, where the left tail is longer and most observations are on the right, left tail probabilities are used. Conversely, for positively skewed dimensions, right tail probabilities are considered. This skewness-corrected version of ECOD, resulting in automatic outlier scores O_{auto}, incorporates this adjustment. The skewness of a dimension d is estimated using the following coefficient (Li et al., 2022):

$$\mu_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(j)} - \overline{X^{(j)}}\right)^3}{\sqrt{\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^{(j)} - \overline{X^{(j)}}\right)^2\right)^3}} \tag{7}$$

In this context, where the mean $\overline{X^{(j)}} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)}$ is used to determine the skewness coefficient μ_j. If μ_j is >0, points in the right tail are more likely to be outliers. Conversely, if μ_j is <0, samples in the left tail are considered potential outliers. The following equations are used to compute outlier scores (Li et al., 2022):

For the left-tail only score:

$$O_{L-only}(X_i) := -log\widehat{F}_L(X_i) = -\sum_{j=1}^{d} log\left(\widehat{F}_L^{(j)}\left(X_i^{(j)}\right)\right) \tag{8}$$

For the right-tail only score:

$$O_{R-only}(X_i) := -log\widehat{F}_R(X_i) = -\sum_{j=1}^{d} log\left(\widehat{F}_R^{(j)}\left(X_i^{(j)}\right)\right) \tag{9}$$

For the skewness-corrected score:

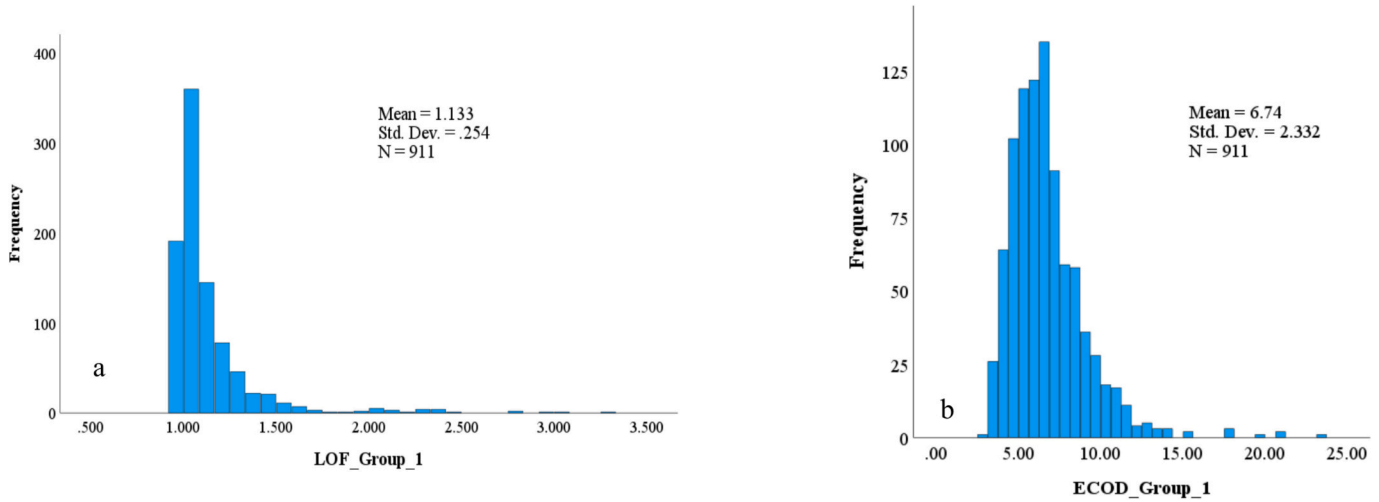

Fig. 1. Histograms of anomaly scores from LOF (a) and ECOD (b) applied to the Group_1 dataset.

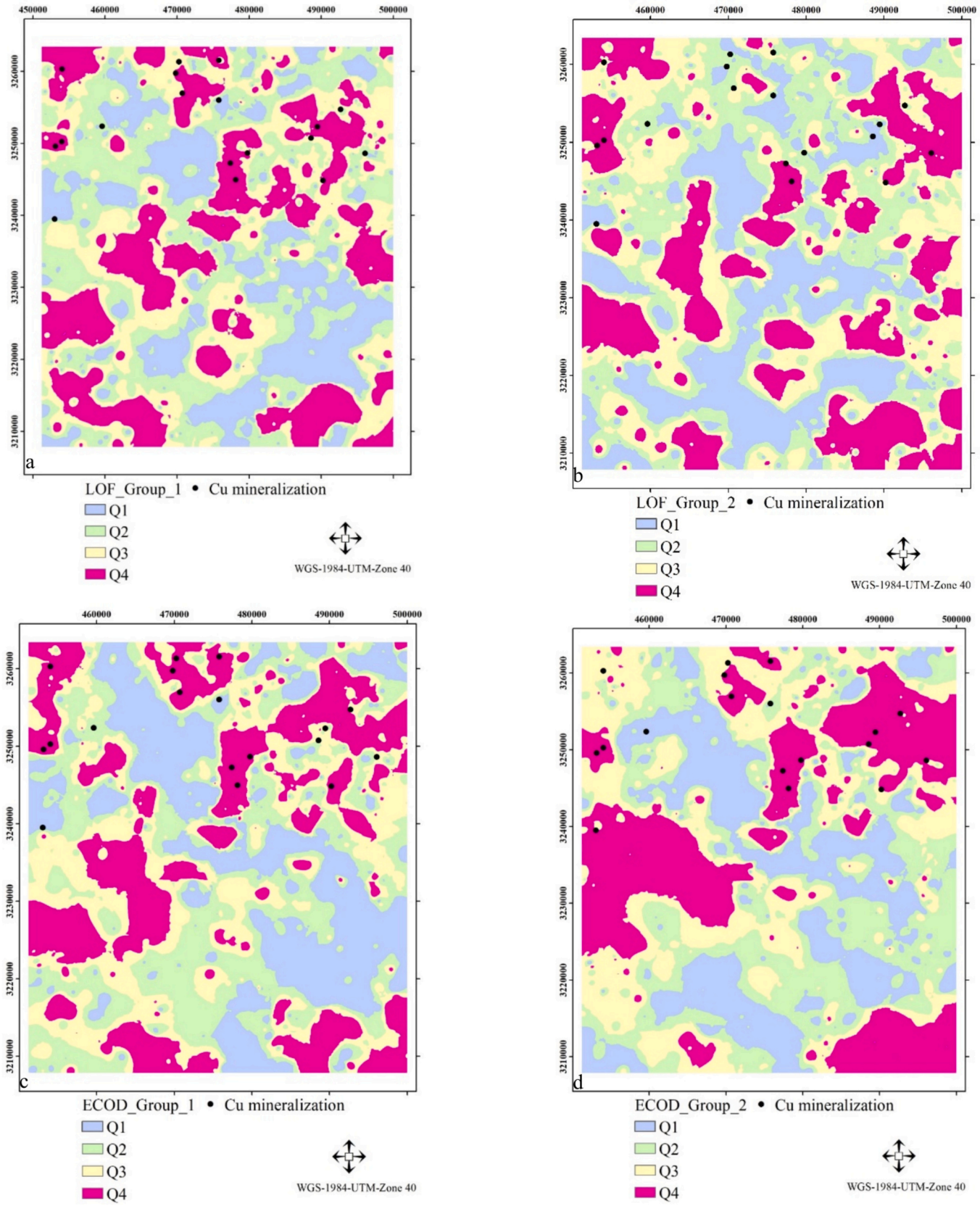

Fig. 2. Geochemical anomaly maps derived from ECOD and LOF outlier detection methods applied to multivariate stream sediment geochemical data from the Baft area. (a) LOF_Group_1, (b) LOF_Group_2, (c) ECOD_Group_1, (d) ECOD_Group_2.

Table 3
Comparison of the performances of LOF and ECOD in delineating known Cu mineralization in the Baft area.

