Journal of Mining and Environment (JME) Vol. 17, No. 3, 2026, 801-828

Shahrood University of Technology

Journal of Mining and Environment (JME)

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

Iranian Society of Mining Engineering (IRSME)

Comparison of Unsupervised Multivariate Clustering Methods for the Geochemical and Geospatial Characterization of Mining Tailings

Marco Antonio Cotrina-Teatino[1]*, Jairo Jhonatan Marquina-Araujo[1], Jose Nestor Mamani-Quispe[2], Solio Marino Arango-Retamozo[1], and Joe Alexis Gonzalez-Vasquez[3]

1. Department of Mining Engineering, Faculty of Engineering, National University of Trujillo, Trujillo, Peru
2. Faculty of Chemical Engineering, National University of the Altiplano of Puno, Puno, Peru
3. Department of Industrial Engineering, Faculty of Engineering, National University of Trujillo, Trujillo, Peru

Article Info

Received 25 July 2025
Received in Revised form 12 September 2025
Accepted 4 October 2025
Published online 4 October 2025

DOI: 10.22044/jme.2025.16568.3239

Keywords

Mining tailings
Multivariate clustering
Riemannian K-Means
Geochemical characterization

Abstract

The geochemical and spatial characterization of legacy mine tailings is essential for identifying reprocessing opportunities and informing environmental management. However, the high compositional complexity of polymetallic tailings requires robust multivariate approaches. This study evaluates and compares the performance of four unsupervised clustering algorithms Euclidean K-Means, Riemannian K-Means, Gaussian Mixture Model (GMM), and Agglomerative Clustering applied to 927 samples from the Quiulacocha tailings deposit in Peru, using six major elements (Zn, Pb, Cu, Fe, Ag, Au) and spatial coordinates. All methods consistently identified three main geochemical domains. Cluster 1 was enriched in Cu and Au, Cluster 2 in Pb and Fe, and Cluster 3 in Zn, Ag, and Fe. Covariance-based methods (Riemannian K-Means and Agglomerative Clustering) outperformed others in internal validation (Silhouette scores up to 0.58) and consistency (Adjusted Rand Index = 1.00), offering more interpretable and geologically coherent partitions. CLR transformation reduced clustering performance, highlighting the importance of preserving raw geochemical variance for spatial segmentation. These findings demonstrate the effectiveness of multivariate clustering for unraveling compositional heterogeneity in tailings and delineating domains of potential economic value. The approach provides a quantitative framework for supporting reprocessing decisions, reducing risk, and guiding future research on mine waste valorization.

1. Introduction

The large-scale processing of ores in the mining industry generates substantial quantities of solid waste annually, with mine tailings being among the most prominent byproducts [1]. These tailings, consisting of finely ground rock particles (D90 ≈ 100 μm) [2], result from the physicochemical treatment of ores and exhibit significant geochemical reactivity due to their high surface-area-to-volume ratio and behavior in surface environments [3]. The oxidation of sulfide minerals within the tailings promotes the redistribution of trace elements and fosters the formation of secondary minerals, such as sulfates and iron oxyhydroxides [4], progressively altering their geochemical composition over time.

In addition to their potential environmental impact, mine tailings also represent an underexploited economic resource. They may contain residual concentrations of valuable elements such as Au and Ag, alongside metals and metalloids including As, Pb, Cu, Zn, Cd, Co, and Ni [1], whose geochemical behavior is governed by factors such as redox potential, water saturation, and pH conditions [5, 6]. When present in economically significant concentrations, tailings can be reprocessed as a secondary source of metals [7–9]. For such recovery operations to be feasible

Corresponding author: mcotrinat@unitru.edu.pe (M.A. Cotrina-Teatino)

and efficient, it is essential to understand the spatial distribution and geochemical behavior of these elements within the tailings deposit [10].

However, characterizing the spatial and compositional variability of mine tailings remains a complex challenge. Georeferenced multivariate geochemical data exhibit high-dimensional structure, strong interdependencies among variables, and heterogeneous spatial patterns [11]. A key goal is to segment the deposit into homogeneous and spatially contiguous geochemical domains, which is vital for both environmental management and assessing metallurgical recovery potential. Furthermore, geochemical data derived from mine tailings are inherently compositional, reflecting relative concentrations constrained by a constant-sum limitation. Applying classical multivariate methods to such data may lead to biased or misleading interpretations [12]. Log-ratio transformations and compositional data analysis (CoDA) provide more robust alternatives for preserving the intrinsic structure of geochemical datasets [13–16]. Despite their relevance, these techniques are still rarely integrated into the spatial analysis of tailings. Their combination with clustering algorithms could enhance both the interpretability and statistical robustness of geochemical domain delineation.

Comparable challenges in spatial segmentation have been encountered in other geoscientific disciplines, such as the delineation of climate zones [17], land use areas [18], archaeological sites [19], and mineralogical typologies [20, 21]. Traditionally, geochemical analyses have relied on univariate approaches or methods based on simplistic statistical assumptions, including mean ±2σ thresholding [22], classical univariate analysis [23], conventional multivariate analysis [24, 25], or geostatistical techniques [26, 27]. Although useful, these methods are generally suited to datasets with relatively homogeneous geochemical backgrounds and assume known distributional patterns. More complex and nonlinear models, such as fractal and multifractal approaches [28–30], have also been used to represent complex geochemical systems, albeit with their own methodological limitations.

Recent research underscores the potential of integrating fractal and machine learning techniques into geoscientific characterization. Pourgholam et al. [31] applied a deep learning fractal-wavelet approach to identify iron-apatite mineralizations in Tarom, Iran. Similarly, Farhadi et al. [32] showed that ensemble models, such as StackingC, significantly improve lithological classification. Saadati et al. [33] utilized stepwise fractal modeling with ASTER data to map alteration zones, while Samadi et al. [34] applied C-V and N-S fractal models to delineate zones of effective porosity in a reservoir. In the context of mine tailings, Cotrina-Teatino et al. [35] conducted a geochemical and mineralogical characterization of critical elements in gold tailings from La Cienega, Peru, and assessed their reusability using ordinary kriging. Additionally, Cotrina-Teatino et al. [36] evaluated the strategic potential of lanthanum and scandium through unsupervised geochemical-lithological analysis in southern Ecuador.

The emergence of unsupervised learning techniques has opened new possibilities for analyzing multivariate geochemical data. Ahmed et al. [37] examined geochemical relationships using raster maps derived from LA-ICP-MS data, combined with logarithmic transformations. Nascimento et al. [3] studied mineralogy and metal mobility in altered deltaic sediments, emphasizing how natural weathering affects reprocessing potential. Wang and Chen [38] demonstrated the utility of a DAGMM model, based on deep autoencoders, in detecting complex geochemical anomalies.

In parallel, numerous clustering-based approaches have been applied to geochemical analysis. Zhou and Maerz [39] incorporated orientation and spacing into cluster criteria. Stumpe and Marschner [40] combined K-Means and Random Forest to evaluate the thermal properties of tailings piles. Santos et al. [41] used multivariate techniques to analyze contaminant dispersion in Pb-Zn tailings. Jin et al. [42] proposed grouped factor analysis to overcome traditional factor analysis limitations, enabling the detection of spatial dependency patterns. Baragilly et al. [43] developed a non-parametric clustering method based on spatial rank functions. Xiao et al. [44] applied spatial clustering using self-organizing maps (SOM, Geo-SOM) to model geological risks.

Interest continues to grow in methods that explicitly incorporate spatial information and multivariate dependency structures. Fouedjio [45] proposed a hierarchical method tailored to multivariate geostatistical datasets. Riquelme and Ortiz [46] employed Riemannian geometry to delineate complex spatial domains. Cotrina-Teatino et al. [47] used a Riemannian K-Means model for classifying mineral resources in a copper deposit in Peru. Martin and Boisvert [48] integrated multivariate compactness metrics with spatial contiguity constraints in their clustering algorithm. Templ et al. [49] addressed the challenges of clustering geochemical data with non-normal or

multimodal distributions. Hajihosseinlou et al. [50] compared OPTICS, GMM, and K-Means for anomaly detection. Sadeghi et al. [51] and Jansson et al. [52] combined PCA and K-Means to identify prospective mineralization zones. Morales et al. [53] applied Genetic K-Means to soil gas data in geothermal studies.

Other advances include Bayesian hierarchical clustering techniques [54], MCACA-based pattern recognition [55], clustering for geological domain definition [56], geodiversity assessment [57], and advanced deep clustering methods [58]. Tokuda et al. [59] and Li et al. [60] explored hierarchical variants and ensemble clustering strategies. Marquina-Araujo et al. [61] integrated autoencoders with K-Means to define geostatistical domains, while Martin and Boisvert [62] applied spatial clustering to evaluate stationarity decisions in geostatistics. Moreira et al. [63] combined machine learning and geostatistics for geological domain modeling.

This review highlights that, despite significant advances in the application of multivariate clustering techniques to geochemical data, there remains a notable lack of systematic and rigorous comparisons among different clustering methods particularly those incorporating Riemannian geometry in the specific context of the geochemical and spatial characterization of mine tailings deposits. Furthermore, few studies explicitly address the consistency between methods or the robustness of the resulting segmentations, both of which are critical for practical implementation.

To address this gap, the present study carries out a comprehensive comparison of unsupervised multivariate clustering methods, including Euclidean K-Means, Riemannian K-Means, Gaussian Mixture Model (GMM), and Agglomerative Clustering. These methods are applied to the geochemical and geospatial characterization of a mine tailings deposit, using both raw geochemical data and data transformed via the centered log-ratio (CLR) technique. This dual approach enables the evaluation of clustering performance under both conventional and compositional data structures.

The structure of the manuscript is as follows: first, the geological context and geographical location of the study area are presented; second, the methodology and data processing are described; next, the clustering results are presented and analyzed; and finally, the conclusions and future perspectives are discussed.

2. Geology and geographical location

This study is based on a geochemical and geospatial exploration campaign conducted at the Quiulacocha tailings deposit, located in the Pasco Department of central Peru (Figure 1). The site lies at over 4,300 meters above sea level in the Eastern Cordillera of the central Peruvian Andes, within the Cerro de Pasco mining district one of the most significant polymetallic districts in the Central Andes.

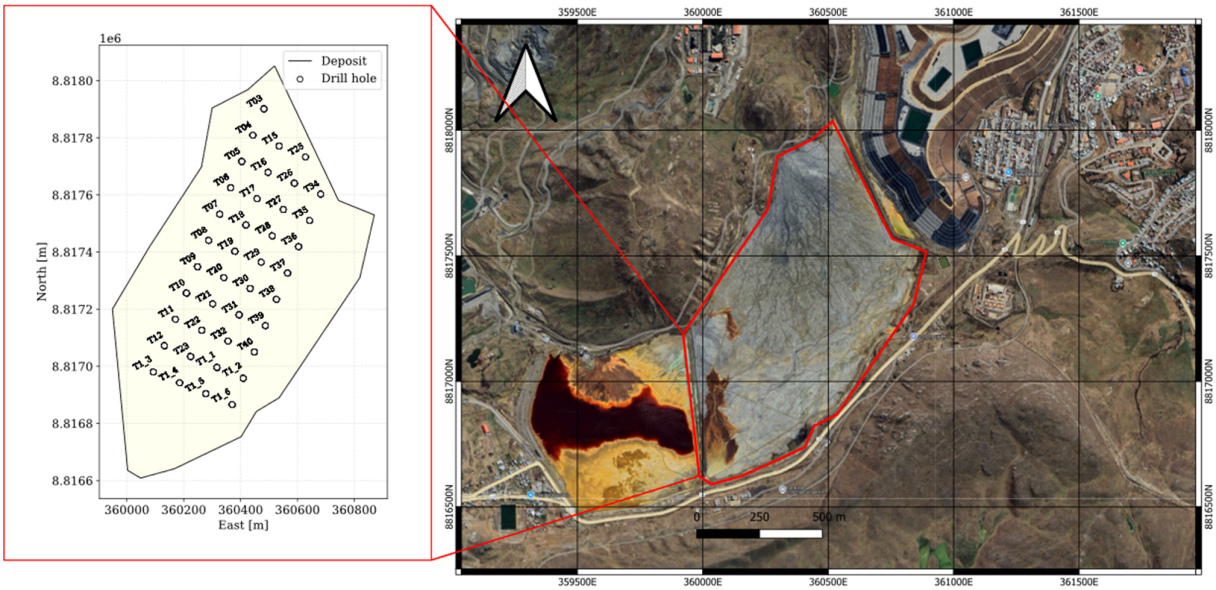

Figure 1. Location of the Quiulacocha tailings deposit in Peru and the sampling drill holes conducted.

The regional geology of the Cerro de Pasco area is defined by a stratigraphic succession composed primarily of sedimentary rocks from the Pucará Group (Upper Triassic to Lower Jurassic limestones), overlain by volcanic and intrusive units emplaced during the late Oligocene to early Miocene. The magmatic-hydrothermal evolution of the region is associated with the intrusion of dacitic domes and porphyritic bodies, which led to the development of Cordilleran-type epithermal systems.

The Cerro de Pasco deposit is classified as a polymetallic epithermal system genetically related to porphyry intrusions. Mineralization occurs in veins, replacement bodies, and stratiform mantos, and is mainly composed of pyrite, sphalerite, galena, and enargite. These assemblages are commonly enriched with silver and copper, and are locally associated with trace amounts of bismuth, arsenic, and antimony. The metallogenic zonation reflects a proximal-to-distal gradient relative to the intrusive centers, with copper-rich zones near the cores transitioning outward into domains enriched in lead, zinc, and silver (see Figure 2).

The Quiulacocha tailings deposit hosts approximately 70 million tonnes of historical mine waste spread over an area of 115 hectares. Tailings were generated by the processing of ores extracted from both open-pit and underground operations at the Cerro de Pasco mine, starting in the 17th century. Continuous tailings deposition at Quiulacocha began in the early 20th century, initially from copper, silver, and gold extraction processes with reported grades of up to 10% Cu, 4 g/t Au, and over 300 g/t Ag and later from zinc, lead, and silver ore bodies, with average grades of 7.41% Zn, 2.77% Pb, and 90.33 g/t Ag. A more recent estimate (2012) by BO Consulting reported resources of approximately 2.9 Mt with average grades of 1.43% Zn, 0.79% Pb, 43.1 g/t Ag, and 0.04% Cu.

The diversity in mineralogical and geochemical composition within the tailings reflects the complex metallurgical history of the site, making it a relevant case study for the application of unsupervised clustering methods to identify distinct geochemical domains for both environmental assessment and potential reprocessing.

