Ore Geology Reviews 181 (2025) 106591



Contents lists available at ScienceDirect

Ore Geology Reviews

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



Review



# Clustering in geo-data science: Navigating uncertainty to select the most reliable method

Behnam Sadeghi [a,b]

[a] CSIRO Mineral Resources, Australian Resources Research Centre, Kensington, WA 6151, Australia
[b] Earth and Sustainability Science Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052, Australia

ARTICLE INFO

*Keywords:*
Clustering
Geo-data science
Mineral exploration
Uncertainty
Decision-making

ABSTRACT

Clustering is a fundamental technique in unsupervised learning that groups data points based on their similarities, enabling the discovery of underlying structures and patterns without the need for labeled examples. In geo-data science, clustering plays a pivotal role in applications such as exploratory data analysis, geochemical anomaly detection, and dimensionality reduction for multivariate data analysis. Mineral exploration, in particular, involves various uncertainties that significantly affect decision-making processes. A key source of such uncertainties arises from the choice of clustering methods and their evaluation metrics. Among the widely used clustering methods, K-Means, K-Medoids, Silhouette and Hierarchical Clustering are prominent, with K-Means being a more popular choice in geosciences. Similarly, evaluation techniques such as Silhouette, Davies-Bouldin, Calinski-Harabasz, Elbow, and Bayesian Information Criterion are employed, with the Elbow method being a frequent favorite in geo-data science. However, questions remain regarding the efficiency and suitability of these methods in different contexts. Should we rely solely on K-Means and Elbow, or should we adopt a more case-specific approach, comparing uncertainties and selecting methods that minimize them? This research provides a critical review of clustering methods and evaluation metrics in geo-data science. By developing the *pyClusterWise* Python package and using illustrative examples, the importance of tailoring these adjustments to the data and selecting clustering techniques and evaluation metrics based on their associated uncertainties is demonstrated. By doing so, the aim is to reduce overall uncertainty and improve decision-making outcomes in mineral exploration.

## 1. Introduction

Clustering is a fundamental technique in unsupervised learning, aimed at grouping data points based on their inherent similarities (Kaufman and Rousseeuw, 1990; Templ et al., 2008; Sadeghi et al., 2024). Unlike supervised learning, which depends on labeled data, clustering allows us to uncover hidden patterns and structures in datasets without relying on predefined categories. This capability to identify meaningful groups is crucial across many disciplines, including market segmentation, image recognition, and especially in geo-data science, where datasets are often complex and involve multiple variables (Jain et al., 1999; Xu and Tian, 2015).

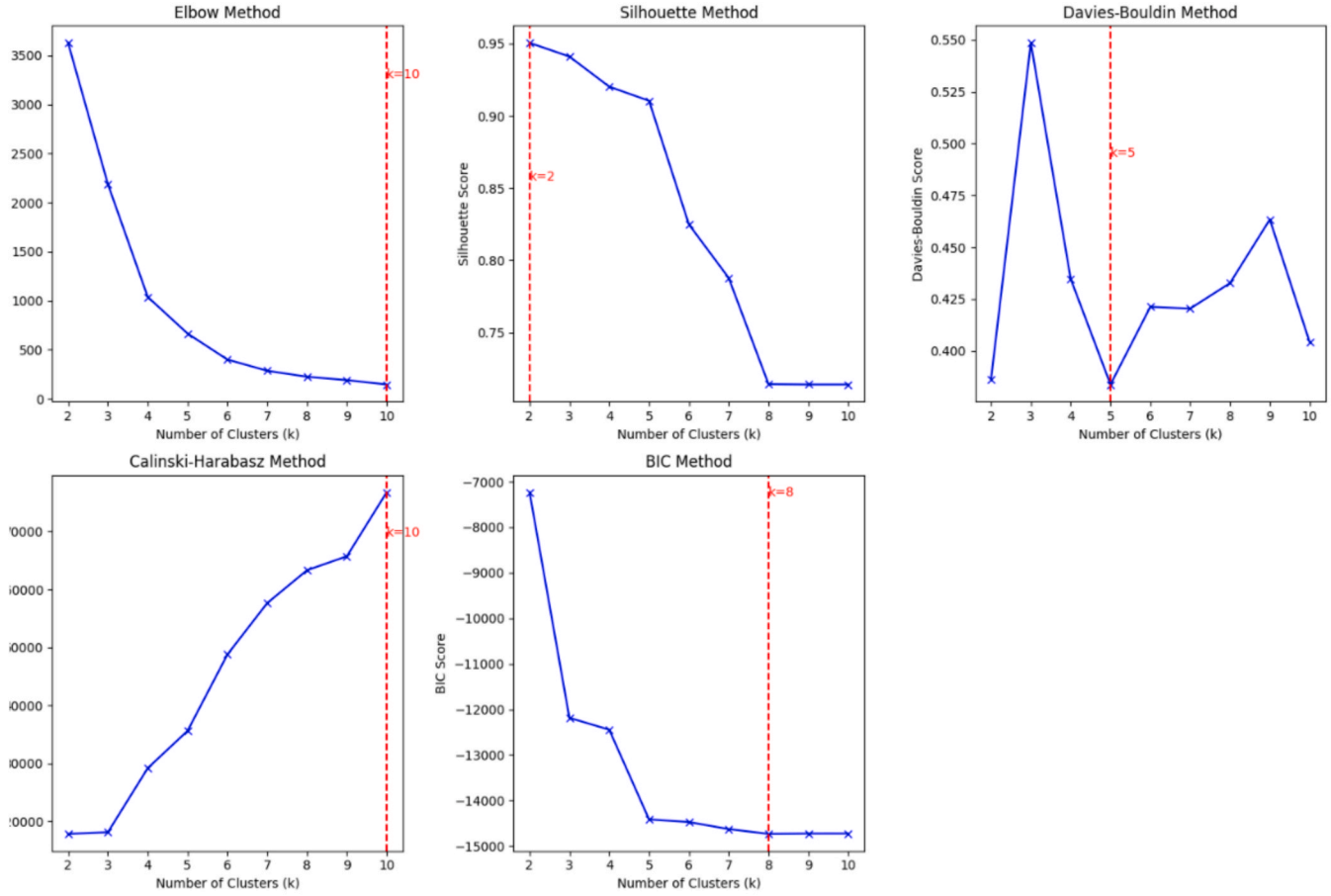
In geo-data science, clustering is essential for tasks like exploratory data analysis, geochemical anomaly detection, and dimensionality reduction (Sadeghi et al., 2024). For example, in geochemical studies, clustering can pinpoint compositional anomalies that may signal mineral deposits or other geological features (Reimann et al., 2008). When analyzing multivariate data, clustering simplifies datasets by reducing dimensionality, maintaining key structures while streamlining further modeling. These applications demonstrate how clustering extracts valuable insights from noisy, heterogeneous datasets, making it particularly useful in areas like mineral exploration.