Method	Class	Number of known deposits (out of 18)	Percentage
ECOD_Group_1	Q1	2	11.1
	Q2	2	11.1
	Q3	3	16.7
	Q4	11	61.1
ECOD_Group_2	Q1	1	5.6
	Q2	1	5.6
	Q3	2	11.1
	Q4	14	77.7
LOF_Group_1	Q1	2	11.1
	Q2	3	16.7
	Q3	4	22.2
	Q4	9	50.0
LOF_Group_2	Q1	3	16.7
	Q2	6	33.3
	Q3	3	16.7
	Q4	6	33.3

$$O_{auto}(X_i) = -\sum_{j=1}^{d}\left[\mathrm{I}\left\{\mu_j < 0\right\} log\left(\widehat{F}_L^{(j)}\left(X_i^{(j)}\right)\right) + I\left\{\mu_j \geq 0\right\} log\left(\widehat{F}_R^{(j)}\left(X_i^{(j)}\right)\right)\right] \quad (10)$$

The final anomaly score O_F is determined by selecting the highest value among the left-tail only score, right-tail only score, and skewness-corrected score for each sample Xi (Li et al., 2022):

$$O_F = max\left\{O_{l-only}(X_i), O_{L-only}(X_i), O_{R-only}(X_i)\right\} \quad (11)$$

ECOD demonstrates robustness in high-dimensional data through its strategic use of empirical cumulative distribution functions (ECDFs) computed independently for each feature, effectively addressing challenges associated with the curse of dimensionality. In high-dimensional settings, calculating a joint ECDF that accurately represents all variables becomes increasingly difficult, as the joint ECDF converges to the true cumulative distribution function (CDF) at a slower rate with more dimensions, complicating reliable joint probability estimation. ECOD avoids this issue by computing a univariate ECDF for each feature separately rather than relying on a full multivariate ECDF. The outlier score for each data point is derived by assessing tail probabilities across all features based on the assumption of independence between features. This method evaluates "outlyingness" by multiplying tail probabilities from each univariate ECDF and considers both left and right tails of each feature, with calculations performed in log space to stabilize computation and integrate contributions from all feature tails. Although the approach assumes independence between features—a potentially restrictive assumption—it has been demonstrated to be highly effective in practice (Naaman, 2021; Li et al., 2022). By using feature-wise ECDFs and tail probability aggregation, ECOD circumvents the limitations of distance-based methods, which tend to suffer from distance uniformity in high dimensions, thereby preserving robustness and scalability as dimensionality increases.

4. Results and discussion

Multivariate geochemical data are reported as compositional, meaning each element's concentration is constrained by a constant sum (e.g., total concentrations must equal 100 %). This compositional constraint introduces interdependencies among elements, resulting in spurious correlations that may not represent genuine geochemical associations but are artifacts of the compositional structure (e.g., Aitchison and Egozcue,

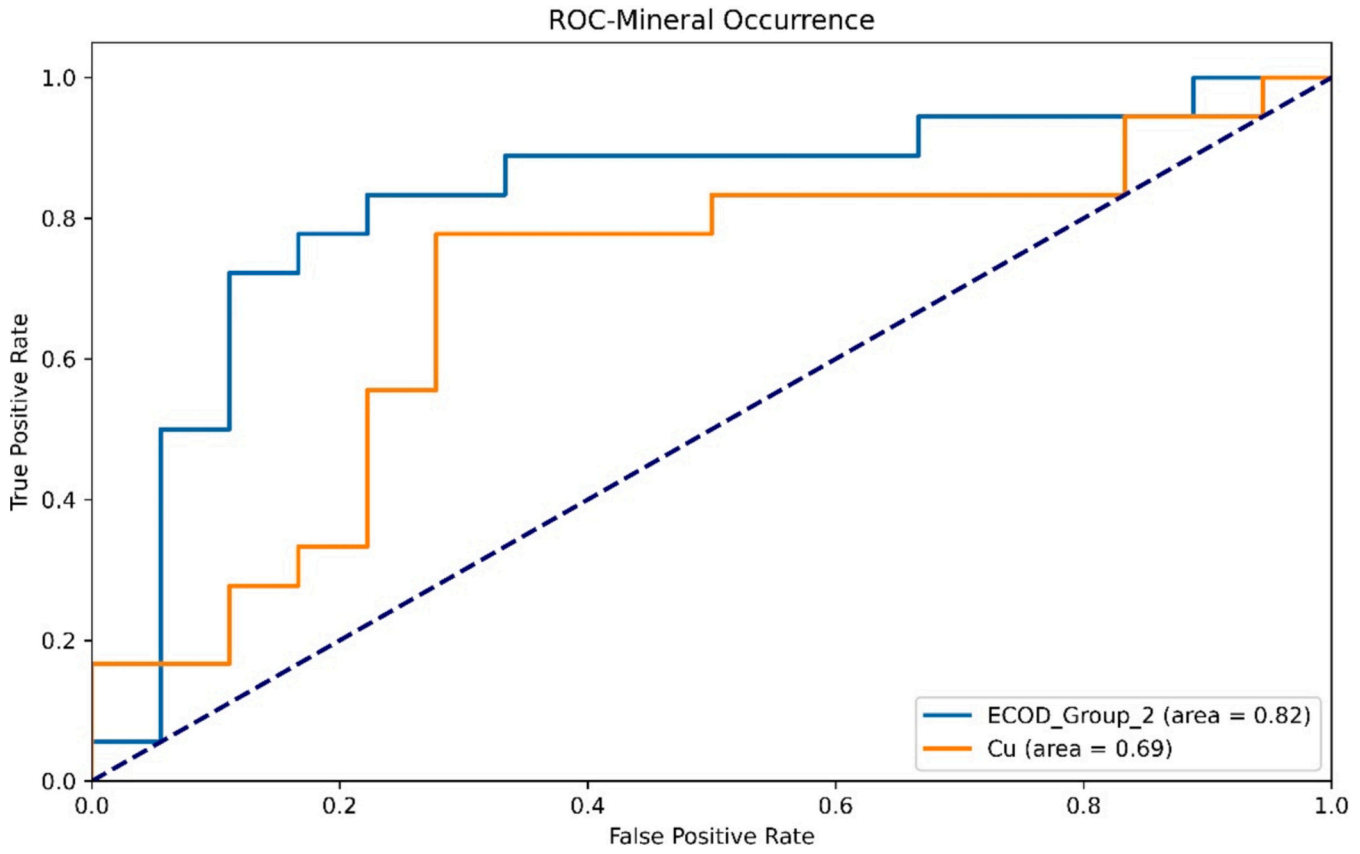

Fig. 3. ROC curves for comparing the efficiency of univariate Cu data and multivariate outlier detection method (ECOD) in delineating geochemical anomalies derived from anomalous sources in the study area.