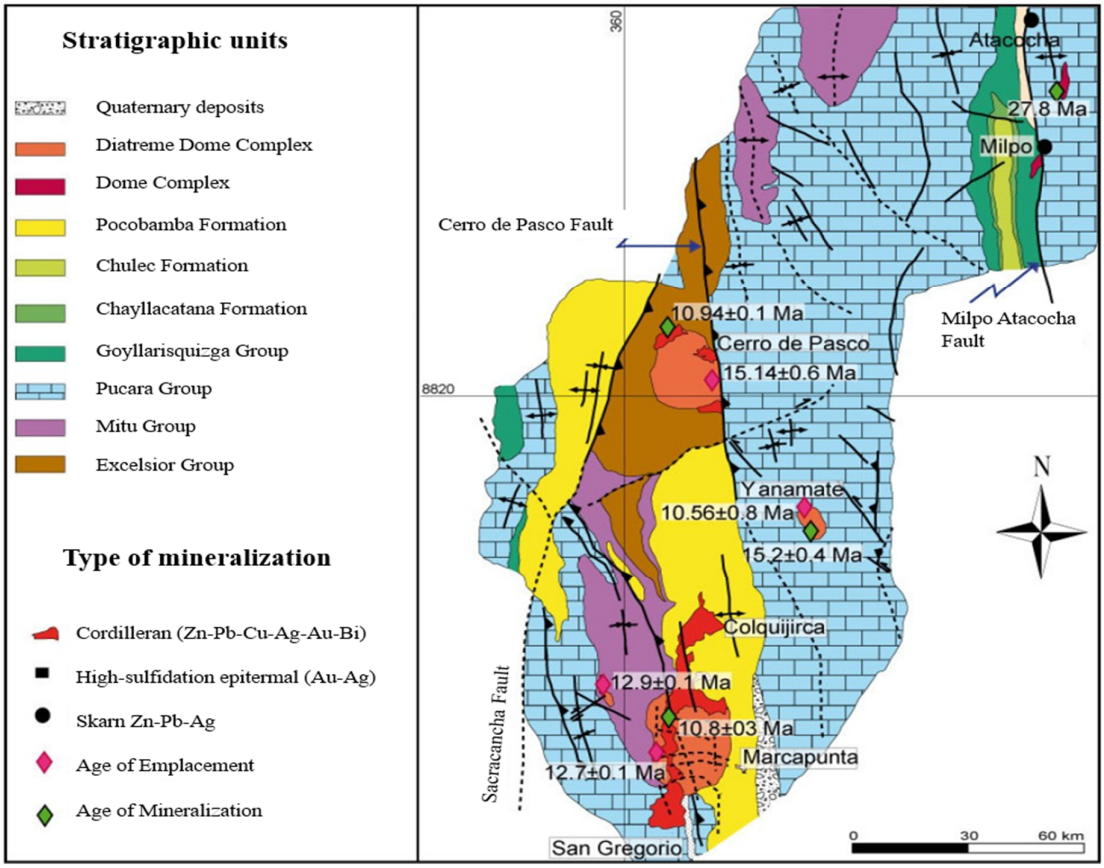

Figure 2. Geological and metallogenic map of the Cerro de Pasco region, showing major stratigraphic units, fault systems, and types of mineralization in the principal deposits [64].

3. Methodology

3.1. Overview of the Approach

A comparative analysis of unsupervised clustering methods was conducted to support the geochemical and geospatial characterization of mine tailings. Four clustering algorithms were implemented: Euclidean K-Means [52, 61], Riemannian K-Means [46, 47, 65, 66], Gaussian Mixture Models (GMM) [50], and Agglomerative Clustering [67]. All algorithms were applied to covariance matrix-based data representations to capture both the interdependence between geochemical variables and their spatial variability.

The input dataset, comprising both elemental concentrations and spatial coordinates, underwent a preprocessing stage to ensure quality and consistency. Clustering performance was assessed using internal validation indices and visual interpretation techniques, allowing for a comprehensive evaluation of segmentation quality and robustness.

3.2. Database Description

The geochemical dataset was obtained from 40 vertical drill holes distributed over the Quiulacocha tailings deposit in a regular grid pattern with an approximate spacing of 100 meters. A total of 927 samples were collected at 1-meter intervals, providing high-resolution vertical coverage throughout the deposit. The study area spans approximately 589 meters in the east-west direction and 1,035 meters in the north-south direction, with drilling depths reaching up to 36 meters.

Table 1 presents the descriptive statistics for the six primary geochemical variables analyzed: silver (Ag), zinc (Zn), lead (Pb), copper (Cu), gold (Au), and iron (Fe). The dataset reveals substantial compositional heterogeneity, with highly skewed distributions and elevated kurtosis values for several elements indicative of heterogeneous geochemical processes and enrichment patterns within the deposit.

Table 1. Descriptive statistics of the six main geochemical variables (Ag, Zn, Pb, Cu, Au, Fe) from 927 drill-hole samples collected at the Quiulacocha tailings deposit.

Statistic	Ag (g/t)	Zn (%)	Pb (%)	Cu (%)	Au (g/t)	Fe (%)
Mean	51.62	1.49	0.88	0.09	0.10	25.77
Std. Dev.	12.63	0.67	0.47	0.07	0.16	5.60
Variance	159.64	0.45	0.22	0.00	0.03	31.36
Minimum	4.75	0.03	0.08	0.01	0.01	13.78
Median	49.14	1.38	0.73	0.07	0.05	26.94
Maximum	168.00	7.90	3.41	0.62	1.34	45.90
IQR	11.51	0.51	0.44	0.06	0.05	4.63
Skewness	2.88	3.48	1.86	2.43	4.68	-0.58
Kurtosis	17.72	23.91	4.18	9.46	26.61	-0.26

3.3. Data Processing

Since the variables included in the multivariate analysis are expressed in different units grams per tonne (g/t) for Ag and Au, and percentage (%) for Zn, Pb, Cu, and Fe a standardization procedure was necessary prior to the application of distance-based and correlation-sensitive clustering algorithms. To this end, z-score normalization was applied, transforming each variable to have zero mean and unit standard deviation [68]. This step ensures that all variables contribute equally to the multivariate analysis by mitigating the influence of differences in scale, while preserving the relative structure among observations.

To properly address the compositional nature of the geochemical data, two separate versions of the dataset were used for the clustering experiments. The first consisted of the raw, untransformed geochemical values, which maintain the original measurement units and facilitate interpretability. The second dataset was transformed using the centered log-ratio (CLR) method, which is commonly employed in compositional data analysis (CoDA) to eliminate spurious correlations and to satisfy the mathematical assumptions of Euclidean geometry. This dual approach enables a robust evaluation of clustering performance under both conventional and compositionally appropriate data structures.

All geochemical analyses were performed by an accredited laboratory using standardized procedures for sample preparation and elemental quantification. Analytical accuracy and precision were ensured through the use of certified reference materials, calibration standards, and replicate measurements.

3.4. Feature Representation and Feature Space

The multivariate analysis was based on the combined representation of geochemical and

spatial variables within a feature space designed to capture both their individual variability and inter-variable relationships. As an initial step, correlation matrices were computed to examine dependencies among variables and to guide a more robust understanding of the dataset's underlying structure [69]. The correlation matrix $\mathbb{R}$ is defined as:

$$R_{ij} = \frac{Cov(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}, \quad (1)$$

where $Cov(X_i, X_j)$ is the covariance between variables X_i and X_j, and $\sigma_{X_i}\sigma_{X_j}$ are their respective standard deviations. This formulation enables the identification of highly correlated variable groups, as well as potential informational redundancies that may influence clustering outcomes.

For distance-based algorithms such as Euclidean K-Means and Agglomerative Clustering, the classical Euclidean distance metric was employed:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}, \quad (2)$$

which is appropriate under the assumption of independence or low correlation among dimensions a condition reasonably satisfied following normalization. However, given the presence of residual correlations and the potential for complex geometric structure among the variables, the analysis was extended into feature spaces defined by covariance matrices. This motivated the use of clustering approaches that account for the geometry of data dispersion, such as Riemannian K-Means and GMM [70].

Covariance matrices encode the multivariate dispersion structure of the samples, allowing each cluster to be represented not as a sphere (as assumed in Euclidean K-Means), but as an oriented ellipsoid in feature space. The distance between two covariance matrices C_1 and C_2 in the Riemannian space is defined using the affine-invariant Riemannian metric [71]:

$$d_R(C_1, C_2) = \left(\sum_{i=1}^{p} log^2 \lambda_i \right)^{1/2}, \quad (3)$$

where λ_i are the eigenvalues of $C_1^{-1}C_2$, and p is the number of variables. This metric is particularly suited for clustering contexts where the shape and orientation of the data dispersion carry meaningful information such as in complex geochemical environments with multiscale variability.

3.5. Clustering Methods

For the multivariate classification of geochemical and spatial data from the Quiulacocha tailings deposit, four unsupervised clustering methods were implemented: Euclidean K-Means, Riemannian K-Means, GMM, and Agglomerative Clustering [61, 72-80]. Correlation and covariance matrices were used as core representations to capture both the individual variability of the variables and their internal dependencies an essential consideration in the analysis of complex geochemical systems.

The classical K-Means algorithm partitions observations into k groups by minimizing the within-cluster variance [61, 81, 82]. Formally, the objective function is:

$$\min_{\{S_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \quad (4)$$

where S_i denotes the set of observations assigned to cluster i, and μ_i is the centroid of that cluster. In this study, distances were measured using the standard Euclidean norm in the multivariate feature space. While K-Means is efficient and interpretable, its assumption of spherical clusters with similar sizes may limit its effectiveness in environments with anisotropic or heterogeneous structures, such as those found in mine tailings.

To incorporate the internal dependency structure of the variables and leverage local covariance matrices, a variant of K-Means operating in Riemannian space was adopted [46, 83, 84]. In this framework, each observation is represented by a local covariance matrix, and clusters are formed by minimizing Riemannian distances between these matrices. The objective function becomes:

$$\min_{\{C_i\}_{i=1}^{k}} \sum_{i=1}^{k} \sum_{x_j \in S_i} d_R^2(C_j, C_i), \quad (5)$$

where C_j is the covariance matrix is associated with observation x_j, C_i is the centroid of cluster i and $d_R(.,.)$ is a suitable Riemannian distance function typically the affine-invariant Riemannian metric. It is important to clarify that, in this framework, the cluster centroid C_i is not computed via arithmetic averaging. Instead, it corresponds to the Fréchet mean (also known as the Riemannian barycenter), which minimizes the sum of squared distances to all matrices within the cluster in Riemannian space.

Covariance matrices C_j were estimated locally using a fixed-size spatial neighborhood around each sample x_j, defined by its geochemical and spatial coordinates. Within each neighborhood, empirical covariance matrices were computed from the subset of samples falling within a predefined Euclidean radius, ensuring that the matrices are positive-definite and reflect localized structural variability. This local estimation strategy provides a more faithful representation of heterogeneity within the tailings deposit and enables clustering algorithms to exploit both local dispersion patterns and global geometric structure.

In Riemannian space, the centroid C_i is not obtained by arithmetic averaging, but instead corresponds to the Fréchet mean (also called the Riemannian barycenter), which minimizes the sum of squared distances to all matrices in the cluster:

$$C_i = \underset{C \in P_n}{arg\,min} \sum_{C_j \in S_i} d_R^2(C_j, C), \quad (6)$$

where P_n denotes the space of $n\ x\ n$ symmetric positive-definite matrices. This minimization is typically solved via an iterative gradient-based algorithm, as no closed-form solution exists in general. This definition ensures that the centroid respects the intrinsic geometry of the SPD manifold and enables the modeling of clusters with anisotropic and heterogeneous structures, which are common in tailings deposits.

is a suitable Riemannian metric, such as the affine-invariant metric. This approach enables the modeling of clusters with differentiated and anisotropic internal structures, as commonly encountered in tailings deposits with heterogeneous depositional histories.

The GMM approach models each cluster as a multivariate Gaussian distribution, allowing for flexible representations of clusters with arbitrary shapes [50, 85, 86]. The joint density function is given by:

$$p(x) = \sum_{i=1}^{k} \pi_i N(x \mid \mu_i, \Sigma_i), \quad (7)$$

where π_i are the mixture weights $(\Sigma_i \pi_i = 1)$, μ_i is the mean vector, and Σ_i is the covariance matrix of cluster i. Parameter estimation is performed via the Expectation-Maximization (EM) algorithm, which maximizes the log-likelihood:

$$\mathcal{L} = \sum_{j=1}^{n} log\left(\sum_{i=1}^{k} \pi_i N(x_j|\mu_i, \Sigma_i)\right) \quad (8)$$

GMM is particularly suitable in this context, as it allows clusters of arbitrary shapes to be modeled and can capture gradual transitions between different geochemical domains within the deposit.

Agglomerative (or hierarchical) clustering builds a dendrogram by successively merging the most similar pairs of clusters based on a predefined linkage criterion [59, 60, 67]. In this study, Ward's linkage criterion was used, which minimizes the increase in intra-cluster variance after each merge:

$$\Delta E_{AB} = \frac{n_A n_B}{n_A + n_B} \|\mu_A - \mu_B\|^2 \quad (9)$$

where n_A and n_B are the cardinalities of clusters A and B, and μ_A, μ_B are their respective centroids. This method is particularly suitable for revealing hierarchical relationships and multi-scale structures in multivariate data.

3.6. Validation and Comparison of Methods

To objectively assess the quality of the clusterings produced by the different algorithms, three complementary internal validation metrics were employed: the Silhouette Index (SI), the Davies–Bouldin Index (DBI), and the Calinski–Harabasz Index (CHI). These metrics quantify two key aspects of clustering performance intra-cluster cohesion and inter-cluster separation which are critical for the multivariate geochemical and spatial characterization of tailings deposits.

The SI evaluates clustering quality at the level of individual observations by integrating cohesion (the proximity of a point to others in the same cluster) and separation (its distance from the nearest neighboring cluster) [79, 87, 88]. The silhouette score $s(i)$ for each observation i is computed as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \quad (10)$$

where $a(i)$ is the average distance between i and other points in its own cluster, and $b(i)$ is the average distance between i and points in the nearest neighboring cluster. The value of $s(i)$ ranges from -1 to 1; values close to 1 indicate well-defined clusters, while values near 0 or negative suggest overlap or misclassification. In the geochemical context, the SI provides a useful validation that geochemical and spatial populations are coherently separated an essential aspect for identifying zones with distinct compositional characteristics.

The Global Silhouette Index (GSI) is then computed as the mean of all individual silhouette scores across the dataset:

$$GSI = \frac{1}{n}\sum_{i=1}^{n} s(i), \quad (11)$$

where n is the total number of observations. In the context of geochemical clustering, the GSI serves as a global measure of how coherently the samples are grouped in terms of both compositional and spatial similarity. It is particularly useful for comparing the quality of different clustering solutions or determining the optimal number of clusters.

The DBI offers a global evaluation of clustering quality by measuring both cluster compactness and separation [76]. For a clustering solution with k clusters, it is defined as:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max\, j \neq i \left(\frac{s_i + s_j}{d_{ij}}\right), \quad (12)$$

where s_i is the average intra-cluster distance for cluster i (compactness), and d_{ij} is the distance between the centroids of clusters i and j (separation). Lower DBI values indicate better-defined clusterings. In the context of tailings deposits, DBI helps validate whether the groupings correspond to well-differentiated geochemical and spatial domains, minimizing overlap between populations.