Mineral exploration poses distinct challenges, particularly due to the high levels of uncertainty arising from sparse and incomplete data (see Caers, 2011; Scheidt et al., 2018; Sadeghi et al., 2023; Wang et al., 2023). Decision-making in this field often hinges on clustering results, which can vary significantly depending on the methods and evaluation metrics used. An inappropriate clustering approach might overlook the actual underlying patterns, resulting in misleading interpretations and suboptimal decisions. Therefore, carefully selecting suitable clustering techniques and evaluation metrics is crucial to reducing uncertainty and improving the reliability of results (McKinley et al., 2016).

K-Means and K-Medoids are among the most commonly used clustering methods in geo-data science, valued for their simplicity and

*E-mail addresses:* behnam.sadeghi@csiro.au, z5218858@zmail.unsw.edu.au.

https://doi.org/10.1016/j.oregeorev.2025.106591
Received 20 December 2024; Received in revised form 24 March 2025; Accepted 29 March 2025
Available online 15 April 2025
0169-1368/Crown Copyright © 2025 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).




A

**Fig. 1.** The plots and *k* values obtained by different clustering methods applied to the A) Perth and B) Upper North Yilgarn hydrogeochemical data.

computational efficiency. K-Means, in particular, is popular in geosciences for its ability to handle large datasets efficiently (Lloyd, 1982; MacQueen, 1967, Grunsky, 2010; Jansson et al., 2022, Soltani and Imamalipour, 2022, Cardoso-Fernandes et al., 2022; Goncalves et al., 2024). However, its dependence on Euclidean distances and sensitivity to initial conditions can sometimes result in suboptimal clustering outcomes. In contrast, K-Medoids employs a more robust distance metric, making it less affected by outliers, though at the cost of higher computational demands. These trade-offs underscore the need to carefully align the choice of clustering method with the specific characteristics of the dataset and the problem being addressed.