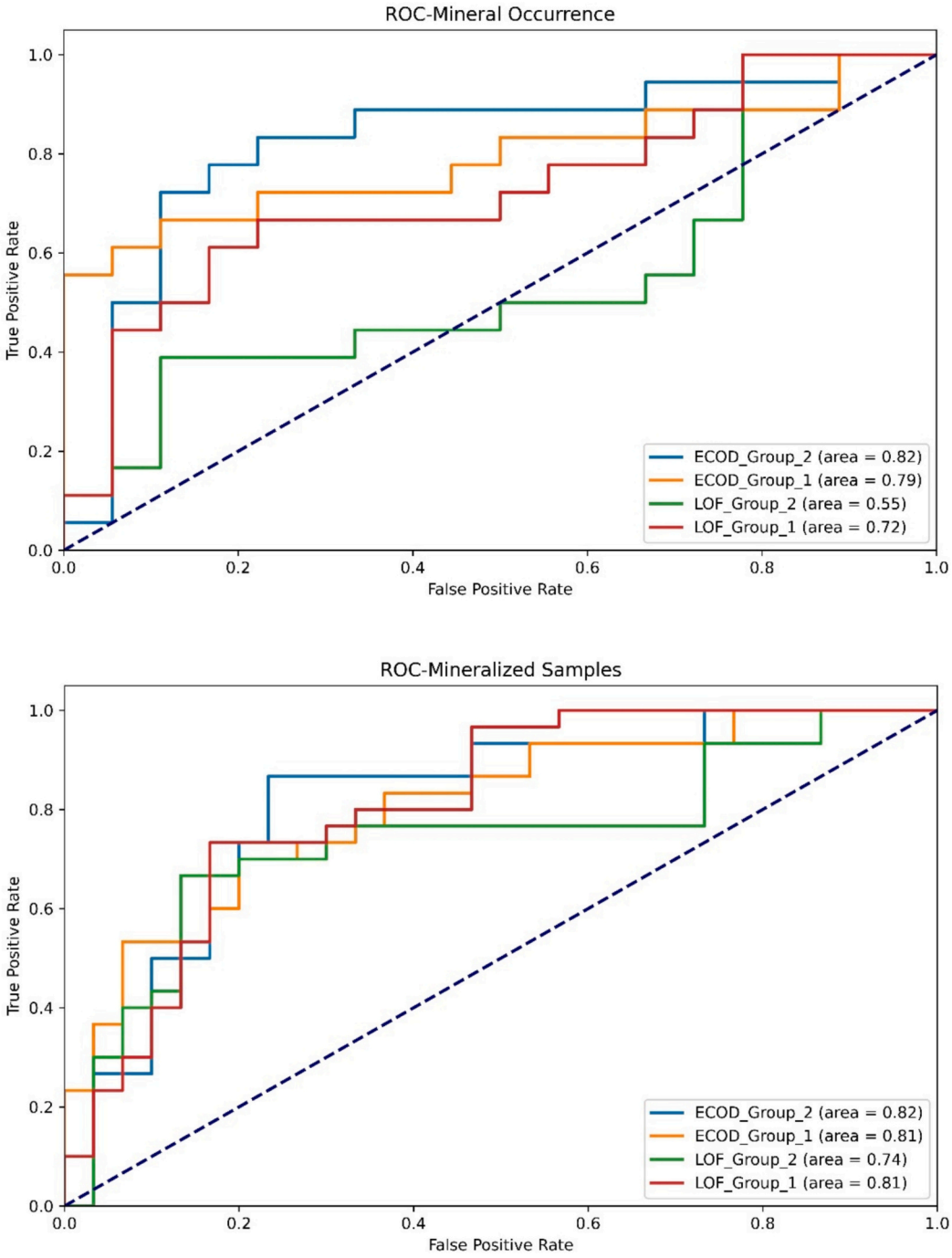

Fig. 4. ROC curves for comparing the efficiency of multivariate outlier detection methods (ECOD and LOF) on two subgroups in delineating geochemical anomalies derived from anomalous sources in the study area (up: known mineral occurrences, down: mineralized samples).

2005). These dependencies complicate the application of statistical methods like ECOD, which assume independence among variables. Without appropriate transformations, such as clr transformation, these artificial correlations can compromise anomaly detection, potentially leading to misinterpretations in geochemical analysis. In this regard, multivariate geochemical data first underwent clr-transformation to tackle the compositional characteristics of the geochemical dataset (Aitchison, 1986). In this study, the feature space includes two sets of variables. The first set, called Group_1, consists of four main trace elements: Au, Cu, Pb, and Zn. The second dataset, called Group_2, includes the main trace elements and their pathfinders: Ag, As, Au, Bi, Co, Cr, Cu, Mo, Ni, Pb, Sb, and Zn. These datasets were selected to consider the impact of different feature space sizes on the performance of the algorithm and to examine if adding new elements in the data analysis procedure improves the efficiency of the anomaly detection methodology. To evaluate the performance of the ECOD method, the emerging outlier scores are compared with those from a well-established outlier detection method called local outlier factor (LOF). The LOF method is a density-based anomaly detection technique that identifies local outliers by comparing the local density of a data point to that of its neighbors. It calculates the local reachable density for each point and generates a local outlier factor score. Points with significantly lower local densities compared to their neighbors are classified as outliers (Breunig et al., 2000). For more details on the LOF method, readers are referred to

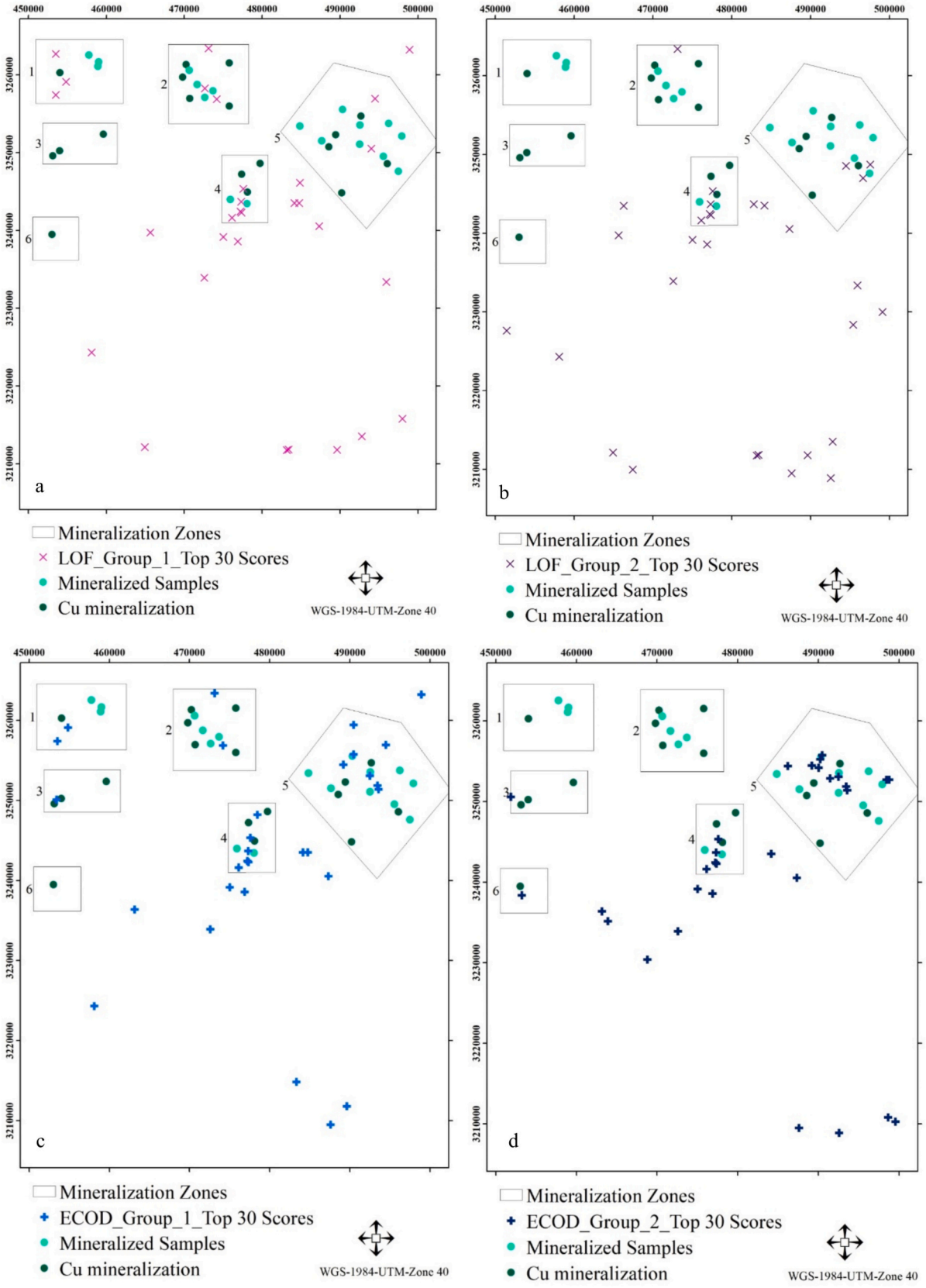

Fig. 5. Spatial distribution of top 30 outlier scores of (a) LOF_Group_1, (b) LOF_Group_2, (c) ECOD_Group_1, (d) ECOD_Group_2 outlier scores relative to the location of known mineral occurrences and anomaly checking mineralized samples.