The CHI, also known as the variance ratio criterion, quantifies the ratio of between-cluster dispersion to within-cluster dispersion [89]. It is calculated as:

$$CHI = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)}, \quad (13)$$

where $Tr(B_k)$ is the trace of the between-cluster dispersion matrix, $Tr(W_k)$ is the trace of the within-cluster dispersion matrix, k is the number of clusters, and n is the total number of observations. Higher CHI values indicate compact and well-separated clusterings. In multivariate geochemical analysis, this metric is particularly useful for optimizing clustering algorithm configurations and for selecting the appropriate number of clusters to reflect the true underlying structure of the data.

4. Results and discussion

Before applying multivariate clustering algorithms, a comprehensive exploratory analysis was carried out on the geochemical and spatial dataset from the Quiulacocha tailings deposit. This step aimed to uncover preliminary trends, correlations, and spatial patterns to better understand the internal structure and compositional variability of the system.

4.1. Multivariate Exploratory Analysis

Three-dimensional concentration maps (Figure 3) reveal pronounced spatial heterogeneity across the deposit. A clear vertical stratification is observed for Cu and Au, with elevated concentrations concentrated in the deeper horizons of the tailings. This trend is consistent with the initial deposition phases, which were likely dominated by residues from the processing of high-sulfidation epithermal ores, particularly enargite- and chalcopyrite-rich sulfide assemblages [9]. These early metallurgical residues are known to retain substantial amounts of Cu and Au, especially in unoxidized or poorly leached zones. In contrast, Zn and in Fe display more laterally variable distributions, with less pronounced vertical zonation. These patterns may be linked to polymetallic sulfide residues, especially from sphalerite- and galena-dominant ore bodies, which were more commonly processed in later mining phases. The relatively homogeneous distribution of Fe may reflect its occurrence in both primary sulfides (e.g., pyrite) and secondary iron oxides formed during tailings oxidation and weathering [4]. Ag and Pb exhibit more irregular and patchy distributions, with localized enrichment zones. These patterns suggest potential superposition of multiple mineralization styles, including secondary enrichment processes or selective deposition from heterogeneous metallurgical feeds, as has been observed in other polymetallic tailings contexts [1].

Bivariate scatter plots (Figure 4) provide insights into inter-element relationships. Notable positive correlations are observed between Zn and Pb, as well as Zn and Fe, consistent with the known paragenesis of sphalerite-galena and iron-bearing sulfides [41, 48]. Meanwhile, the relationship between Cu and Au is characterized by low background levels for most samples, but with a subset of data showing co-enrichment, which supports the hypothesis of a Cu–Au-rich sulfide domain derived from high-sulfidation epithermal systems [3, 54]. These geochemical associations highlight the heterogeneous nature of the deposit, suggesting the coexistence of distinct metallogenic contributions preserved in the tailings.

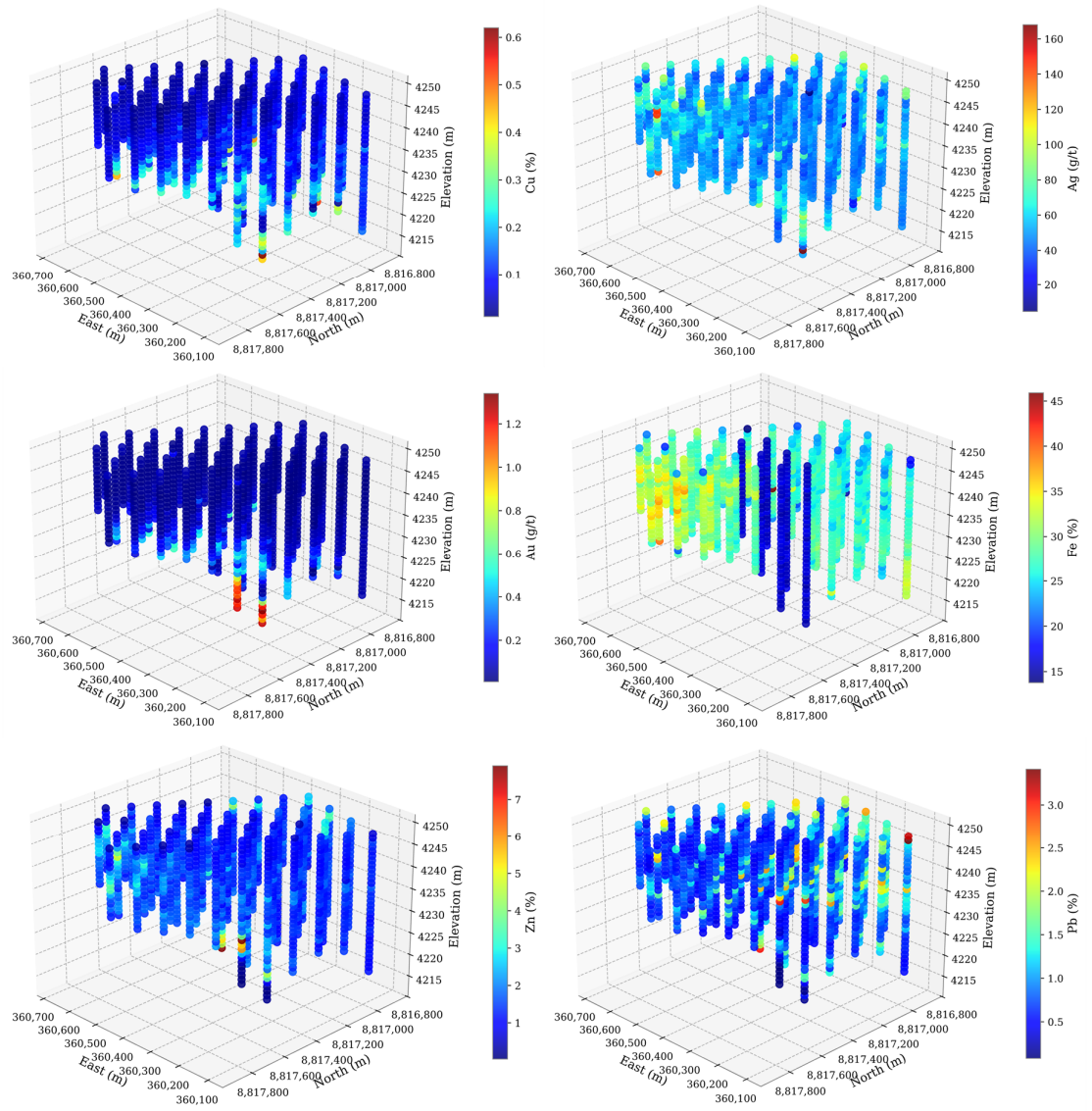

Figure 3. Three-dimensional spatial distribution of eight geochemical variables (Cu, Ag, Au, Fe, Zn, Pb) in the Quiulacocha tailings deposit. The plots show both vertical (elevation) and horizontal (East and North) distributions of elemental concentrations.

The bivariate density plots of normalized data (Figure 5) allow a more detailed examination of distribution patterns. Variables such as Ag, Cu, and Au exhibit strong positive skewness, indicating the presence of minor subpopulations with anomalously high values. Meanwhile, strong density clusters in the Zn–Pb and Zn–Ag spaces suggest the presence of at least two major geochemical groupings likely reflecting residues from the processing of (i) Cu–Au-rich high-sulfidation ores, and (ii) Zn–Pb–Ag polymetallic ores. The dispersion of Fe in these plots is more widespread and poorly correlated with other metals, reflecting its dual presence as both a matrix component and a product of secondary processes such as oxidation. Altogether, the exploratory analysis confirms that the Quiulacocha tailings deposit exhibits high spatial and compositional complexity, shaped by the sequential deposition of tailings derived from multiple ore types. These insights strongly justify the application of multivariate clustering to objectively delineate geochemically distinct domains within the deposit.

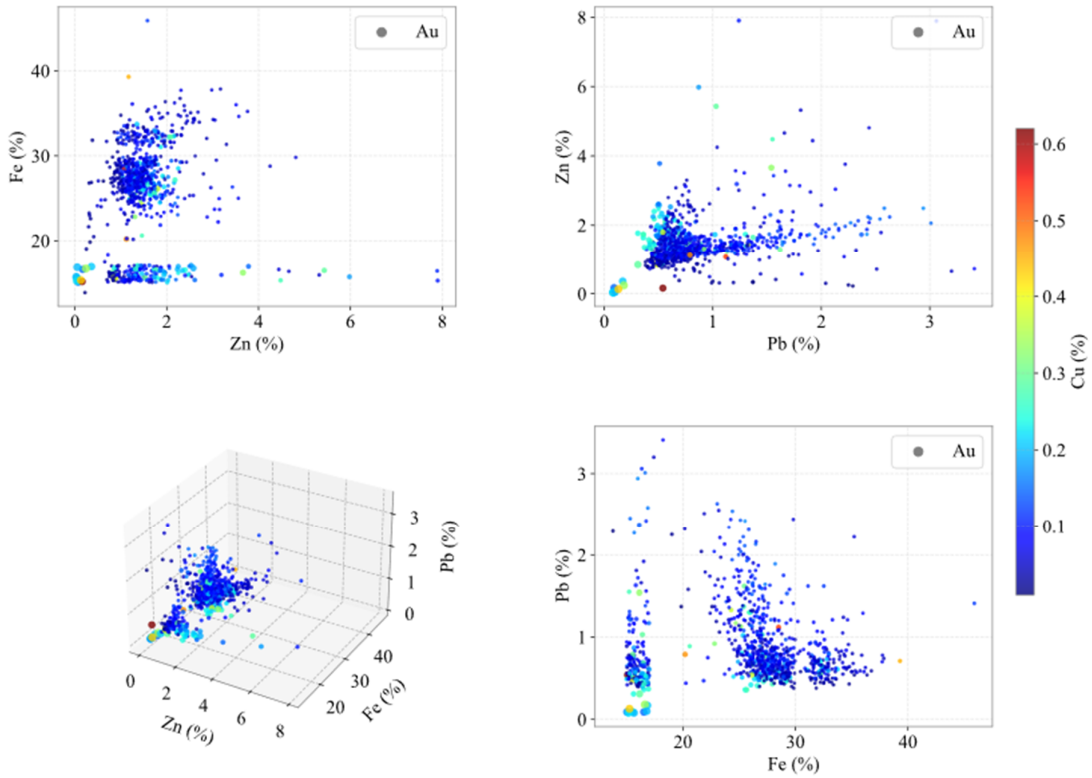

Figure 4. Bivariate visualization of geochemical data from the Quiulacocha tailings deposit. Combinations of Fe, Zn, Pb and Cu are plotted on the axes, with color coding for Cu and marker size proportional to Au content.

4.2. Multivariate Clustering (Raw data)

The multivariate clustering algorithms were applied to the geochemical and spatial dataset from the Quiulacocha tailings deposit with the aim of identifying latent compositional structures and delineating distinct mineralized domains within the deposit. The classification was performed using four algorithms: Euclidean K-Means, Riemannian K-Means, GMM, and Agglomerative Clustering. The optimal number of clusters, k = 3, was selected based on the joint evaluation of three internal validation metrics: the SI, the CHI, and the DBI, which are widely used in spatial geochemical analysis to quantify intra-cluster cohesion and inter-cluster separation [48, 49]. The resulting segmentation, shown in Figure 6, reveals consistent spatial patterns across the algorithms.

The three clusters obtained correspond to geochemically and mineralogically distinct zones within the deposit, reflecting different mineralization styles and processing histories. Cluster 1, predominantly found at intermediate elevations, is characterized by elevated concentrations of Zn, Pb, and Ag, suggesting its association with the residues of polymetallic sulfide ore processing, particularly those containing sphalerite (ZnS), galena (PbS), and argentiferous tetrahedrite ($Cu_{12}Sb_4S_{13}$). This mineral assemblage is consistent with the exploitation of carbonate-hosted Pb–Zn–Ag mineralization that typifies the later stages of mining at Quiulacocha [9].

In contrast, Cluster 2 dominates the uppermost portions of the tailings and exhibits high Fe contents with low trace element concentrations, indicating a relatively homogeneous and oxidized domain. This geochemical signature likely corresponds to a weathered residual matrix formed by post-depositional leaching of sulfides and subsequent iron oxide precipitation (e.g., goethite, hematite), a process commonly observed in exposed tailings under oxidizing conditions [4]. The geochemical homogeneity and low metallic content of this cluster support its interpretation as a leached cap over the more metal-rich zones below.

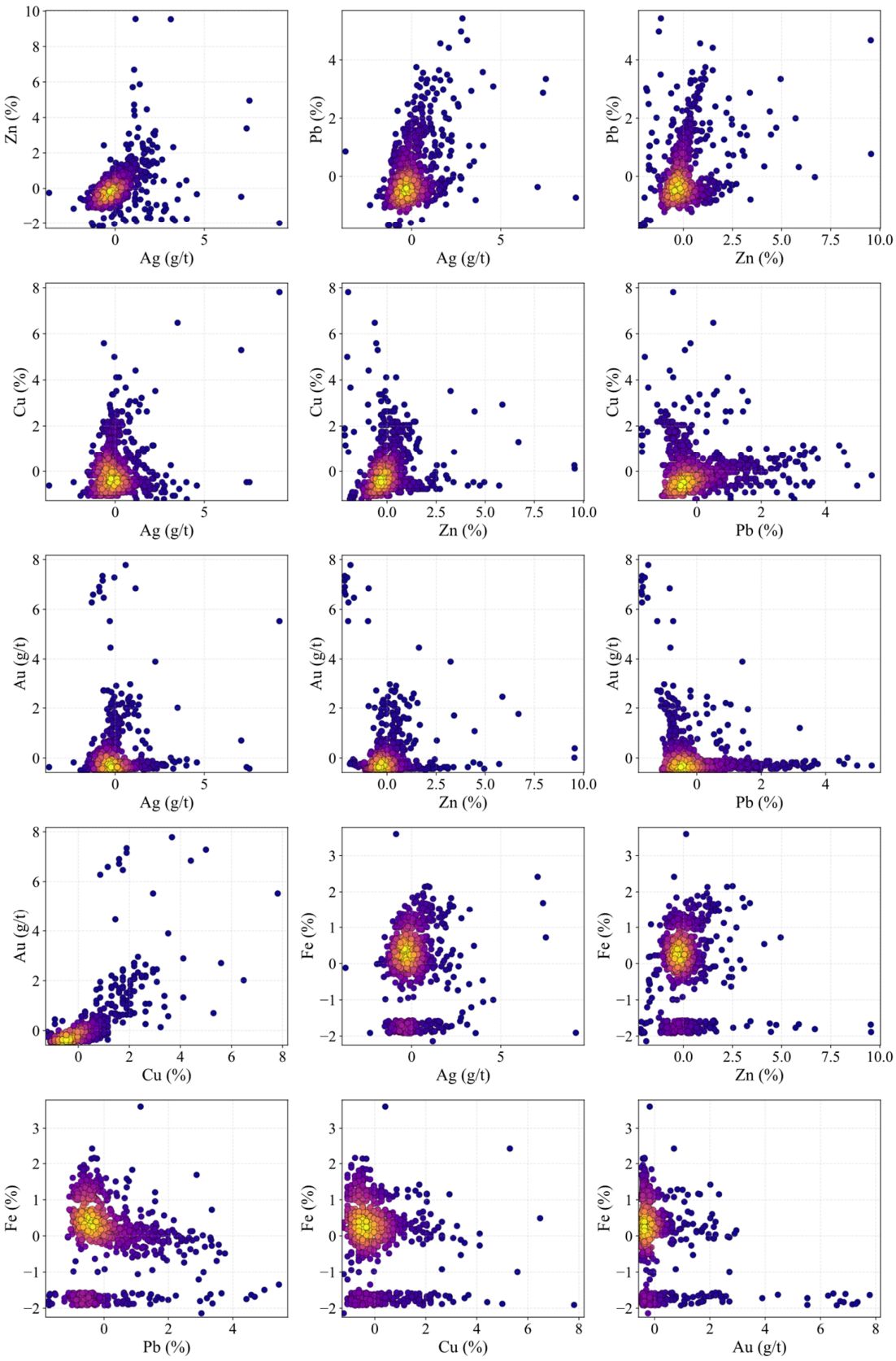

Figure 5. Bivariate density plots for the normalized geochemical variables from the Quiulacocha tailings deposit.

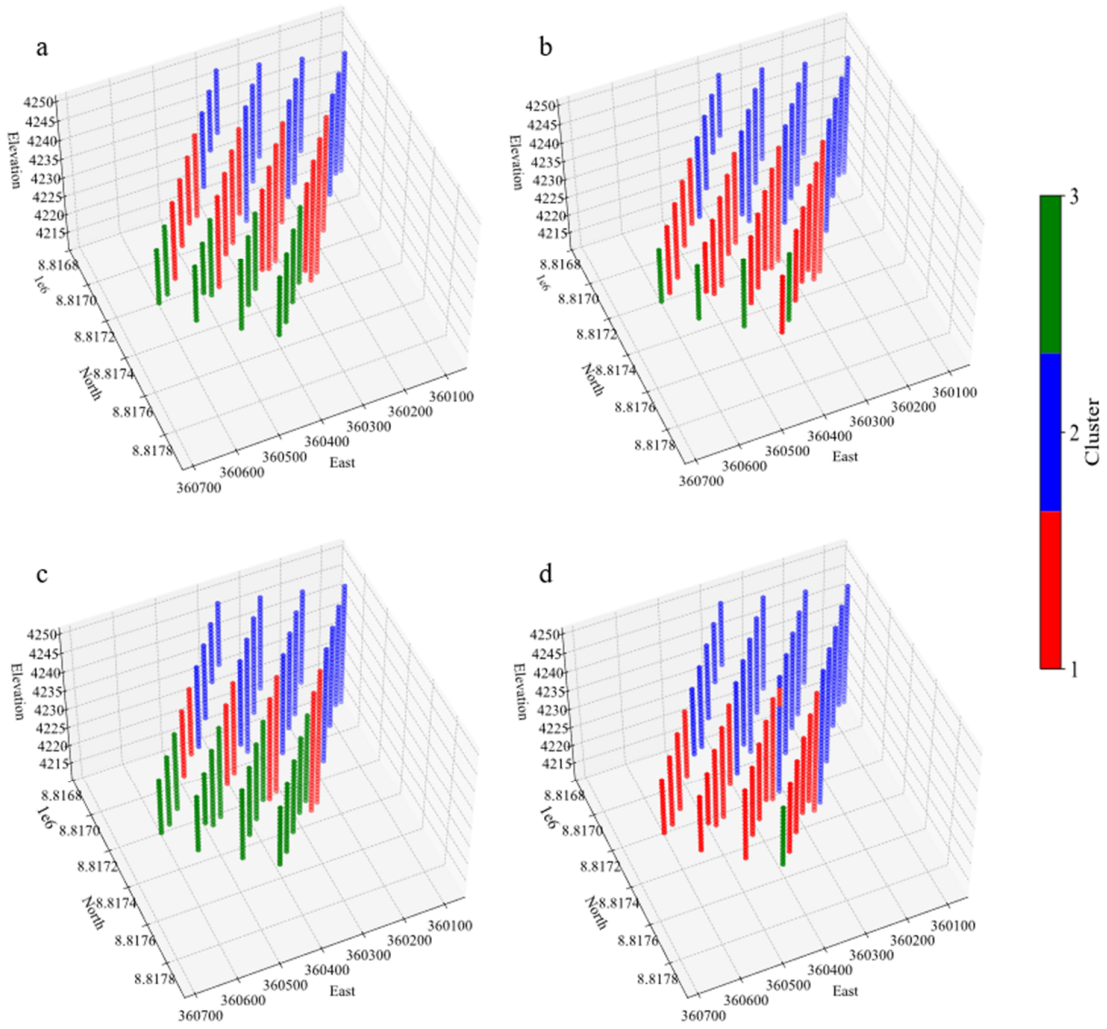

Figure 6. Multivariate classification of the Quiulacocha tailings deposit into 3 clusters: (a) Euclidean K-Means, (b) Riemannian K-Means, (c) Gaussian Mixture Model, (d) Agglomerative Clustering.

Finally, Cluster 3, concentrated in the deepest sections of the deposit, displays enrichment in Cu and Au, which are strongly indicative of tailings derived from the processing of high-sulfidation epithermal mineralization. These early deposition phases were characterized by the flotation of minerals such as enargite (Cu_3AsS_4), pyrite (FeS_2), and minor chalcopyrite ($CuFeS_2$), yielding residues with high Cu–Au content. This is consistent with historical records indicating that the initial phases of mining at Quiulacocha targeted Cu–Au-rich ores [9]. The vertical zonation of clusters thus reflects the metallurgical sequence and mineralization types historically processed at the site.

The validity of the $k = 3$ configuration is quantitatively supported by the internal validation metrics summarized in Table 2. For the Riemannian K-Means algorithm, the SI reaches 0.58, the CHI 1129.88, and the DBI a local minimum of 0.56. Similarly, Euclidean K-Means achieves a Silhouette Score of 0.53 and a CHI of 890.79 at $k = 3$. Although GMM and Agglomerative Clustering yield lower silhouette values (0.33 and 0.41, respectively), they still provide consistent spatial patterns. The convergence of all three metrics around $k = 3$ suggests a robust clustering solution that balances cohesion, separation, and geochemical interpretability.

Table 2. Internal validation metrics (Silhouette, Calinski-Harabasz, Davies-Bouldin) for different clustering configurations (k = 2 to 7) and methods applied in the multivariate classification of the Quiulacocha tailings deposit.

Metric/k	2	3	4	5	6	7
Euclidean K-Means						
Silhouette Score	0.51	0.53	0.58	0.66	0.69	0.74
Calinski-Harabasz	856.76	890.79	963.74	1,193.14	1,687.04	2,689.77
Davies-Bouldin	0.85	0.86	0.69	0.57	0.45	0.35
Riemannian K-Means						
Silhouette Score	0.56	0.58	0.46	0.51	0.65	0.60
Calinski-Harabasz	1354.61	1129.88	931.72	973.68	2504.34	2731.22
Davies-Bouldin	0.66	0.56	0.87	0.68	0.45	0.45
GMM						
Silhouette Score	0.38	0.33	0.39	0.44	0.54	0.54
Calinski-Harabasz	435.15	359.90	346.25	532.49	774.92	720.58
Davies-Bouldin	1.24	1.25	1.20	0.77	0.65	0.58
Agglomerative Clustering						
Silhouette Score	0.38	0.41	0.42	0.48	0.56	0.62
Calinski-Harabasz	435.15	454.54	483.81	578.99	759.07	992.84
Davies-Bouldin	1.24	0.85	0.92	0.82	0.61	0.50

Additionally, a sensitivity analysis of the Riemannian K-Means algorithm (Table 3) highlights the importance of the "radius" parameter used in covariance estimation. At small radii (300–400 m), clusters are less cohesive (DBI > 1.0) and more overlapping. Optimal performance is reached between 700 and 900 m, where Silhouette scores increase (up to 0.88) and DBI values drop (to 0.36), indicating compact and well-separated clusters. However, at larger radii, the number of validated points decreases, suggesting a trade-off between spatial resolution and cluster interpretability.

Table 3. Sensitivity analysis of the "radius" parameter in Riemannian K-Means classification.

Radius	N° validate points	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
300	927.00	0.54	983.80	0.69
400		0.48	544.09	1.14
500		0.58	1129.88	0.56
600		0.55	1914.66	0.79
700		0.60	2236.18	0.59
800		0.61	1241.16	0.45
900		0.88	5262.51	0.36
1000		0.94	1345.62	0.12

4.3. Multivariate Clustering (CLR-transformed data)

The clustering analysis was also performed using the geochemical dataset transformed via the CLR method, which is particularly suitable for addressing the compositional nature of geochemical data. Figure 7 displays the spatial distribution of the resulting clusters using k = 3, which was again selected as the optimal number of groups based on the internal validation metrics and the interpretability of the segmentation.

As observed in the clustering with raw data, the CLR-transformed analysis also produces a clear vertical geochemical stratification within the Quiulacocha tailings deposit, reinforcing the robustness of the clustering structure. Cluster 1 (red) occupies intermediate levels of the deposit and is characterized by high concentrations of Cu (~0.6%) and Au (~1.2 g/t). These values are consistent with the deeper portions of the deposit, suggesting that Cluster 1 captures domains enriched in Cu–Au-bearing sulfide minerals, likely related to early stages of mining activities targeting high-sulfidation epithermal mineralization [9]. Cluster 2 (blue) dominates the upper sections of the deposit and presents relatively high Fe contents (15–45%) with low concentrations of trace metals, such as Zn, Pb, Ag, and Cu. This geochemical signature suggests that Cluster 2 corresponds to a homogeneous, oxidized zone likely the result of post-depositional processes such as sulfide leaching and iron oxide precipitation. These conditions are characteristic of exposed tailings under prolonged weathering, which leads to the formation of an iron-rich residual matrix [4]. In contrast, Cluster 3 (green) is predominantly located at the deepest levels of the deposit and exhibits high concentrations of Zn (up to ~7%), Pb (up to ~3%), and Ag (>120 g/t). This composition aligns with the

presence of polymetallic mineralization residues, including sphalerite (ZnS), galena (PbS), and argentiferous sulfosalts, suggesting that this cluster represents the Pb–Zn–Ag-rich mineralized domains processed during the later phases of the mine's production history.

The CLR-based clustering effectively delineates three mineralization zones within the tailings: a Cu–Au-rich sulfide domain (Cluster 1), an Fe-rich, weathered oxidized cap (Cluster 2), and a Zn–Pb–Ag polymetallic zone (Cluster 3). This segmentation not only confirms the geochemical zoning suggested by the raw data analysis but also reinforces the mineralogical interpretation of the deposit's depositional sequence and metallurgical history.

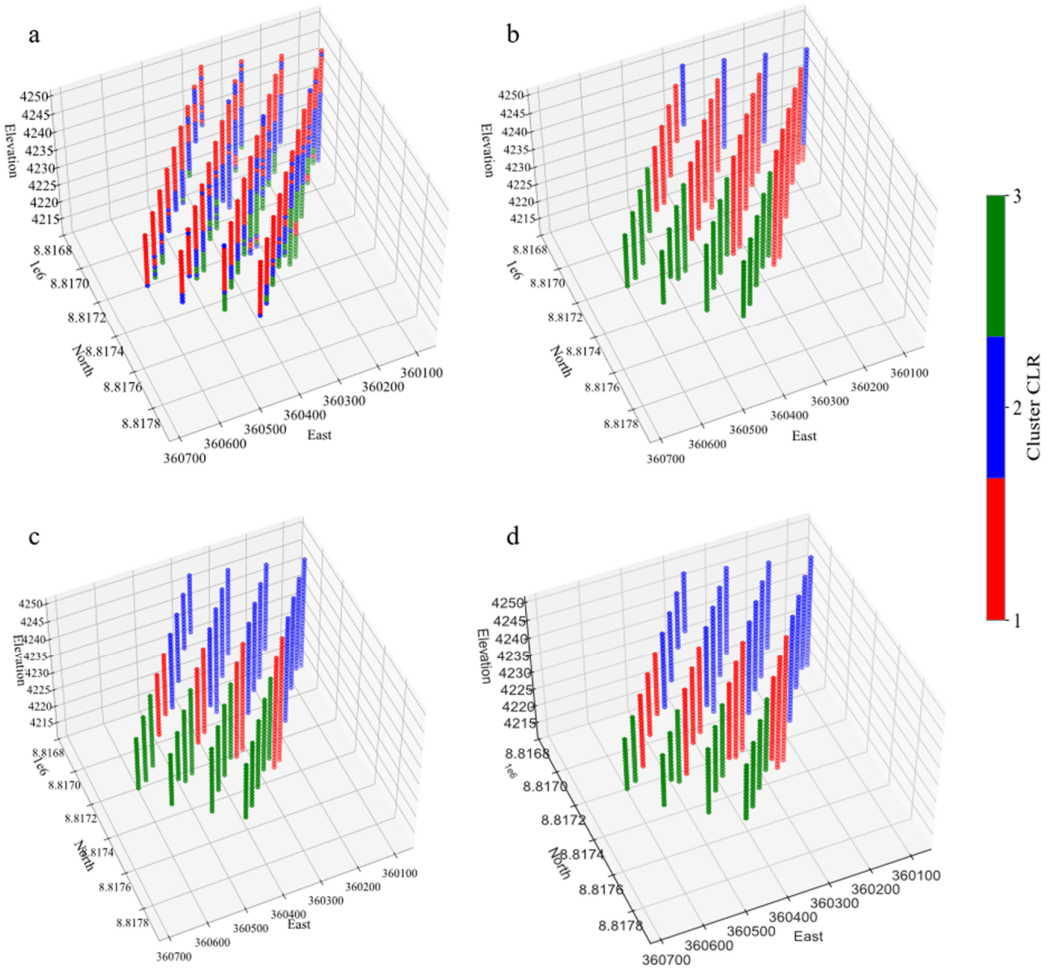

Figure 7. Multivariate classification of the Quiulacocha tailings deposit into three clusters using data transformed with centered log-ratio (CLR): (a) Euclidean K-Means, (b) Riemannian K-Means, (c) Gaussian Mixture Model, (d) Agglomerative Clustering.

Regarding clustering performance, the internal validation metrics (Table 4) highlight some differences between algorithms. The Silhouette scores tend to be lower than those obtained with raw data, ranging from 0.30 to 0.34 for Euclidean and Riemannian K-Means at $k = 3$, indicating reduced cluster cohesion. However, GMM and Agglomerative Clustering show slightly better results, with Silhouette scores of 0.43 and 0.40, respectively, and Davies-Bouldin indices around 1.0, suggesting moderate inter-cluster separation.

The CHI shows a consistent increase with increasing k, particularly for GMM, which reaches 1527.82 at $k = 7$, and for Agglomerative Clustering, which reaches 1523.57. Among all methods, Agglomerative Clustering exhibits the highest stability and interpretability, achieving a Silhouette score of 0.63 and a DBI of 0.48 at k = 7. This suggests that hierarchical methods may be better suited for compositional data under the CLR framework, as they are capable of capturing nested and irregular geochemical structures.