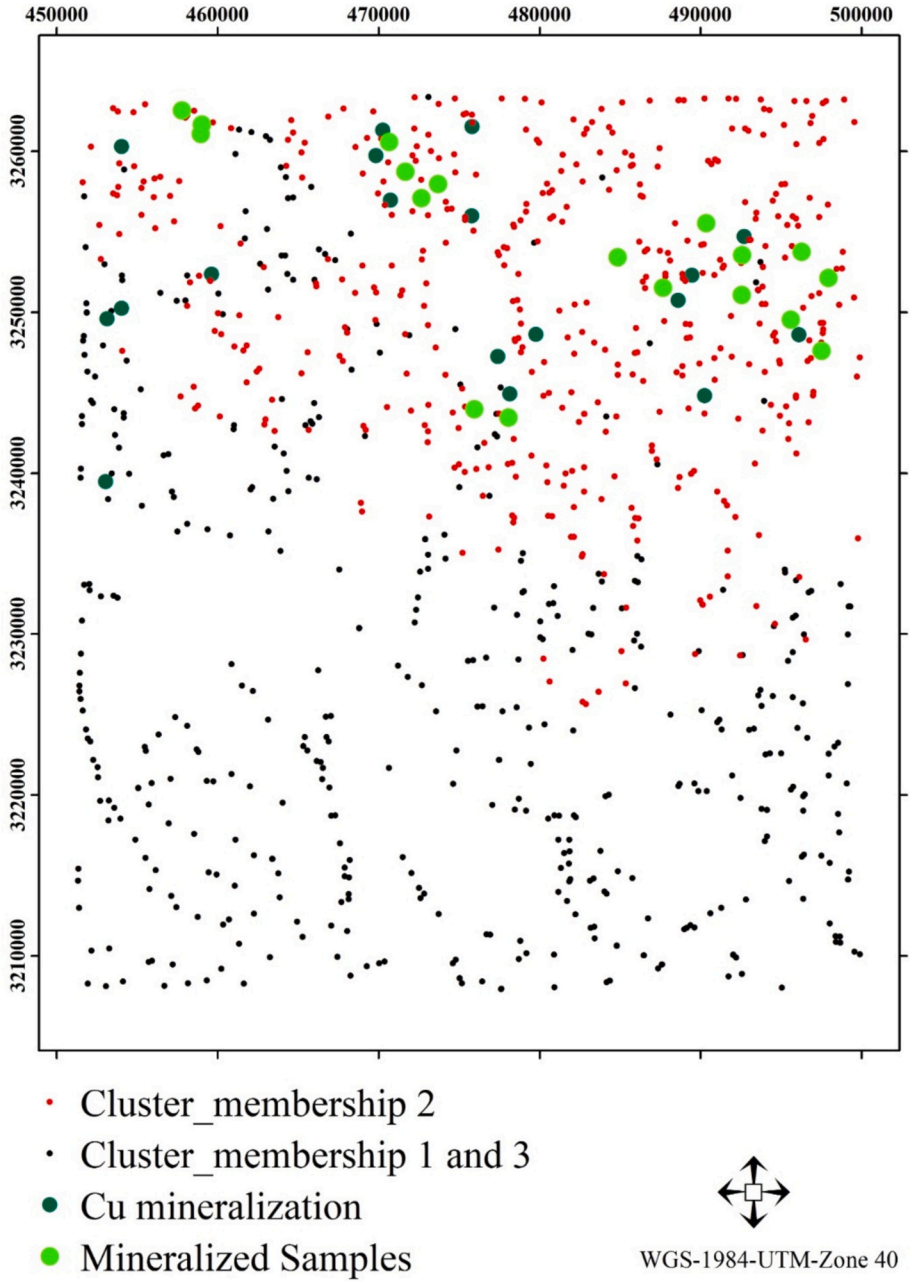

Fig. 6. Categorization of stream sediment samples based on their cluster membership, as determined by the k-means clustering method applied to the Group_2 dataset.

Breunig et al. (2000), Shahrestani and Carranza (2024) and Puchhammer et al. (2024). In this paper, the Python implementation of ECOD (https://github.com/yzhao062/pyod/blob/master/pyod/models/ecod.py) (Li et al., 2022) is used to determine the anomaly scores. The ECOD method is parameter-free and does not require hyperparameter tuning. The LOF method is implemented using the Python 'sklearn' library.

Table 2 shows the correlation between anomaly scores from Group_1 and Group_2 using the ECOD and LOF methods. The data shows that the anomaly scores from the ECOD method are more closely correlated than those from the LOF method. This suggests that ECOD produces more stable anomaly scores when dimensionality changes, compared to LOF. While all four sets of anomaly scores are correlated to some extent, noticeable differences between the scores lead to varying performances across the two datasets and the two algorithms. Fig. 1 shows the histograms of anomaly scores from LOF and ECOD applied to the Group_1 dataset. The ECOD method exhibits a greater range of anomaly scores (2.99 to 23.34) compared to the LOF (0.94 to 3.29). This greater range demonstrates that different samples can be more effectively contrasted, providing more information from the ECOD anomaly scores than from the LOF.

Fig. 2 depicts the interpolated map of anomaly scores resulting from the implementation of ECOD and LOF on Group_1 and Group_2 datasets. To evaluate the proficiency of ECOD in delineating geochemical anomalies derived from mineralization zones, three methodologies are considered. In the first comparison, the number of known mineral occurrences classified in each quantile is compared. Quantile-based mapping of geochemical anomalies is effective, particularly when working with outlier scores from various methods that exhibit diverse distributions. By dividing the data into equal portions based on rank instead of absolute values, quantiles provide a distribution-independent approach. This method ensures a uniform distribution of data across each quartile, regardless of whether the underlying score distribution is normal,

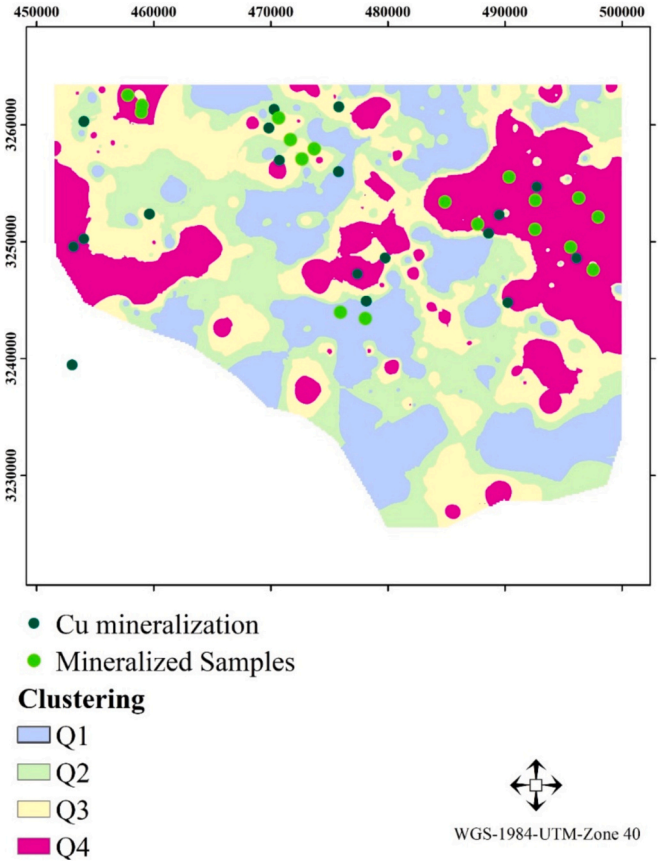

Fig. 7. Anomaly map based on interpolated distance to cluster center resulting from k-mean clustering method applied on stream sediment samples from the study area.

skewed, or multimodal (e.g., Liu et al., 2024). Table 3 presents the overall performance of ECOD and LOF in delineating 18 known Cu mineral occurrences in the study area. The fourth quantile class (Q4) of ECOD_Group_2 encompasses the highest number of delineated mineral occurrences among other realizations, accounting for about 77 % (14 out of 18). Overall, ECOD outperforms LOF in delineating geochemical anomalies related to known Cu mineral occurrences. LOF_Group_1 shows better performance relative to LOF_Group_2, as LOF, being a density-based method, is more sensitive to the number of dimensions. In other words, the concepts of proximity and distance are less pronounced as dimensionality increases. The superiority of ECOD_Group_2 over ECOD_Group_1 can be attributed to the additional information provided by variables such as Co, Ni, Pb, and Mo in delineating volcanic-related or vein-type Cu mineralization. In the case of LOF, the performance significantly decreases when considering 12 trace elements. The interpolated map fails to differentiate between anomalous and non-anomalous zones due to high sensitivity of LOF to the size of the feature space (i.e., curse of dimensionality). Additionally, geochemical anomalies can originate from lithological variations, and in the current case study, the significance of trace elements is under investigation.