Table 4. Internal validation metrics (Silhouette, Calinski-Harabasz, Davies-Bouldin) for different clustering configurations (k = 2 to 7) and methods applied in the multivariate classification of the Quiulacocha tailings deposit using CLR-transformed data.

Metric/k	2	3	4	5	6	7
Euclidean K-Means (CLR-transformed data)						
Silhouette Score	0.40	0.30	0.30	0.30	0.31	0.26
Calinski-Harabasz	577.09	489.48	471.07	452.70	451.93	431.04
Davies-Bouldin	1.02	1.19	1.03	1.05	0.98	1.09
Riemannian K-Means (CLR-transformed data)						
Silhouette Score	0.44	0.34	0.49	0.38	0.38	0.45
Calinski-Harabasz	745.96	421.60	677.99	538.94	360.29	617.08
Davies-Bouldin	0.92	0.96	0.80	0.80	1.14	0.79
GMM (CLR-transformed data)						
Silhouette Score	0.38	0.43	0.41	0.46	0.54	0.60
Calinski-Harabasz	486.20	575.92	500.78	631.03	1169.61	1527.82
Davies-Bouldin	1.15	1.01	1.09	0.71	0.60	0.57
Agglomerative Clustering (CLR-transformed data)						
Silhouette Score	0.45	0.40	0.44	0.53	0.58	0.63
Calinski-Harabasz	541.99	557.22	753.86	924.52	1191.02	1523.57
Davies-Bouldin	0.97	0.98	0.73	0.62	0.56	0.48

4.4. Validation and performance analysis

Figure 8 displays the silhouette coefficient distributions for the clustering results with $k = 3$ using the raw geochemical and spatial data. Among the evaluated algorithms, Riemannian K-Means exhibits the highest mean silhouette score (0.58), followed closely by Euclidean K-Means (0.53). These results indicate well-separated and internally cohesive clusters, with the majority of samples showing silhouette values above 0.4. In contrast, the Gaussian Mixture Model (GMM) and Agglomerative Clustering exhibit lower average silhouette values of 0.33 and 0.41, respectively, suggesting less compact groupings and more overlapping boundaries. However, even these lower-performing models reveal discernible geochemical domains, which adds value when analyzing systems with complex mineralization such as polymetallic tailings.

The convergence of the three internal validation metrics Silhouette, Calinski-Harabasz, and Davies-Bouldin around $k = 3$, supports this configuration as the most robust. It effectively balances cohesion and separation, avoids overfitting, and provides a meaningful segmentation of the deposit that aligns with known metallogenic structures. These findings are consistent with recent studies emphasizing the importance of integrated validation in multivariate geochemical modeling of mine tailings and other spatially heterogeneous systems [53, 54, 70].

Figure 9 presents the silhouette profiles for the same clustering configuration ($k = 3$) using CLR-transformed data. In general, silhouette values are lower compared to those from raw data, with average scores of 0.30 for Euclidean K-Means and 0.47 for Riemannian K-Means, indicating reduced intra-cluster cohesion. The GMM shows slightly improved performance with a mean of 0.43, while Agglomerative Clustering achieves 0.40, reflecting moderate separation. This reduction in silhouette scores is attributed to the CLR transformation, which, while mitigating scale-related biases, also tends to compress variance structure, reducing the contrast between clusters. Nonetheless, Riemannian K-Means and Agglomerative Clustering retain better performance under CLR transformation, especially in terms of interpretability and alignment with geologically meaningful mineralization zones. These methods appear more robust to transformations and better suited for identifying domains with compositional complexity, such as those present in tailings deposits with mixed mineralogical origins.

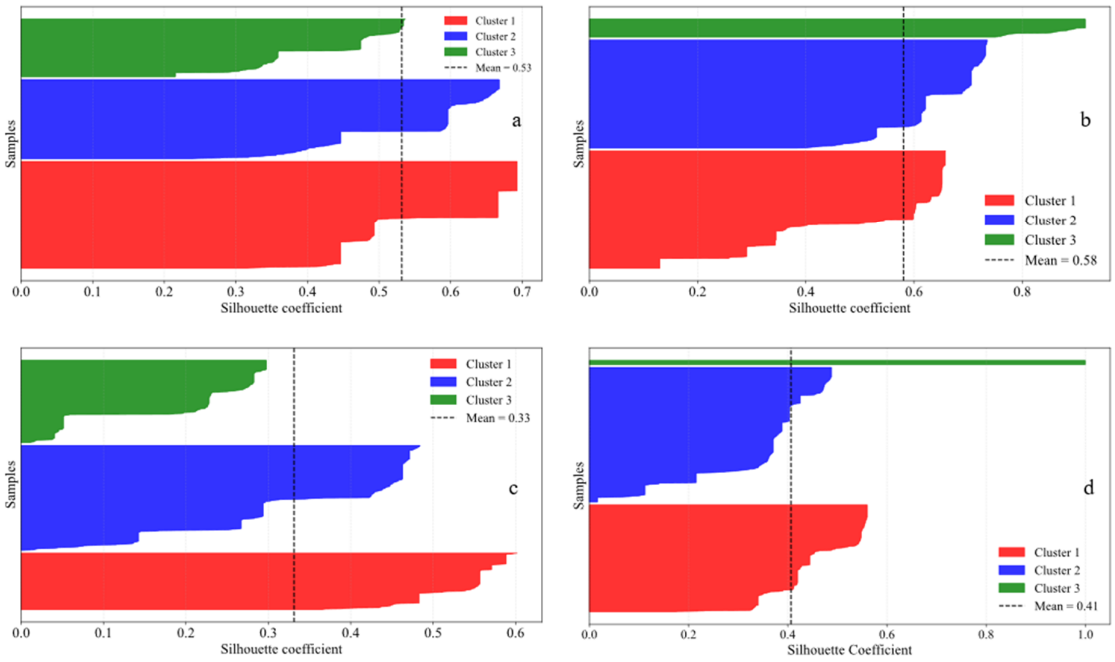

Figure 8. Silhouette coefficient distribution for the *k*=3 clustering solution. (a) Euclidean, (b) Riemannian, (c) GMM, (d) Agglomerative Clustering.

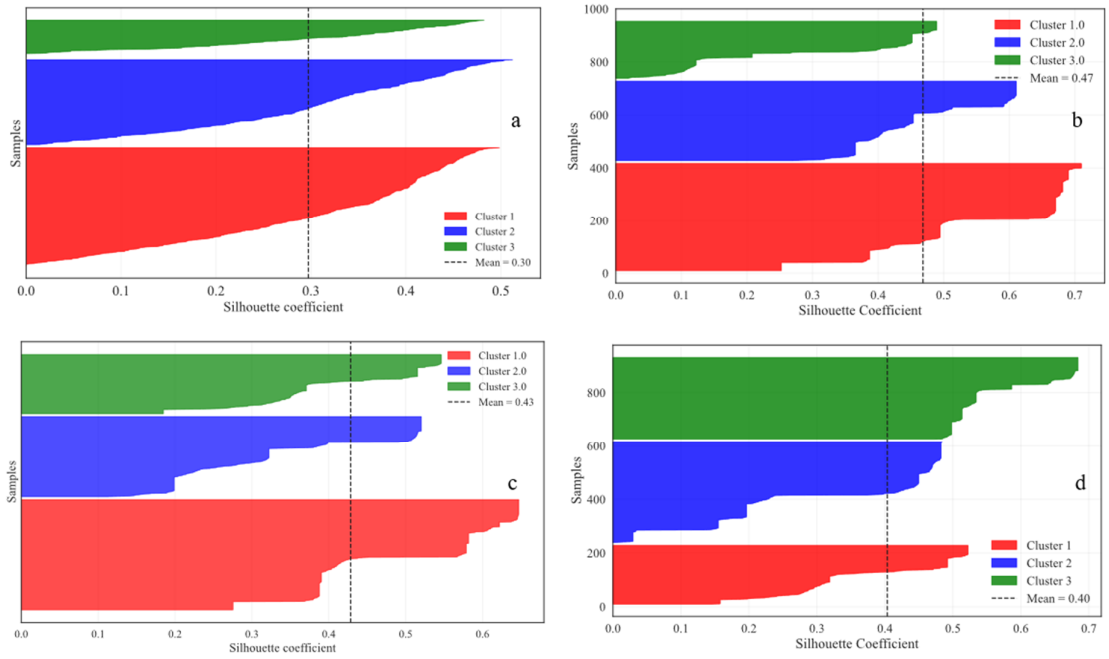

Figure 9. Silhouette coefficient distribution for the *k*=3 clustering solution using CLR-transformed data: (a) Euclidian, (b) Riemannian, (c) GMM, (d) Agglomerative Clustering.

4.5. Geochemical characterization of clustering

The visualization of classified samples in bivariate and three-dimensional geochemical spaces (Zn–Pb, Zn–Pb–Fe), with Cu included as an additional variable, is presented in Figure 10. Across the four algorithms, the segmentation patterns consistently reflect the inherent compositional heterogeneity of the deposit, in agreement with previous studies of polymetallic tailings [1]. Figure 11 shows the corresponding results with CLR-transformed data, which generally reproduce the same domains but with weaker separation, particularly between Clusters 1

and 2. This overlap highlights the tendency of CLR to compress variance structures, which complicates the identification of distinct mineralized domains. For this reason, the final geochemical interpretation relies primarily on the clustering results obtained from raw data.

The combined analysis of Table 5 confirms the presence of three well-differentiated compositional domains, robust across clustering algorithms despite some numerical variations. Cluster 1 is consistently enriched in Cu (≈0.08–0.11%) and Au (≈0.13–0.18 g/t), while showing intermediate levels of Ag (≈49–53 g/t), Zn (≈1.46–1.58%), and Pb (≈0.74–0.81%), together with the lowest Fe contents (≈18–26%). This signature corresponds to Cu–Au–rich domains, reflecting the earliest tailings deposition derived from the processing of chalcopyrite- and gold-bearing sulfides [9]. Cluster 2 systematically records the highest Pb grades (≈1.00–1.10%), intermediate Zn (≈1.39–1.40%) and Ag (≈49–50 g/t), and relatively stable Fe contents (≈24–27%), while Cu and Au remain low (≈0.10% and ≈0.06–0.09 g/t). This composition is consistent with Pb–Zn–Ag–rich domains, associated with intermediate processing phases of galena–sphalerite ore bodies at Cerro de Pasco [41], and further stabilized by surface oxidation processes that preserved Fe as secondary oxides [4, 5]. Cluster 3 stands out for its highest Zn (≈1.67–1.87%) and Ag (≈55–62 g/t) contents, accompanied by high Fe (≈30–32%) and relatively low Cu (≈0.05–0.08%) and Au (≈0.04–0.08 g/t). This geochemical profile corresponds to Zn–Ag–Fe–rich domains, reflecting later processing of sphalerite-rich ores and possible remobilization of Fe-bearing phases during post-depositional processes [3, 7].

Although minor variations exist across algorithms, the relative structure of these three geochemical domains is stable, confirming the robustness of the clustering. Agglomerative Clustering shows the strongest selectivity, isolating a small but distinct subgroup in Cluster 3 (only 16 samples) with extreme Zn–Ag–Fe enrichment, whereas partitioning methods such as K-Means distribute samples more evenly. This consistency across methods indicates that clustering not only reproduces compositional variability but also reflects metallogenic zoning inherited from the original mineralized system.

4.6. Distribution analysis (Q-Q plots)

The analysis of distribution patterns within clusters is further refined through Q–Q plots of Ag, Cu, and Au, presented in Figure 12. These metals were selected due to their economic relevance and their strong discriminative capacity across clusters. For Ag, Clusters 2 and 3 display distributions shifted toward higher values, with Cluster 3 showing more pronounced heavy tails, consistent with its association with Zn–Ag–enriched residues. For Cu and Au, Cluster 1 dominates the upper tails of the distributions, reflecting its link with early Cu–Au sulfide processing phases. These trends are consistent across Euclidean K-Means, Riemannian K-Means, GMM, and Agglomerative Clustering, reinforcing the robustness of the multivariate clustering approach in capturing geochemically meaningful patterns in polymetallic tailings. Notably, the hierarchical method identified a very small population of samples with extreme compositions, underscoring its selectivity in isolating rare but geochemically significant domains.

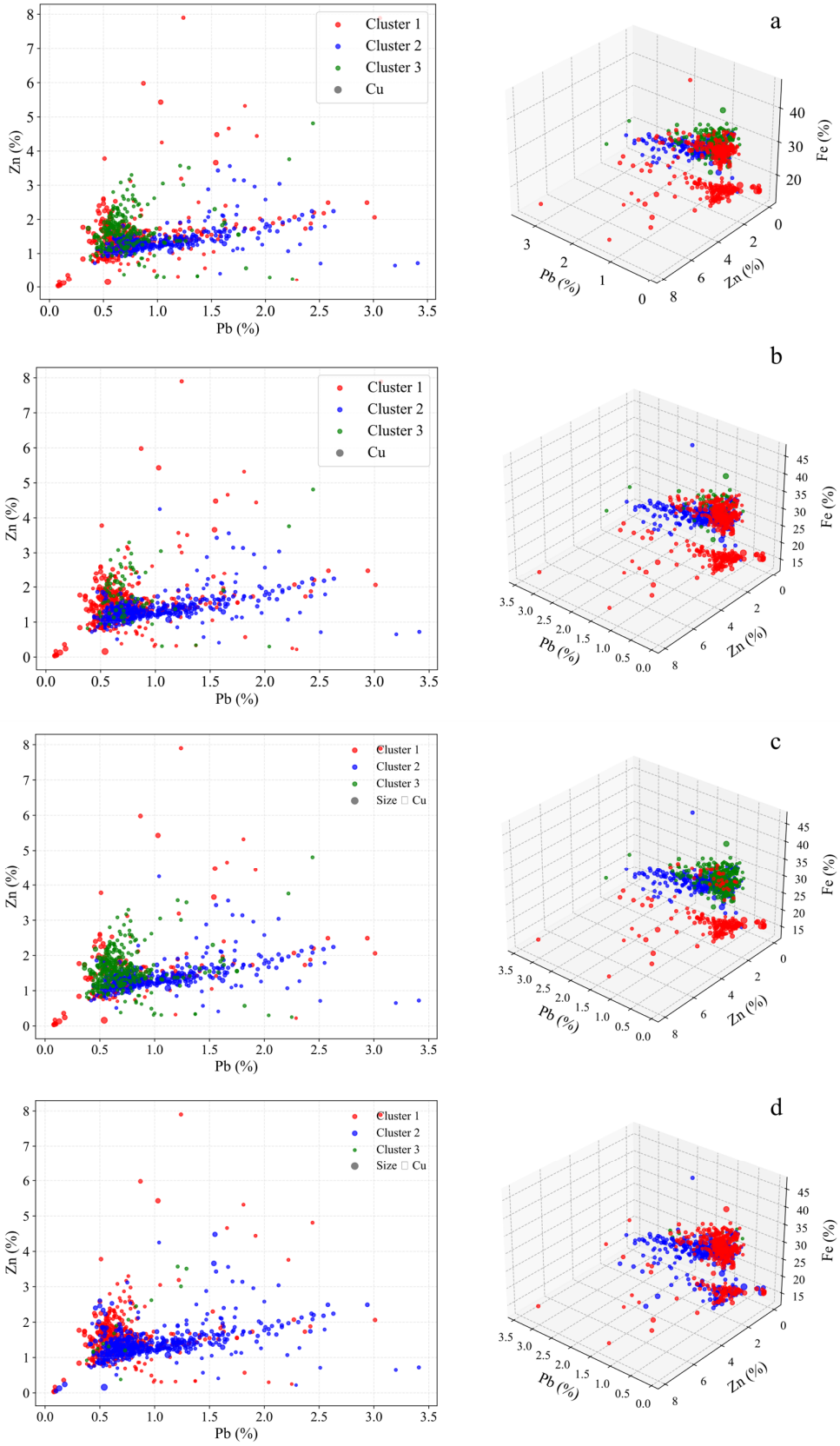

Figure 10. Visualization of classified samples in geochemical feature space. (a) Euclidean K-Means, (b) Riemannian K-Means, (c) GMM, (d) Agglomerative Clustering.

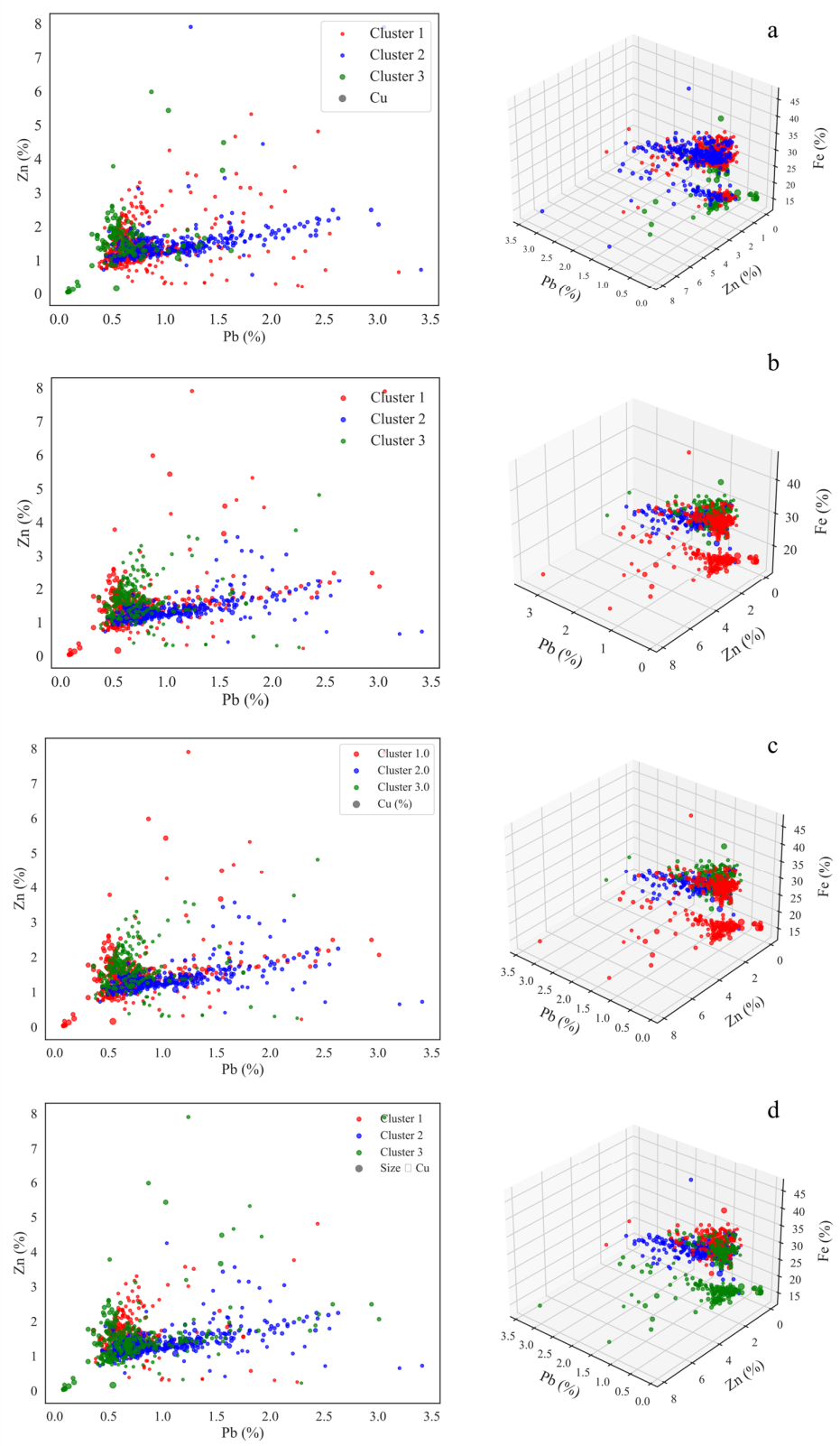

Figure 11. Visualization of classified samples in geochemical feature space using CLR-transformed data: (a) Euclidean K-Means, (b) Riemannian K-Means, (c) GMM, (d) Agglomerative Clustering.

Table 5. Mean grades of the clusters obtained with different clustering methods.

Method / Cluster	Count	Ag (g/t)	Zn (%)	Pb (%)	Cu (%)	Au (g/t)	Fe (%)
Euclidean K-Means							
1	408	49.21	1.46	0.80	0.10	0.14	22.64
2	300	50.56	1.39	1.10	0.10	0.06	26.31
3	219	57.55	1.67	0.75	0.07	0.08	30.83
Riemannian K-Means							
1	456	52.02	1.52	0.76	0.09	0.13	24.10
2	403	49.40	1.39	1.04	0.10	0.07	26.56
3	68	62.08	1.87	0.79	0.07	0.07	32.21
GMM							
1	213	50.90	1.54	0.81	0.11	0.18	18.00
2	403	49.40	1.39	1.04	0.10	0.07	26.56
3	311	54.99	1.59	0.72	0.08	0.08	30.05
Agglomerative Clustering							
1	401	53.44	1.58	0.74	0.08	0.11	26.81
2	510	49.97	1.40	1.00	0.10	0.09	24.75
3	16	58.67	1.85	0.74	0.05	0.04	32.01

When applying CLR-transformed data (Figure 13), the separation among clusters becomes less distinct. In the case of Ag, there is a strong overlap between Clusters 2 and 3, which limits the capacity to clearly isolate Zn–Ag–enriched domains. Similarly, for Cu and Au, the dominance of Cluster 1 in the upper tails is less evident, suggesting that the CLR transformation, while effective in reducing skewness and mitigating the effect of extreme values, also attenuates natural contrasts that are critical for geochemical discrimination. Consequently, although CLR provides statistical homogenization, clustering results based on raw data preserve more interpretable and geologically consistent patterns, particularly for economically relevant metals.

4.7. Consistency Across Clustering Methods

The consistency of the clustering solutions was evaluated using the Adjusted Rand Index (ARI), which quantifies the degree of similarity between two partitions while correcting for chance agreement. Figure 14 presents the ARI matrix for the raw data, showing that Riemannian K-Means and Agglomerative Clustering achieve perfect agreement (ARI = 1.00). This result indicates that both methods capture essentially the same latent structure in the dataset, which is consistent with their shared ability to incorporate covariance matrices and thus model inter-variable dependencies [46]. A relatively high agreement is also observed between Riemannian K-Means and GMM (ARI = 0.5759), suggesting that both approaches identify similar geochemical domains, although GMM introduces greater flexibility by accounting for probabilistic assignments rather than deterministic partitions. In contrast, Euclidean K-Means exhibits only moderate to low consistency with the other methods (ARI ≈ 0.17–0.44), reflecting its reliance on Euclidean distances in the standardized feature space, which tends to overlook correlation structures that are fundamental in multivariate geochemical systems [47, 60]. These results highlight that methods explicitly accounting for data dispersion geometry and covariance-based dependencies are more robust for delineating mineralized zones within tailings deposits characterized by high compositional heterogeneity.

When considering the CLR-transformed dataset (Figure 15), the consistency patterns shift. Once again, Riemannian K-Means and Agglomerative Clustering display perfect agreement (ARI = 1.00), and both also show full agreement with GMM, confirming that covariance-based methods converge to nearly identical partitions after CLR transformation. However, Euclidean K-Means remains poorly consistent with the other approaches, with ARI values close to zero (0.02–0.04). This outcome demonstrates that the CLR transformation does not improve the reliability of Euclidean K-Means relative to correlation-sensitive algorithms. From a mineralogical standpoint, this reinforces that robust delineation of Cu–Au–rich zones and Zn–Pb–Ag–enriched domains requires methods capable of exploiting covariance structures, while purely distance-based methods risk oversimplifying the true geochemical variability of the Quiulacocha tailings.

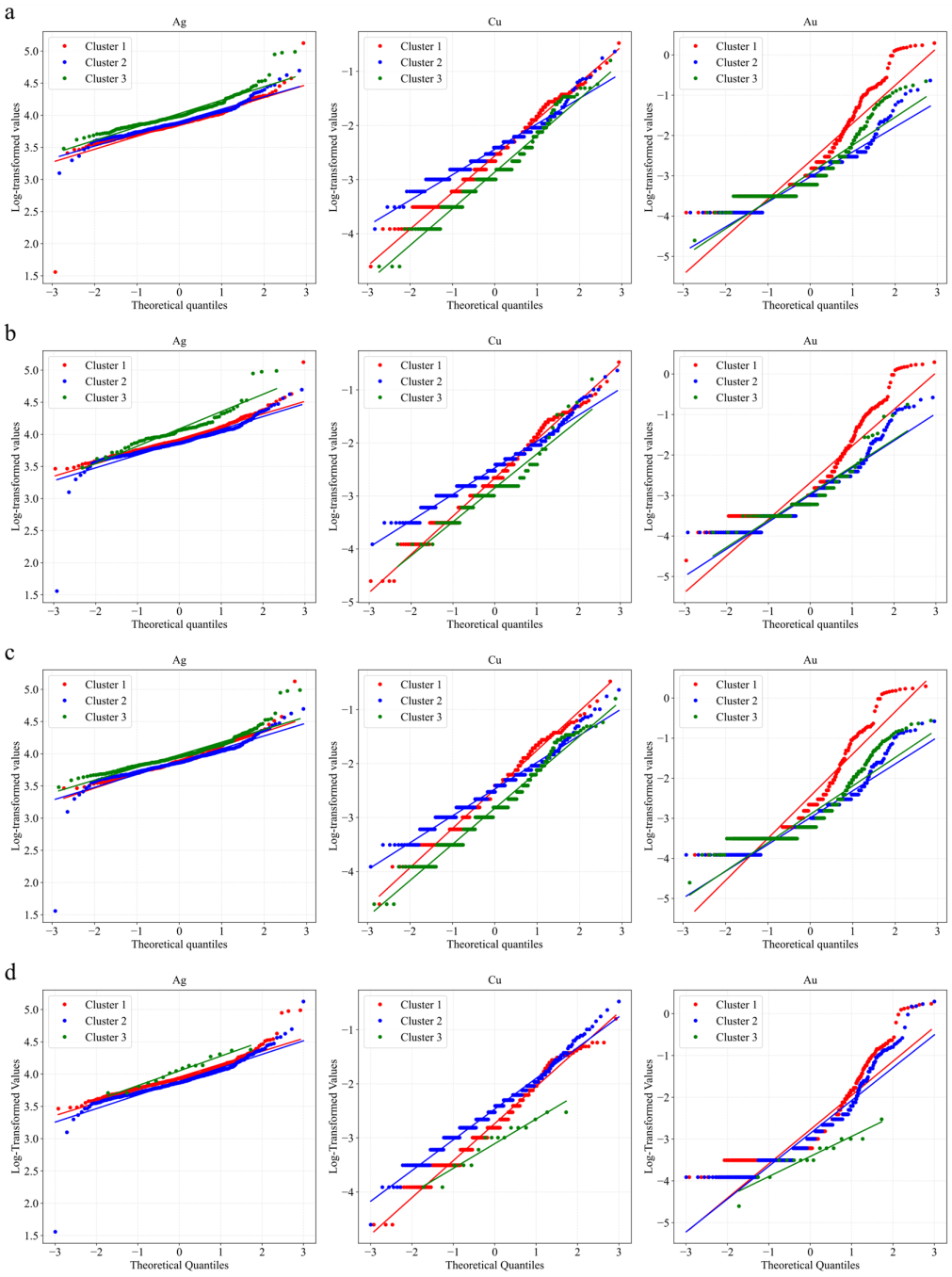

Figure 12. Q–Q plots of geochemical variables for the obtained clusters. Cluster distributions are shown for Ag, Cu, and Au. (a) Euclidean K-Means, (b) Riemannian K-Means, (c) GMM, (d) Agglomerative Clustering.

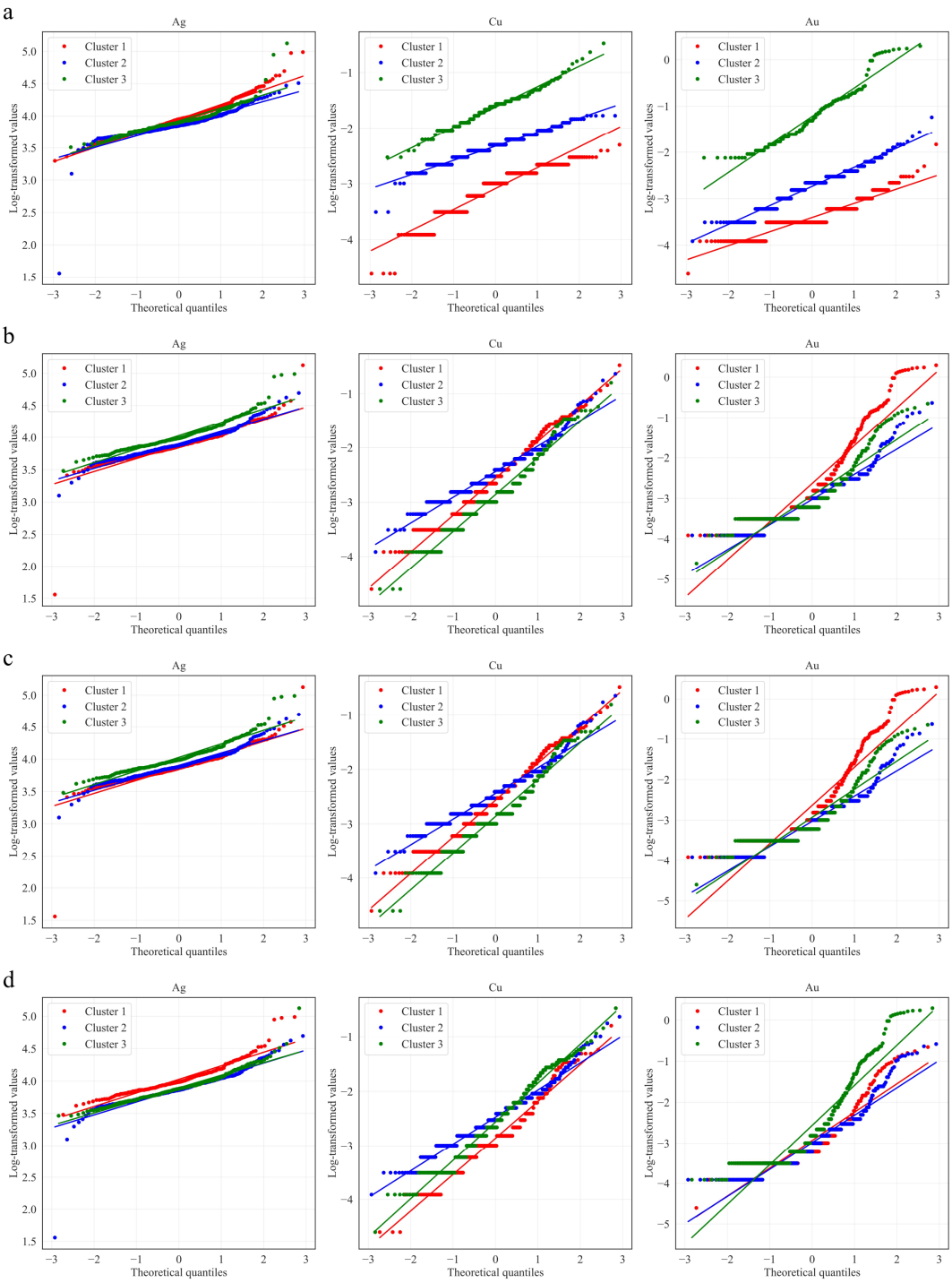

Figure 13. Q–Q plots of geochemical variables for the obtained clusters using CLR-transformed data. Cluster distributions are shown for Ag, Cu, and Au. (a) Euclidean K-Means, (b) Riemannian K-Means, (c) GMM, (d) Agglomerative Clustering.