One common approach to examining the performance of anomaly detection methods is using the receiver operating characteristic (ROC) curves, where the area under the curve serves as a proxy for performance of the method. A variation of these diagrams incorporates true positive and false positive rates (e.g., Fawcett, 2006). However, these curves require labeled data to evaluate efficiency of the method. In geochemical exploration, the locations of known mineral occurrences or any evidence that proves mineralization can serve as labeled data. In this study, both the locations of known mineral occurrences and 30 samples collected during the anomaly checking stage are used to depict the ROC curves. Additionally, false positive samples are randomly selected, with their corresponding values sampled from interpolated geochemical anomalies. The use of random locations provides a contrast between anomalies that are known to be related to mineral occurrences and those scores derived from locations that may or may not exhibit mineralization. Additionally, in the context of outlier detection, the output anomaly scores may have meaningful ranges for samples near known mineralization, and using random locations may help highlight these relationships. However, it is important to acknowledge that definitively designating false positives is challenging, particularly in regional-scale geochemical studies where proving the absence of mineralization is difficult. In the first comparison, a univariate Cu map is contrasted with the ECOD_Group_2 anomaly map (Fig. 3). The implementation of multivariate anomaly detection considerably outperforms the univariate Cu anomaly map. The failure to delineate known Cu mineralization using univariate Cu anomaly maps can be attributed to relatively high copper mobility, dilution of Cu geochemical signals, high background levels of Cu content, and insufficient sampling density.

In the next comparison, ROC curves are determined for four multivariate anomaly detection methods, using known Cu mineral occurrences and mineralized samples as validation points (Fig. 4). Among the anomaly maps, the ECOD method using 12 trace elements (ECOD_Group_2) demonstrates superiority over the other subset (Group_1) and the LOF method. Additionally, ECOD and LOF show similar performances when only Cu, Au, Pb, and Zn are considered. As expected, increasing the number of dimensions reduces the effectiveness of density-based outlier detection in the LOF method, as evidenced by the number of detected known mineral occurrences. For checking anomalies against mineralized samples, three anomaly maps including ECOD_Group_1, ECOD_Group_2, and LOF_Group_1 exhibit corroboration with the locations of mineralized samples. However, the LOF_Group_2 anomaly map demonstrates that increasing the number of variables reduces the contrast between geochemical anomalies and the background.

To delve into the anomaly scores emerging from the two subsets and the two methods, the top 30 anomaly scores from each method are extracted. The spatial distribution of these top anomaly scores relative to mineral occurrences and mineralized samples is depicted in Fig. 5, with the six mineralization zones also highlighted. There is no strong consistency between the top anomaly scores and validation points for LOF_Group_2, where most of the top 30 scores are concentrated in the southern part of the study area, except in zone 4, which shows a spatial correlation between validation points and scores. LOF_Group_1 indicates conformity between validation points and top anomaly scores in zones 1, 2, 4, and 5. However, in zone 5, which has several known mineralization, only two of the top 30 scores are observed. For ECOD_Group_1, there is a notable conformity between anomaly scores and known mineralized samples in zones 1, 2, 3, 4, and 5. Additionally, the top scores are near the validation points in zones 3, 4, and 5, demonstrating the efficiency of ECOD in capturing mineralization-related geochemical anomalies. High corroboration between validation points and the top 30 anomaly scores is also observed in zones 4, 5, and 6 when considering 12 trace elements in the ECOD algorithm. However, for the other top anomaly scores emerging from ECOD_Group_1, ECOD_Group_2, and LOF_Group_1 in the southern part, there is no proven evidence of mineralization, and differentiating between false and true positive geochemical anomalies is not practical for these anomaly scores.

To investigate the efficiency of the ECOD multivariate anomaly detection in delineating geochemical anomalies, the ECOD_Group_2 anomaly map is compared with the k-means clustering method. First, k-means clustering is performed on 12 elemental concentrations, considering 3 to 6 clusters. Among these, a clear distinction is observed between the northern part of the study area, which contains known mineral occurrences, and the southern part, with three distinct clusters (Fig. 6). Stream sediments within the northern cluster are retained, and the distance to the cluster center is interpolated as a proxy for the outlierness of samples (Fig. 7). Similarly, outlier scores from the

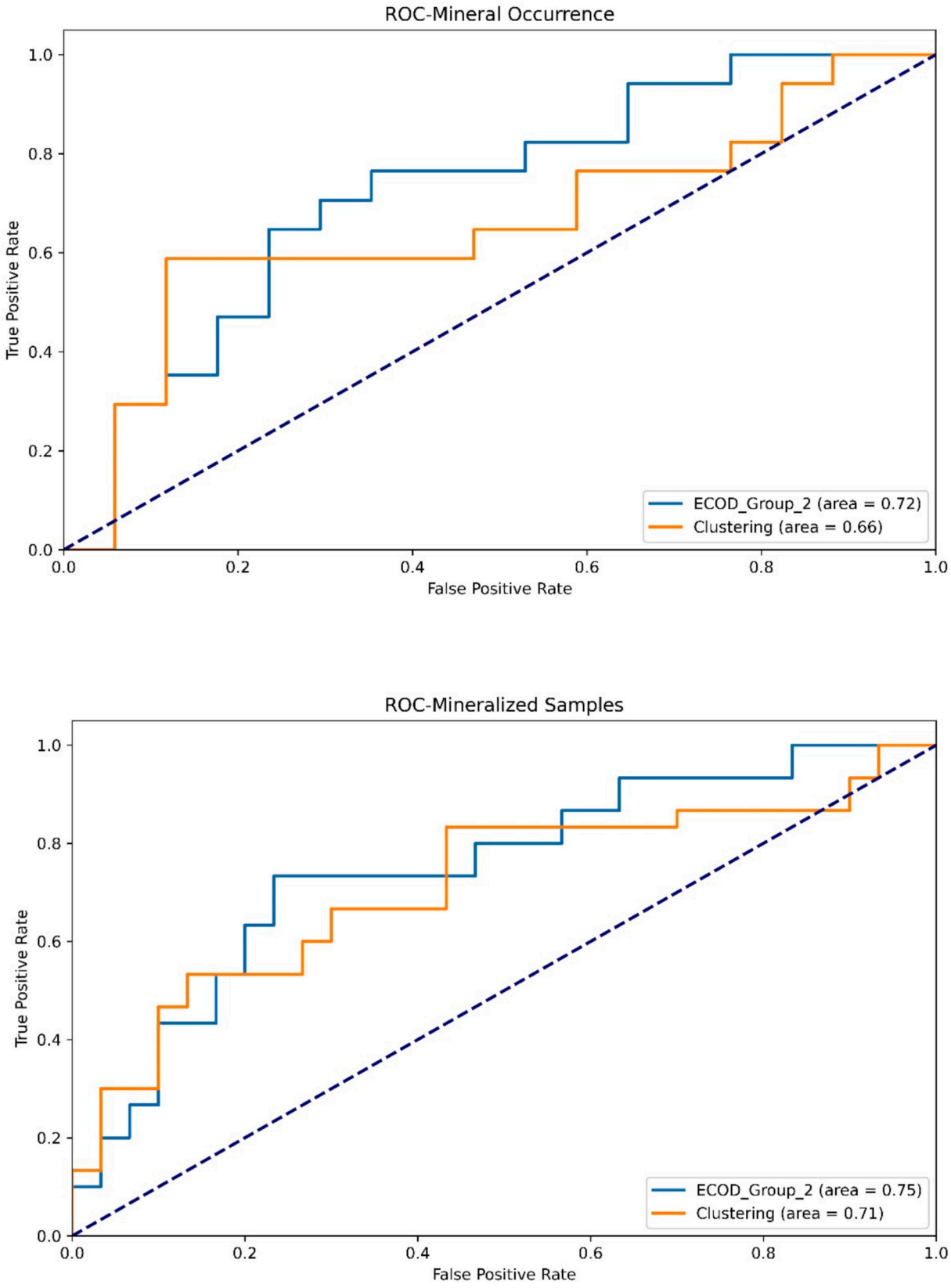

Fig. 8. ROC curves comparing the efficiency of the multivariate outlier detection method (ECOD) applied to subgroup 2 with the k-means clustering approach in delineating geochemical anomalies in the study area (Top: known mineral occurrences; Bottom: mineralized samples).

Table 4
Ranking of elements based on skewness values and correlation with ECOD_Group_2 anomaly scores.

Rank	Correlation with ECOD_Group_2 scores	Element	Rank	Skewness	Element
1	0.437	Pb	1	3.99	Pb
2	0.367	Ag	2	1.87	Zn
3	0.344	Sb	3	1.82	Au
4	0.194	Zn	4	1.61	Ag
5	0.119	Bi	5	1.08	Sb
6	0.078	Au	6	1.07	Bi
7	0.027	As	7	0.94	Ni
8	0.002	Cu	8	0.93	Mo
9	−0.014	Mo	9	0.91	Cu
10	−0.065	Cr	10	0.74	As
11	−0.066	Ni	11	0.65	Cr
12	−0.089	Co	12	0.29	Co

ECOD_Group_2 approach are interpolated separately for samples with membership 2. ROC curves are drawn using known mineral occurrences and mineralized samples as validation points (Fig. 8). As shown in the ROC curves, the area under the curve is higher for the ECOD_Group_2 anomaly map compared to the clustering approach.

ECOD considers both the left and right tails of the cumulative distribution function across various dimensions. The decision to consider either the right or left tail is based on the skewness of the variables. This raises the question of the relationship between the degree of skewness in a dimension and its contribution to the final ECOD anomaly score. To address this, the skewness value of each dimension is calculated (Table 4). Additionally, the correlation values between clr-transformed independent variables and the resulting ECOD_Group_2 anomaly scores are determined (Table 4). In Table 4, elements are ranked based on their skewness and their correlation with ECOD_Group_2 scores. Pb ranks highest in both correlation (0.437) and skewness (3.99), indicating a strong influence on the anomaly scores. Similarly, Ag, Sb, and Zn also show high correlations and skewness values, highlighting their significant contributions to the anomaly detection. Elements like Bi and Au, despite their moderate skewness, have lower correlations, suggesting a lesser impact. Conversely, elements such as As, Cu, Mo, Cr, Ni, and Co exhibit negative or near-zero correlations, coupled with lower skewness, implying minimal influence on the ECOD_Group_2 scores. Overall, a similar trend can be observed in the elemental rankings, indicating that elements with higher skewness values are more likely to contribute significantly to the emerging anomaly scores. This has both advantages and disadvantages. Therefore, it is crucial to conduct a preselection stage before applying the ECOD method. Elements with no geochemical affinity with mineralized sources but with high skewness may have a larger impact on anomaly scores compared to those with relatively lower skewness but an apparent connection to mineralization. Thus, ECOD should be applied after a feature selection stage that combines statistical approaches and incorporates the geochemical characteristics of mineral deposits. In the current study, mineralization includes known porphyry-type mineral deposits, where the connection between most of the geochemical anomalies of 12 elements and anomalous sources is validated. It is important to note that all the input dimensions have positive skewness values ($\mu_j \geq 0$), indicating that the right-tail parts of the cumulative distribution functions are used to assign scores to samples in the ECOD method.

While the implementation of ECOD showed promising results on clr-transformed multivariate geochemical data, one strong assumption of ECOD is that each feature should be statistically independent, a

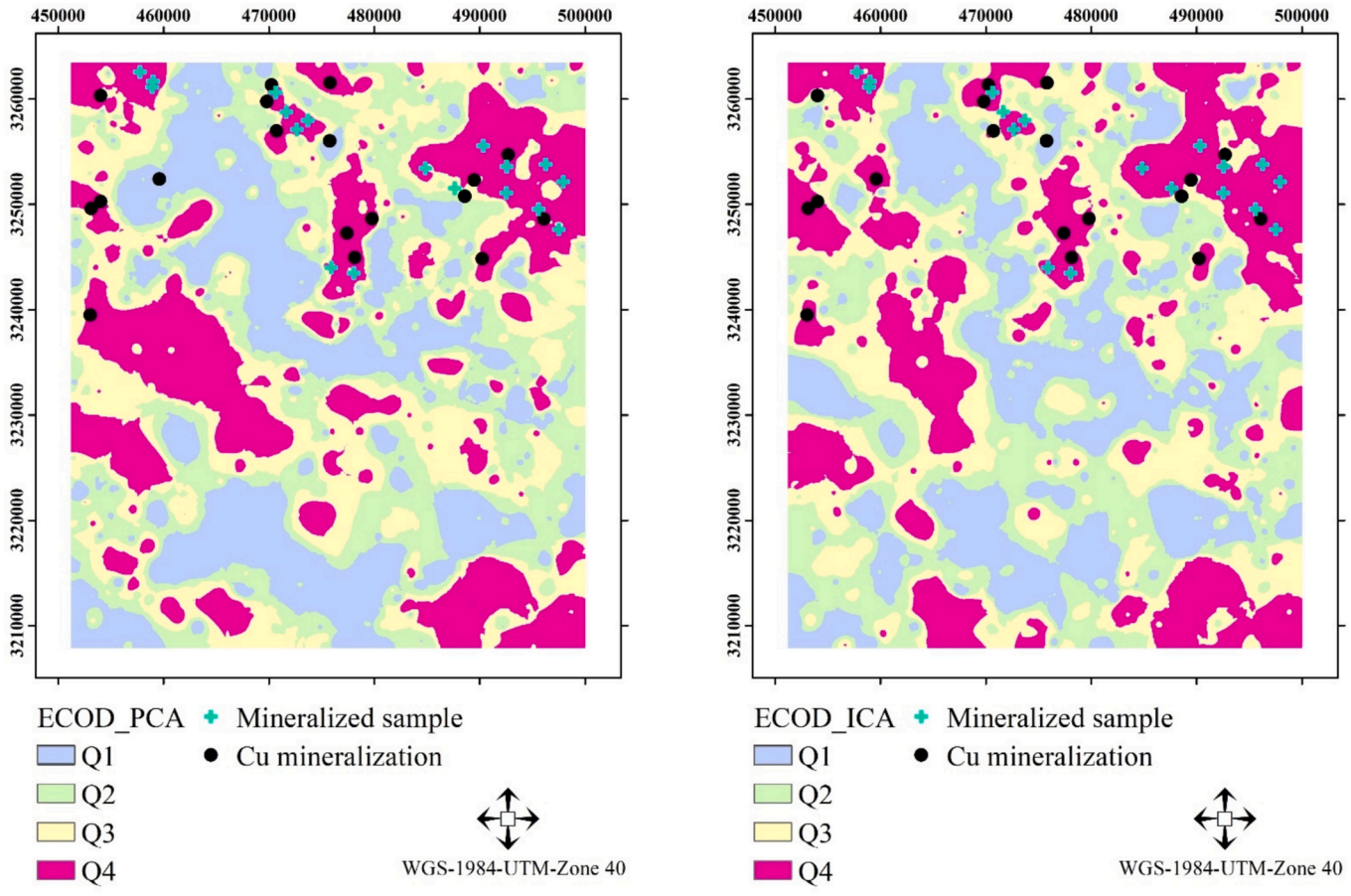

Fig. 9. Geochemical anomaly maps emerging from ECOD considering new feature space by using PCA (left) and ICA (right) dimension reduction methods.