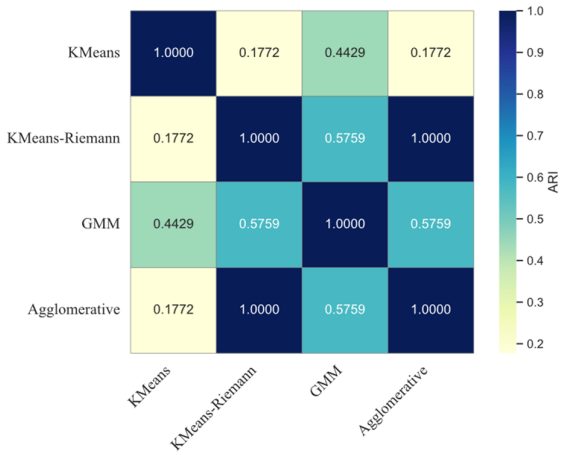

Figure 14. ARI matrix between the applied clustering methods. Higher values indicate greater consistency between the partitions produced by different algorithms.

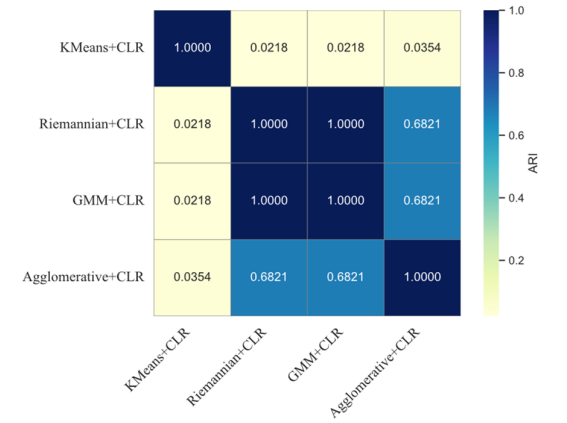

Figure 15. ARI matrix between the applied clustering methods using CLR-transformed data. Higher values indicate greater consistency between the partitions produced by different algorithms.

5. Conclusions

This study presents a comprehensive multivariate analysis of the Quiulacocha tailings deposit, comparing four unsupervised clustering algorithms Euclidean K-Means, Riemannian K-Means, GMM, and Agglomerative Clustering applied to high-dimensional geochemical data. The clustering results consistently support a three-domain structure (k = 3), which aligns with known mineralization and processing histories. In the case of raw geochemical data, the best-performing methods (Riemannian K-Means and Agglomerative Clustering) achieved high internal validation scores, with average Silhouette coefficients of 0.58 and 0.41, respectively, and Davies-Bouldin indices below 0.60 for k = 3. These methods also showed strong geospatial coherence and reproducibility, reaching an ARI of 1.00 between them and 0.5759 when compared with GMM, indicating robust and stable segmentation of latent geochemical structures. In contrast, Euclidean K-Means showed lower cohesion (Silhouette = 0.53) and moderate ARI scores (0.17–0.44), highlighting its limited sensitivity to the covariance geometry inherent to compositional datasets.

Geochemical characterization of the clusters revealed three distinct mineralization domains. Cluster 1 consistently exhibited the highest contents of Cu (0.08–0.11%) and Au (0.13–0.18 g/t), along with intermediate levels of Ag (49–53 g/t), Zn (1.46–1.58%), and Pb (0.74–0.81%), and the lowest Fe (18–26%), indicating an association with Cu–Au sulfide processing residues. Cluster 2 showed the highest Pb concentrations (1.00–1.10%), intermediate Zn (1.39–1.40%) and Ag (49–50 g/t), and low Au (0.06–0.09 g/t), along with a relatively homogeneous Fe content (24–27%), reflecting domains related to Pb–Zn–Ag ores and oxidation-stabilized Fe phases. Cluster 3 presented the highest Zn (1.67–1.87%) and Ag (55–62 g/t) concentrations, with low Cu (0.05–0.08%) and Au (0.04–0.08 g/t), and the highest Fe contents (30–32%), suggesting a link with late-stage processing or remobilization. While the CLR transformation improved the treatment of compositional constraints, it reduced clustering performance (Silhouette < 0.34, increased overlap), confirming that raw-data clustering yielded more geochemically interpretable results.

Despite the robustness of the results, several limitations are recognized. The analysis focused on a selected suite of major and economically relevant elements, and did not incorporate detailed mineralogical or textural information. The lack of external validation using metallurgical or mineralogical data also limits the ability to directly confirm cluster-domain interpretations. Moreover, while z-score normalization and CLR transformation were tested, future studies should evaluate alternative preprocessing approaches (e.g., Box-Cox, ILR transformations) to improve cluster compactness and interpretability. The application of spatially explicit or deep-learning clustering models may further enhance resolution in structurally complex deposits. Future work should also consider time-dependent or process-based clustering to better understand post-depositional changes and support predictive modeling for resource recovery and environmental risk assessment.

References

[1]. Falagán C, Grail BM, Johnson DB (2017) New approaches for extracting and recovering metals from mine tailings. *Miner Eng*. https://doi.org/10.1016/j.mineng.2016.10.008

[2]. Kermani M, Hassani FP, Aflaki E, Benzaazoua M, Nokken M (2015) Evaluation of the effect of sodium silicate addition to mine backfill, Gelfill - Part 1. *Journal of Rock Mechanics and Geotechnical Engineering*. https://doi.org/10.1016/j.jrmge.2015.03.006

[3]. Nascimento SC, Cooke DR, Cracknell MJ, Miller CB, Parbhakar-Fox A (2025) Mineralogical and geochemical characterization of mine tailings in the King river delta, Western Tasmania: Implications for long-term stability of trace elements. *Applied Geochemistry* 184:106366

[4]. Elghali A, Benzaazoua M, Bussière B, Kennedy C, Parwani R, Graham S (2019) The role of hardpan formation on the reactivity of sulfidic mine tailings: A case study at Joutel mine (Québec). *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2018.11.066

[5]. Gäbler HE (1997) Mobility of heavy metals as a function of pH of samples from an overbank sediment profile contaminated by mining activities. *J Geochem Explor*.

[6]. Anju M, Banerjee DK (2010) Comparison of two sequential extraction procedures for heavy metal partitioning in mine tailings. *Chemosphere*.

[7]. Cook NJ, Ciobanu CL, Pring A, Skinner W, Shimizu M, Danyushevsky L, Saini-Eidukat B, Melcher F (2009) Trace and minor elements in sphalerite: A LA-ICPMS study. *Geochim Cosmochim Acta*.

[8]. Yin Z, Sun W, Hu Y, Zhang C, Guan Q, Wu K (2018) Evaluation of the possibility of copper recovery from tailings by flotation through bench-scale, commissioning, and industrial tests. *J Clean Prod*.

[9]. Antonijević MM, Dimitrijević MD, Stevanović ZO, Serbula SM, Bogdanovic GD (2008) Investigation of the possibility of copper recovery from the flotation tailings by acid leaching. *J Hazard Mater*. https://doi.org/10.1016/j.jhazmat.2008.01.063

[10]. Wanhainen C, Palsson BI, Martinsson O, Lahaye Y (2017) Rare earth mineralogy in tailings from Kiirunavaara iron ore, northern Sweden: Implications for mineral processing. *Minerals and Metallurgical Processing*. https://doi.org/10.19150/mmp.7859

[11]. Schuenemeyer JH, Drew LJ (2010) Statistics for Earth and Environmental Scientists. *Statistics for Earth and Environmental Scientists*. https://doi.org/10.1002/9780470650707

[12]. Aitchison J (1982) The Statistical Analysis of Compositional Data. J R Stat Soc Series B Stat Methodol. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

[13]. Filzmoser P, Hron K, Templ M (2012) Discriminant analysis for compositional data and robust parameter estimation. Comput Stat. https://doi.org/10.1007/s00180-011-0279-8

[14]. Shanmugam R (2019) Applied compositional data analysis: with worked examples in R. J Stat Comput Simul. https://doi.org/10.1080/00949655.2019.1628880

[15]. Pacifico LR, Guarino A, Iannone A, Albanese S (2025) Accounting for the Compositional Nature of Geochemical Data to Improve the Interpretation of Their Univariate and Multivariate Spatial Patterns: A Case Study from the Campania Region (Italy). Geosciences (Basel) 15:20

[16]. Somma R, Ebrahimi P, Troise C, De Natale G, Guarino A, Cicchella D, Albanese S (2021) The first application of compositional data analysis (CoDA) in a multivariate perspective for detection of pollution source in sea sediments: The Pozzuoli Bay (Italy) case study. *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2021.129955

[17]. Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci. https://doi.org/10.5194/hess-11-1633-2007

[18]. Foley JA, DeFries R, Asner GP, et al (2005) Global consequences of land use. Science (1979). https://doi.org/10.1126/science.1111772

[19]. Verhoeven G (2011) Taking computer vision aloft - archaeological three-dimensional reconstructions from aerial photographs with photoscan. Archaeol Prospect. https://doi.org/10.1002/arp.399

[20]. Carranza EJM (2009) Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Handbook of Exploration and Environmental Geochemistry. https://doi.org/10.1016/S1874-2734(09)70004-X

[21]. Koutsaki E, Vardakis G, Papadakis N (2023) Spatiotemporal Data Mining Problems and Methods. Analytics. https://doi.org/10.3390/analytics2020027

[22]. HAWKES HE, WEBB JS (1963) Geochemistry in Mineral Exploration. Soil Sci. https://doi.org/10.1097/00010694-196304000-00016

[23]. Ilgen E, Levsen K, Angerer J, Schneider P, Heinrich J, Wichmann HE (2001) Aromatic hydrocarbons in the atmospheric environment - Part II: Univariate and multivariate analysis and case studies of indoor concentrations. Atmos Environ. https://doi.org/10.1016/S1352-2310(00)00490-8

[24]. Shahrestani S, Cohen DR, Mokhtari AR (2024) A comparison of PCA and ICA in geochemical pattern recognition of soil data: The case of Cyprus. J Geochem Explor 264:107539

[25]. Dominech S, Yang S, Aruta A, Gramazio A, Albanese S (2022) Multivariate analysis of dilution-corrected residuals to improve the interpretation of geochemical anomalies and determine their potential sources: The Mingardo River case study (Southern Italy). J Geochem Explor. https://doi.org/10.1016/j.gexplo.2021.106890

[26]. Goovaerts P (1997) Geostatistics for Natural Resources Evaluation (Applied Geostatistics). Oxford University Press, New York

[27]. Aghahadi MH, Jozanikohan G, Asghari O, Talesh Hosseini S, Emery X, Rezaei M (2024) Geochemical anomaly separation based on geology, geostatistics, compositional data and local singularity analyses: A case study from the kuh panj copper deposit, Iran. Applied Geochemistry 173:106135

[28]. Cheng Q (2007) Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol Rev. https://doi.org/10.1016/j.oregeorev.2006.10.002

[29]. Cheng Q, Agterberg FP, Ballantyne SB (1994) The separation of geochemical anomalies from background by fractal methods. J Geochem Explor. https://doi.org/10.1016/0375-6742(94)90013-2

[30]. Li C, Ma. T, Shi J (2003) Application of a fractal method relating concentrations and distances for separation of geochemical anomalies from background. J Geochem Explor. https://doi.org/10.1016/S0375-6742(02)00276-5.

[31]. Pourgholam MM, Adib A, Afzal P, Rahbar K, Gholinejad M (2025) Deep learning and fractal-wavelet techniques for magnetite-apatite exploration in Tarom Iran. Scientific reports 15(1), 31907. https://doi.org/10.1038/s41598-025-16040-2.

[32]. Farhadi S, Tatullo S, Boveiri M, Afzal P (2024) Evaluating StackingC and ensemble models for enhanced lithological classification in geological mapping. Journal of Geochemical Exploration 260, 107441. https://doi.org/10.1016/j.gexplo.2024.107441.

[33]. Saadati H, Afzal P, Torshizian H, Solgi A (2025) Application of Stepwise Fractal Modeling for Interpretation of Remote Sensing data, NE Iran. Iranian Journal of Earth Sciences 17 (3). https://doi.org/10.57647/j.ijes.2025.16799.

[34]. Samadi S, Afzal P, Arian M, Solgi A, Maleki Z, Seraj M (2025) Detection of effective porosity zones utilizing fractal modeling in an oilfield reservoir, NW Iran. Geopersia 15 (1), 85-95. https://doi.org/10.22059/geope.2024.380851.648770.

[35]. Cotrina-Teatino M, Marquina-Araujo J, Mamani-Quispe J, Chira-Fernandez J, Castillo-Chung A, Arango-Retamozo S, González-Vasquez J, Ortiz-Quintanilla S (2025) Geochemical and mineralogical characterization of critical elements in gold tailings from the La Cienega, Peru, and assessment of their reuse potential. Journal of Environmental Chemical Engineering 13 (5), 118497. https://doi.org/10.1016/j.jece.2025.118497.

[36]. Cotrina-Teatino M, Marquina-Araujo J, Mamani-Quispe J, Guartán J, Castillo-Chung A, Arango-Retamozo S, González-Vasquez J, Ortiz-Quintanilla S (2025) Strategic potential assessment of lanthanum and scandium through geochemical-lithological analysis with unsupervised machine learning in southern Ecuador. Resource Policy 109, 105731. https://doi.org/10.1016/j.resourpol.2025.105731.

[37]. Ahmed AD, Hood SB, Cooke DR, Belousov I (2020) Unsupervised clustering of LA-ICP-MS raster map data for geological interpretation: A case study using epidote from the Yerington district, Nevada. Applied Computing and Geosciences. https://doi.org/10.1016/j.acags.2020.100036

[38]. Wang X, Chen Y (2025) Unsupervised detection of multivariate geochemical anomalies using a high-performance deep autoencoder Gaussian mixture model. J Geochem Explor 271:107671

[39]. Zhou W, Maerz NH (2002) Implementation of multivariate clustering methods for characterizing discontinuities data from scanlines and oriented boreholes. Comput Geosci. https://doi.org/10.1016/S0098-3004(01)00111-X

[40]. Stumpe B, Marschner B (2024) Rehabilitated Tailing Piles in the Metropolitan Ruhr Area (Germany) Identified as Green Cooling Islands and Explained by K-Mean Cluster and Random Forest Regression Analyses. Remote Sens (Basel) 16:4348

[41]. Santos NL, Gomes M da CR, Dos Anjos JÂSA, Cunha FG (2020) Multivariate statistical analysis applied to assess the dispersion of contaminants in a mining tailings basin in the semiarid region of bahia – brazil. Revista Ambiente e Agua. https://doi.org/10.4136/AMBI-AGUA.2572

[42]. Jin Y, Wakayama T, Jiang R, Sugasawa S (2025) Clustered factor analysis for multivariate spatial data. Spat Stat 66:100889

[43]. Baragilly MH, Gabr H, Willis BH (2023) Clustering Analysis of Multivariate Data: A Weighted Spatial Ranks-Based Approach. J Probab Stat 2023:1–15

[44]. Xiao W, Zhou Z, Ren B, Deng X (2025) Integrating spatial clustering and multi-source geospatial data for comprehensive geological hazard modeling in Hunan Province. Sci Rep 15:1982

[45]. Fouedjio F (2016) A hierarchical clustering method for multivariate geostatistical data. Spat Stat. https://doi.org/10.1016/j.spasta.2016.07.003

[46]. Riquelme ÁI, Ortiz JM (2024) A Riemannian Tool for Clustering of Geo-Spatial Multivariate Data. Math Geosci. https://doi.org/10.1007/s11004-023-10085-7

[47]. Cotrina-Teatino M, Riquelme Á, Marquina J, Mamani-Quispe J, Arango-Retamozo S, Ccatamayo-Barrios J, Donaires-Flores T, Calla-Huayapa M, González-Vásquez J (2025) KMeans-Riemannian model for classification mineral resources in a copper deposit in Peru. International Journal of Mining, Reclamation and Environment. https://doi.org/10.1080/17480930.2025.2518987.

[48]. Martin R, Boisvert J (2018) Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics. Comput Geosci. https://doi.org/10.1016/j.cageo.2018.08.005

[49]. Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: Problems and possibilities. Applied Geochemistry. https://doi.org/10.1016/j.apgeochem.2008.03.004

[50]. 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. https://doi.org/10.1016/j.chemer.2024.126094

[51]. Sadeghi M, Casey P, Carranza EJM, Lynch EP (2024) Principal components analysis and K-means clustering of till geochemical data: Mapping and targeting of prospective areas for lithium exploration in Västernorrland Region, Sweden. Ore Geol Rev 167:106002

[52]. Jansson NF, Allen RL, Skogsmo G, Tavakoli S (2022) Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden. J Geochem Explor. https://doi.org/10.1016/j.gexplo.2021.106909

[53]. Morales González-Moro Á, D'Auria L, Pérez Rodríguez NM (2025) Genetic K-Means Clustering of Soil Gas Anomalies for High-Enthalpy Geothermal Prospecting: A Multivariate Approach from Southern Tenerife, Canary Islands. Geosciences (Basel) 15:204

[54]. Ellefsen KJ, Smith DB (2016) Manual hierarchical clustering of regional geochemical data using a Bayesian finite mixture model. Applied Geochemistry. https://doi.org/10.1016/j.apgeochem.2016.05.016

[55]. Ghanbari Y, Hezarkhani A, Ataei M, Pazand K (2010) Regional geochemical pattern recognition with multivariate correspondence cluster analysis in the Ravar area, Iran. Transactions of the Institutions of Mining and Metallurgy, Section B: Applied Earth Science. https://doi.org/10.1179/1743275811Y.0000000014

[56]. Moreira G de C, Coimbra Leite Costa JF, Marques DM (2020) Defining geologic domains using cluster analysis and indicator correlograms: a phosphate-titanium case study. Applied Earth Science: Transactions of the Institute of Mining and Metallurgy. https://doi.org/10.1080/25726838.2020.1814483

[57]. Erikstad L, Bakkestuen V, Dahl R, Arntsen ML, Margreth A, Angvik TL, Wickström L (2022) Multivariate Analysis of Geological Data for Regional Studies of Geodiversity. Resources. https://doi.org/10.3390/resources11060051

[58]. Hoseinzade Z, Bazoobandi MH (2024) Deep embedded clustering: Delineating multivariate geochemical anomalies in the Feizabad region. Geochemistry 84:126208

[59]. Tokuda EK, Comin CH, Costa L da F (2022) Revisiting agglomerative clustering. Physica A: Statistical Mechanics and its Applications. https://doi.org/10.1016/j.physa.2021.126433

[60]. Li T, Rezaeipanah A, Tag El Din ESM (2022) An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2022.04.010

[61]. Marquina-Araujo JJ, Cotrina-Teatino MA, Cruz-Galvez JA, Noriega-Vidal EM, Vega-Gonzalez JA (2024) Application of Autoencoders Neural Network and K-Means Clustering for the Definition of Geostatistical Estimation Domains. Mathematical Modelling of Engineering Problems 11:1207–1218

[62]. Martin R, Boisvert J (2020) Performance of clustering for the decision of stationarity; A case study with a nickel laterite deposit. Comput Geosci. https://doi.org/10.1016/j.cageo.2020.104565

[63]. Moreira G de C, Modena RCC, Costa JFCL, Marques DM (2021) A workflow for defining geological domains using machine learning and geostatistics. Tecnol Metal Mater Min. https://doi.org/10.4322/2176-1523.20212472

[64]. Carlotto V, Quispe J, Acosta H, Rodríguez R, Romero D, Cerpa L, Mamani M, Díaz-Martínez E, Navarro P, Jaimes F, Velarde T, Lu S, Cueva E (2009) Geotectonic domain as tool for metallogenetic mapping in Peru. Sociedad Geológica del Perú.

[65]. Barachant A, Bonnet S, Congedo M, Jutten C (2013) Classification of covariance matrices using a Riemannian-based kernel for BCI applications. Neurocomputing. https://doi.org/10.1016/j.neucom.2012.12.039

[66]. Barachant A, Bonnet S, Congedo M, Jutten C (2010) Riemannian geometry applied to BCI classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-642-15995-4_78

[67]. Märzinger T, Kotík J, Pfeifer C (2021) Application of hierarchical agglomerative clustering (Hac) for systemic classification of pop-up housing (puh)

environments. Applied Sciences (Switzerland). https://doi.org/10.3390/app112311122

[68]. Raju VNG, Lakshmi KP, Jain VM, Kalidindi A, Padma V (2020) Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020. https://doi.org/10.1109/ICSSIT48917.2020.9214160

[69]. Corcoran L, Simonetti A, Spano TL, Lewis SR, Dorais C, Simonetti S, Burns PC (2019) Multivariate analysis based on geochemical, isotopic, and mineralogical compositions of uranium-rich samples. Minerals. https://doi.org/10.3390/min9090537

[70]. Li P, Wang Q, Zeng H, Zhang L (2017) Local Log-Euclidean Multivariate Gaussian Descriptor and Its Application to Image Classification. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2016.2560816

[71]. Congedo M, Barachant A, Bhatia R (2017) Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. Brain-Computer Interfaces. https://doi.org/10.1080/2326263X.2017.1297192

[72]. Cotrina-Teatino MA, Marquina-Araujo JJ, Riquelme ÁI (2025) Comparison of Machine Learning Techniques for Mineral Resource Categorization in a Copper Deposit in Peru. Natural Resources Research. https://doi.org/10.1007/s11053-025-10505-x

[73]. Cotrina M, Marquina J, Mamani J (2025) Application of artificial neural networks for the categorization of mineral resources in a copper deposit in Peru. World Journal of Engineering. https://doi.org/https://doi.org/10.1108/WJE-01-2025-0004

[74]. Cotrina M, Marquina J, Mamani J, Arango S, Ccatamayo J, Gonzalez J, Donaires T, Calla M (2025) Categorization of Mineral Resources using Random Forest Model in a Copper Deposit in Peru. Journal of Mining and Environmental 16:947–962

[75]. Shi N, Liu X, Guan Y (2010) Research on k-means clustering algorithm: An improved k-means clustering algorithm. 3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010. https://doi.org/10.1109/IITSI.2010.74

[76]. Tarigan DA (2023) Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification. Media Online) 4:

[77]. Shang M, Li H, Ahmad A, Ahmad W, Ostrowski KA, Aslam F, Joyklad P, Majka TM (2022) Predicting the Mechanical Properties of RCA-Based Concrete Using Supervised Machine Learning Algorithms. Materials. https://doi.org/10.3390/ma15020647

[78]. Waller LA (2012) Detection of Clustering in Spatial Data. The SAGE Handbook of Spatial Analysis. https://doi.org/10.4135/9780857020130.n16

[79]. Batool F, Hennig C (2021) Clustering with the Average Silhouette Width. *Comput Stat Data Anal*. https://doi.org/10.1016/j.csda.2021.107190

[80]. Massing T (2021) Clustering Using Student t Mixture Copulas. *SN Comput Sci*. https://doi.org/10.1007/s42979-021-00503-0

[81]. Vattani A (2011) k-means Requires Exponentially Many Iterations Even in the Plane. *Discrete Comput Geom*. https://doi.org/10.1007/s00454-011-9340-1

[82]. Supajaidee N, Chutsagulprom N, Moonchai S (2024) An Adaptive Moving Window Kriging Based on K-Means Clustering for Spatial Interpolation. Algorithms. https://doi.org/10.3390/a17020057

[83]. Goh A, Vidal R (2008) Clustering and dimensionality reduction on Riemannian manifolds. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR. https://doi.org/10.1109/CVPR.2008.4587422

[84]. Holopainen I, Rickman S (1993) Classification of Riemannian manifolds in nonlinear potential theory. *Potential Analysis*. https://doi.org/10.1007/BF01047672

[85]. Viroli C, McLachlan GJ (2019) Deep Gaussian mixture models. Stat Comput. https://doi.org/10.1007/s11222-017-9793-z

[86]. Chassagnol B, Bichat A, Boudjeniba C, Wuillemin PH, Guedj M, Gohel D, Nuel G, Becht E (2023) Gaussian Mixture Models in R. R Journal.

[87]. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*.

[88]. Shutaywi M, Kachouie NN (2021) Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*.

[89]. Lima SP, Cruz MD (2020) A genetic algorithm using Calinski-Harabasz index for automatic clustering problem. *Revista Brasileira de Computação Aplicada 12*:97–106

انجمن مهندسی معدن ایران
نشریه مهندسی معدن و محیط زیست
نشانی نشریه: www.jme.shahroodut.ac.ir
دانشگاه صنعتی شاهرود

مقایسه روش‌های خوشه‌بندی چند متغیره بدون نظارت برای توصیف ژئوشیمیایی و مکانی باطله‌های معدنی

مارکو آنتونیو کوترینا-تئاتینو[1*]، جایرو جوناتان مارکینا-آرائوخو[1]، خوزه نستور مامانی-کویسپه[2]، سولیو مارینو آرانگو-رتاموزو[1] و جو الکسیس گونزالس-واسکوئز[3]

۱. گروه مهندسی معدن، دانشکده مهندسی، دانشگاه ملی تروخیو، تروخیو، پرو
۲ .دانشکده مهندسی شیمی، دانشگاه ملی آلتیپلانو پونو، پونو، پرو
۳. گروه مهندسی صنایع، دانشکده مهندسی، دانشگاه ملی تروخیو، تروخیو، پرو

اطلاعات مقاله

تاریخ ارسال: ۲۰۲۵/۰۷/۲۵
تاریخ داوری: ۲۰۲۵/۰۹/۱۲
تاریخ پذیرش: ۲۰۲۵/۱۰/۰۴

DOI:10.22044/jme.2025.16568.3239

کلمات کلیدی

باطله‌های معدنی
خوشه‌بندی چند متغیره
میانگین‌های K ریمانی
مشخصات ژئوشیمیایی

چکیده

توصیف ژئوشیمیایی و فضایی باطله‌های معدنی قدیمی برای شناسایی فرصت‌های بازفرآوری و اطلاع‌رسانی در مورد مدیریت زیست‌محیطی ضروری است. با این حال، پیچیدگی بالای ترکیبی باطله‌های چندفلزی نیازمند رویکردهای چند متغیره قوی است. این مطالعه عملکرد چهار الگوریتم خوشه‌بندی بدون نظارت K-Means اقلیدسی، K-Means ریمانی، مدل مخلوط گاوسی (GMM) و خوشه‌بندی تجمعی را که بر روی ۹۲۷ نمونه از کانسار باطله‌های کوئیولاکوچا در پرو اعمال شده است، با استفاده از شش عنصر اصلی (روی، سرب، مس، آهن، نقره، طلا) و مختصات مکانی ارزیابی و مقایسه می‌کند. همه روش‌ها به طور مداوم سه حوزه ژئوشیمیایی اصلی را شناسایی کردند. خوشه ۱ از نظر مس و طلا، خوشه ۲ از نظر سرب و آهن و خوشه ۳ از نظر روی، نقره و آهن غنی شده بود. روش‌های مبتنی بر کوواریانس (K-Means ریمانی و خوشه‌بندی تجمعی) در اعتبارسنجی داخلی (نمرات سیلوئت تا ۰.۵۸) و سازگاری (شاخص رند تعدیل‌شده = ۱.۰۰) از سایر روش‌ها بهتر عمل کردند و پارتیشن‌های قابل تفسیرتر و از نظر زمین‌شناسی منسجم‌تری ارائه دادند. تبدیل CLR عملکرد خوشه‌بندی را کاهش داد و اهمیت حفظ واریانس ژئوشیمیایی خام را برای تقسیم‌بندی مکانی برجسته کرد. این یافته‌ها اثربخشی خوشه‌بندی چند متغیره را برای آشکار کردن ناهمگونی ترکیبی در باطله‌ها و مشخص کردن حوزه‌های ارزش اقتصادی بالقوه نشان می‌دهد. این رویکرد یک چارچوب کمی برای پشتیبانی از تصمیمات بازفرآوری، کاهش ریسک و هدایت تحقیقات آینده در مورد ارزش‌گذاری ضایعات معدنی فراهم می‌کند.

نویسنده مسئول مکاتبات: mcotrinat@unitru.edu.pe (م. کوترینا-تئاتینو)