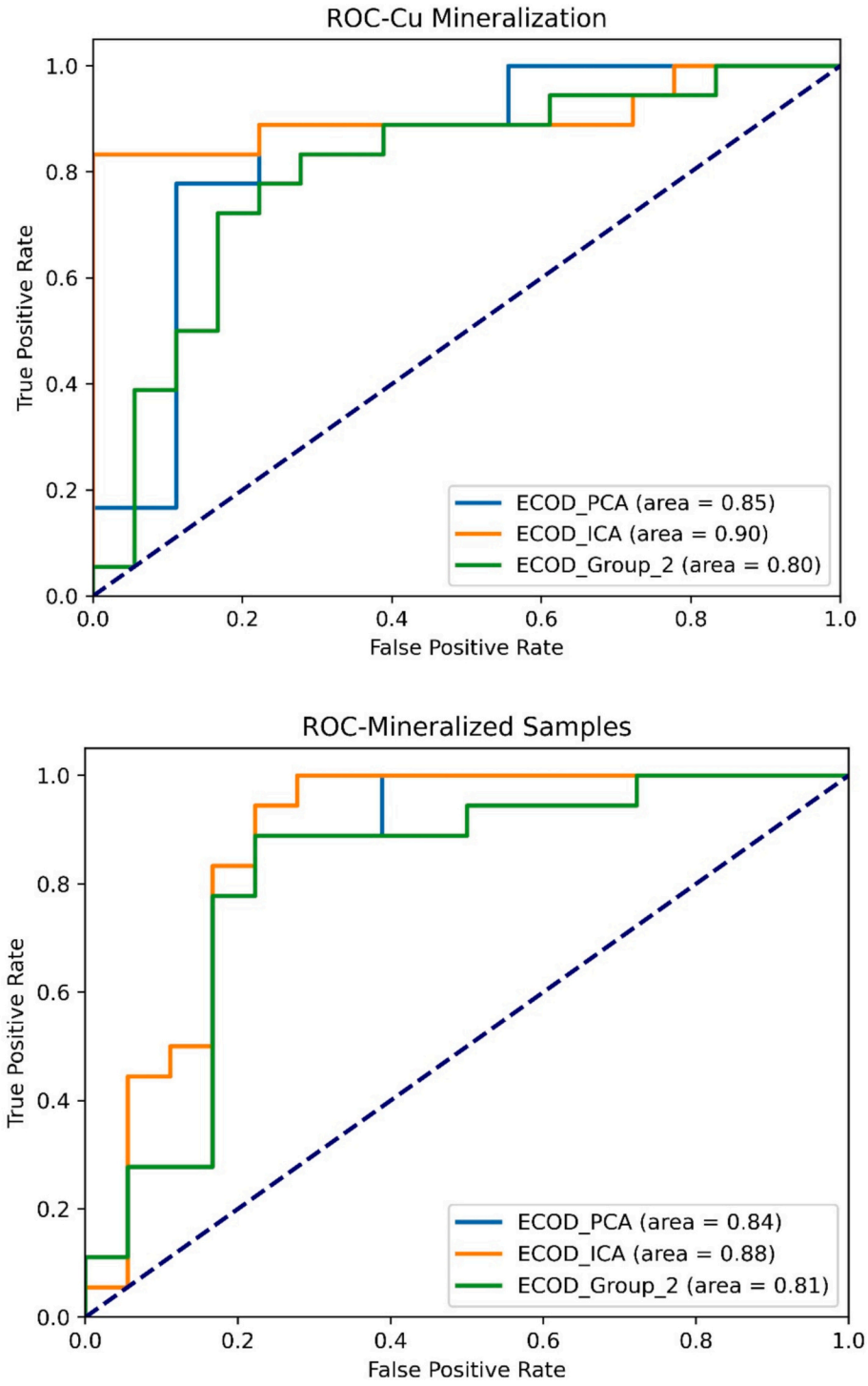

Fig. 10. ROC curves comparing the efficiency of the ECOD method applied to new ICA-based, PCA-based feature spaces and group_2 (all elements) in delineating geochemical anomalies in the study area (Top: known mineral occurrences; Bottom: mineralized samples).

condition rarely met in the context of geochemical exploration based on multivariate regional datasets. To address this, clr-transformed multivariate geochemical data were first transformed into new principal and independent feature spaces using principal component analysis (PCA) and independent component analysis (ICA) (e.g., Shahrestani et al., 2024). Based on the scree plot (not shown here), seven principal and independent components were selected as the new feature space, and ECOD was then applied to these transformed feature spaces, with the resulting geochemical maps presented in Fig. 9. As shown in Fig. 9, converting the primary feature space into a new independent feature space using ICA (ECOD_ICA) enhances the efficiency of the anomaly detection process. Though improvement is also observed with ECOD_PCA in delineating mineralized samples, the ECOD_ICA anomaly maps classify all mineralized samples along with 15 out of 18 Cu mineralization sites in the highest score class (Q4). Additionally, Fig. 10 shows a ROC curve comparing the effectiveness of ECOD applied to the original and transformed feature spaces. This figure highlights the superiority of ECOD_ICA in delineating geochemical anomalies, with both Cu mineralization sites and mineralized samples used as validation points. The effectiveness of ECOD depends on the level of correlation

between primary elemental commodities and their pathfinder elements. For instance, in this case study, Pearson correlation values between Cu and other variables are as follows: Ag (0.078), As (0.256), Au (0.050), Bi (0.078), Co (0.072), Cr (−0.544), Mo (0.095), Ni (−0.292), Pb (−0.014), Sb (0.222), and Zn (0.067). For features that exhibit high correlation, transforming dimensions into independent feature spaces can significantly enhance the efficiency of the ECOD outlier detection method.

While ECOD is designed for high-dimensional data analysis, it has limitations when applied to geochemical exploration. ECOD mainly uses univariate tail probabilities to identify anomalies, which limits its ability to detect multivariate or interaction-based anomalies. In geochemical studies, significant anomalies often come from complex relationships between different elements, where multiple elements integrate to indicate mineralization or other geological features. This means that ECOD might miss important multivariate patterns that point to valuable mineral deposits, leading to incomplete or misleading results. Another challenge is the difficulty in interpreting the anomaly scores produced by ECOD. Understanding why certain samples are flagged as outliers is crucial for decision-making, yet the way tail probabilities are combined can hide the reasons behind anomaly detection. Moreover, ECOD assumes that features are statistically independent, but in geochemical datasets, elements often show strong correlations. This can undermine the reliability of ECOD because these dependencies can hide true anomalies and increase the chances of false positives. To tackle these issues, dimension reduction techniques like PCA and ICA might be necessary. Although these methods help manage correlations and redundancies in geochemical datasets, they can also obscure subtle signals that are important for detecting meaningful anomalies. Additionally, geochemical datasets often contain noise and redundancy, which can grow in higher dimensions. This may lead ECOD to mistake this noise for true anomalies, especially when unrelated features with little geochemical importance are included. Therefore, these limitations require careful consideration and possible adjustments when applying ECOD to geochemical exploration data.

5. Conclusions

- The ECOD method yields more stable and informative anomaly scores than LOF, with a greater range and stronger correlation between scores between two elemental subsets. This suggests that ECOD provides a more effective comparison of samples and performs better across various dimensionalities. Performance of ECOD benefits from its ability to utilize both tails of the cumulative distribution function, with the right-tail being used due to positive skewness in input dimensions. Non-parametric nature of ECOD and its insensitivity to the number of dimensions contribute to its robust performance. Elements with higher skewness significantly influence ECOD anomaly scores, highlighting the importance of careful feature selection to avoid skewness-related artifacts.
- ROC curve analysis demonstrates that ECOD outperforms LOF, with ECOD_Group_2 successfully detecting 77 % of known Cu occurrences. Multivariate anomaly detection with ECOD_Group_2 significantly surpasses the univariate Cu map, which fails due to delineate known Cu mineralization probably due to high copper mobility, signal dilution, and insufficient sampling density.
- ECOD_Group_2, using 12 trace elements, surpasses other methods, including LOF, in detecting known Cu mineral occurrences. When only Cu, Au, Pb, and Zn are used, ECOD and LOF perform similarly. However, increasing dimensionality reduces the effectiveness of LOF, as evidenced by fewer detected mineral occurrences. ECOD maintains effectiveness across different trace elements due to its flexibility and robustness in handling high-dimensional data.
- The top 30 anomaly scores indicate that LOF_Group_2 lacks strong consistency with validation points, with better alignment observed in LOF_Group_1 and ECOD_Group_1. ECOD_Group_2 shows high correlation with validation points in several zones, proving effective in identifying mineralization-related anomalies.
- ECOD_Group_2 outperforms k-means clustering in detecting geochemical anomalies, as evidenced by a higher area under the ROC curve. This demonstrates the superiority of ECOD in identifying and delineating geochemical anomalies compared to k-means clustering.
- While ECOD shows promise for analyzing clr-transformed multivariate geochemical data, its assumption of statistical independence among features is often not met in geochemical exploration. The transformation of clr data into new feature spaces using PCA and ICA enhances anomaly detection; however, the reliance on univariate tail probabilities limits the effectiveness of ECOD in identifying multivariate anomalies stemming from complex inter-element relationships. Additionally, strong correlations within geochemical datasets can result in false positives, emphasizing the importance of applying dimension reduction techniques cautiously, as they may obscure meaningful signals while addressing correlation challenges.

CRediT authorship contribution statement

Shahed Shahrestani: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Ioan Sanislav:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their sincere gratitude to the Geological Survey of Iran for generously providing the regional stream sediment geochemical data from the Baft area, which played a crucial role in this study.

Data availability

The authors do not have permission to share data.

References

Aghazadeh, M., Hou, Z., Badrzadeh, Z., Zhou, L., 2015. Temporal–spatial distribution and tectonic setting of porphyry copper deposits in Iran: constraints from zircon U–Pb and molybdenite Re–Os geochronology. Ore Geol. Rev. 70, 385–406.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. *Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd., London (UK) ((Reprinted in 2003 with additional material by The Blackburn Press), London (UK). 416 p).

Aitchison, J. and J. Egozcue, J., 2005. Compositional data analysis: where are we and where should we be heading?. Math. Geol., 37, pp.829–850.

Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000, May. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104).

Chen, Y., Wu, W., 2017. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. Geochem. Explor. Environ. Anal. 17 (3), 231–238.

Dimitrijevic, M.D., 1973. Geology of Kerman region. Geol. Surv. Iran Rep. YU/52, 247.

Ding, L., Chen, B., Zhu, Y., Dong, H., Chan, G., Zhang, P., 2024. Geo-Hgan: Unsupervised anomaly detection in geochemical data via latent space learning. Comput. Geosci. 192, 105703.

Esmaeiloghli, S., Lima, A., Sadeghi, B., 2024. Lithium exploration targeting through robust variable selection and deep anomaly detection: an integrated application of sparse principal component analysis and stacked autoencoders. Geochemistry, 126111.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Ghasemzadeh, S., Maghsoudi, A., Yousefi, M., 2019. Application of geometric average approach for Cu-porphyry prospectivity mapping in the Baft area, Kerman. Sci. Q. J. Geosci. 29 (113), 231–130.

Ghasemzadeh, S., Maghsoudi, A., Yousefi, M., Mihalasky, M.J., 2022. Information value-based geochemical anomaly modeling: a statistical index to generate enhanced geochemical signatures for mineral exploration targeting. Appl. Geochem. 136, 105177.

Hajihosseinlou, M., Maghsoudi, A., Ghezelbash, R., 2024. A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins. Geochemistry 84, 126094.

Hinton, G.E., Zemel, R., 1993. Autoencoders, minimum description length and Helmholtz free energy. Adv. Neural Inf. Proces. Syst. 6.

Howarth, R.J., Thompson, M., 1976. Duplicate analysis in geochemical practice. Part II. Examination of proposed method and examples of its use. Analyst 101 (1206), 699–709.

Jamali, H., 2017. The behavior of rare-earth elements, zirconium and hafnium during magma evolution and their application in determining mineralized magmatic suites in subduction zones: constraints from the Cenozoic belts of Iran. Ore Geol. Rev. 81, 270–279.

Lazarevic, A., Kumar, V., 2005, August. Feature bagging for outlier detection. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 157–166.

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., Chen, G.H., 2022. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. IEEE Trans. Knowl. Data Eng. 35 (12), 12181–12193.

Liu, Y., Xia, Q., Duan, J., Dai, J., Wu, S., Zhao, Z., 2024. Geochemical anomalies of critical metals in the Eastern Kunlun Orogenic Belt, China: Implications for nickel and cobalt mineral exploration. Ore Geol. Rev., 106168

Moghadam, M.C., Tahmasbi, Z., Ahmadi-Khalaji, A., Santos, J.F., 2018. Petrogenesis of Rabor-Lalehzar magmatic rocks (SE Iran): Constraints from whole rock chemistry and Sr-Nd isotopes. Geochemistry 78 (1), 58–77.

Mou, N., Wang, G., Sun, X., 2023. Identification of geochemical anomalies related to mineralization: a case study from porphyry copper deposits in the Qulong-Jiama mining district of Tibet, China. J. Geochem. Explor. 244, 107126.

Naaman, M., 2021. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. Stat. Probab. Lett. 173, 109088.

Niktabar, S.M., Moradian, A., Ahmadipour, H., Santos, J.F., Mendes, M.H., 2015. Petrogenesis of the Lalezar granitoid intrusions (Kerman Province-Iran). J. Sci. I. R. Iran 26 (4), 333–348.

Pokrajac, D., Lazarevic, A., Latecki, L.J., 2007, March. Incremental local outlier detection for data streams. In: In *2007 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, pp. 504–515.

Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., Filzmoser, P., 2024. A performance study of local outlier detection methods for mineral exploration with geochemical compositional data. J. Geochem. Explor. 258, 107392.

Sepidbar, F., Ao, S., Palin, R.M., Li, Q.L., Zhang, Z., 2019. Origin, age and petrogenesis of barren (low-grade) granitoids from the Bezenjan-Bardsir magmatic complex, southeast of the Urumieh-Dokhtar magmatic belt, Iran. Ore Geol. Rev. 104, 132–147.

Shafiei, B., 2010. Lead isotope signatures of the igneous rocks and porphyry copper deposits from the Kerman Cenozoic magmatic arc (SE Iran), and their magmatic-metallogenetic implications. Ore Geol. Rev. 38 (1–2), 27–36.

Shafiei, B., Haschke, M., Shahabpour, J., 2009. Recycling of orogenic arc crust triggers porphyry Cu mineralization in Kerman Cenozoic arc rocks, southeastern Iran. Mineral. Deposita 44, 265–283.

Shahrestani, S., Carranza, E.J.M., 2024. Effectiveness of LOF, iForest, and OCSVM in Detecting Anomalies in Stream Sediment Geochemical Data. Geochemistry: Exploration, Environment, Analysis geochem 2024–009.

Shahrestani, S., Mokhtari, A.R., Carranza, E.J.M., Hosseini-Dinani, H., 2019. Comparison of efficiency of techniques for delineating uni-element anomalies from stream sediment geochemical landscapes. J. Geochem. Explor. 197, 184–198.

Shahrestani, S., Cohen, D.R., Mokhtari, A.R., 2024. A comparison of PCA and ICA in geochemical pattern recognition of soil data: the case of Cyprus. J. Geochem. Explor. 264, 107539.

Srdic, A., Dimitrijevic, M.N., Cvetic, S., Dimitrijevic, M.D., 1972. Geological Map of Baft. 1/100000 Series, Sheet 7348. Geological survey of Iran.

Yang, J., Tan, X., Rahardja, S., 2023. Outlier detection: how to Select k for k-nearest-neighbors-based outlier detectors. Pattern Recogn. Lett. 174, 112–117.

Zhou, Y., Xia, H., Yu, D., Cheng, J., Li, J., 2024. Outlier detection method based on high-density iteration. Inf. Sci. 662, 120286.