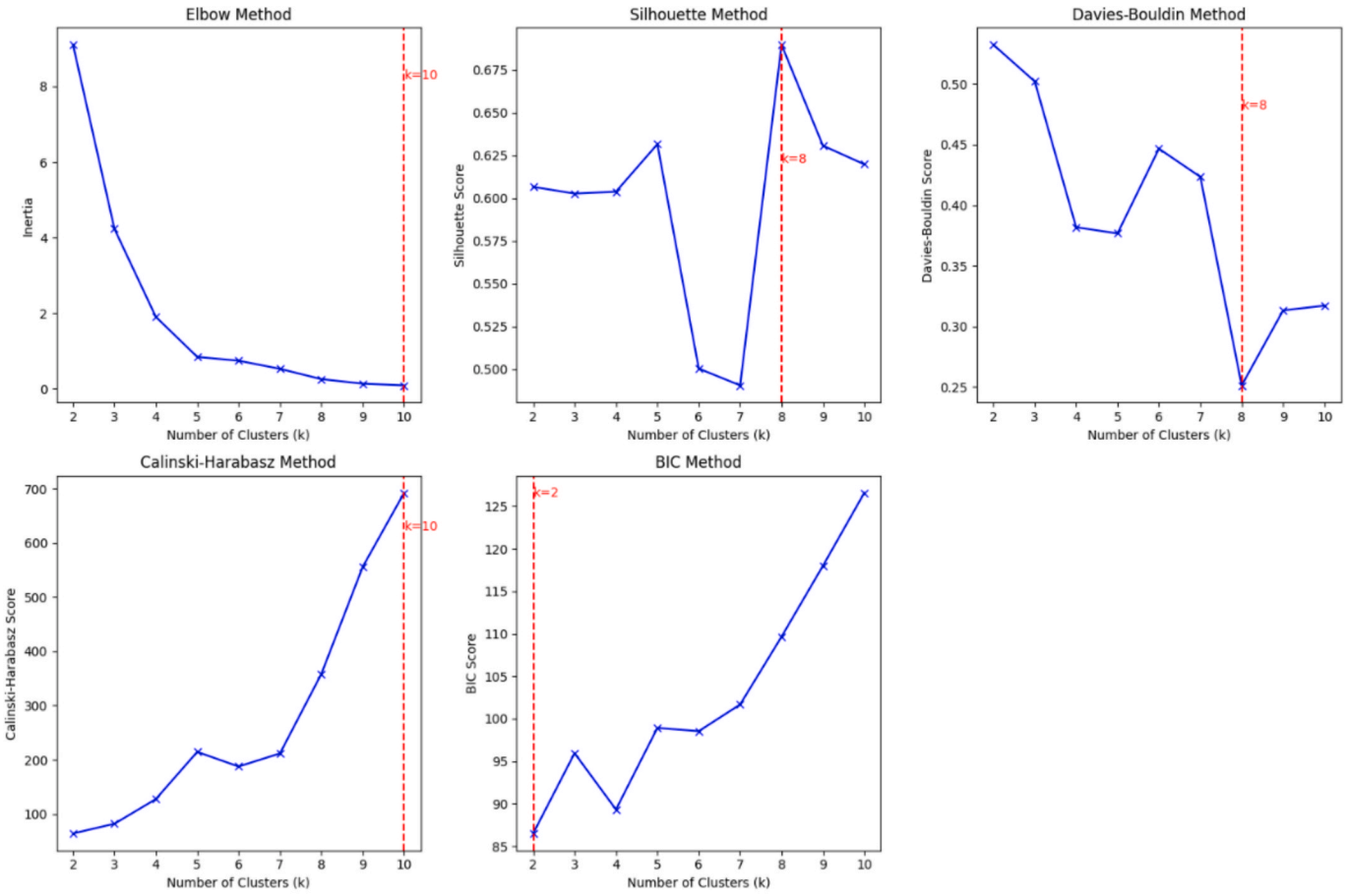
Evaluating clustering results is a critical yet often underemphasized step in the clustering workflow. Metrics such as Silhouette Index (SI), Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Elbow (Thorndike, 1953, Kaiser, 1992), and Bayesian Information Criterion (BIC) offer diverse approaches to assess clustering quality. In geo-data science, the Elbow method is particularly popular for its simplicity and interpretability. However, relying solely on a single evaluation metric—especially one that may not account for the unique complexities of geo-data—can introduce considerable uncertainty into the analysis. This raises a key question: should we continue relying on conventional methods like K-Means and the Elbow method, or should we adopt a more tailored and nuanced approach that considers uncertainties and adapts methods to specific scenarios?

Although low uncertainty is a desirable property indicating that the method produces consistent and stable results across different realizations, other factors must be considered to determine its suitability for a specific geoscientific application.

In geosciences, we also need to consider the following criteria to select the most appropriate clustering method: I) alignment with the objectives (Jain, 2010; Hastie et al., 2009); II) data characteristics such as datasets with high-dimensional features, nonlinearity, or noise. In such cases, a method with low uncertainty but poor adaptability to the data structure may end up with results that are not geologically meaningful (Xu & Tian, 2015); III) interpretability and geological relevance (Lindsay et al., 2019); and IV) uncertainty should be evaluated by metrics that are discussed in continue.

Therefore, in this study, uncertainty quantification was used as one of the key criteria to evaluate the robustness of clustering methods, but it was not the sole factor in determining the suitability of a method. Final selection involved balancing low uncertainty with high clustering validity. This study seeks to address this pivotal question by presenting a comprehensive review of clustering methods and evaluation metrics applied in geo-data science. Using real-world examples from mineral exploration, we demonstrate how uncertainties inherent in different clustering approaches can impact decision-making. This analysis emphasizes the necessity of choosing clustering techniques and evaluation metrics that effectively reduce uncertainty, enhancing the reliability and interpretability of results.

The ultimate goal of this research is to advocate for a more informed and context-specific approach to clustering in geo-data science. By incorporating uncertainty analysis into the selection process, we aim to enhance decision-making frameworks in mineral exploration and related fields. This perspective is especially relevant in light of the




B

**Fig. 1.** (*continued*).

increasing complexity of datasets and the pressing demand for robust analytical methods in the geosciences.

## 2. Methodology

### 2.1. Clustering methods

#### 2.1.1. K-Means clustering method

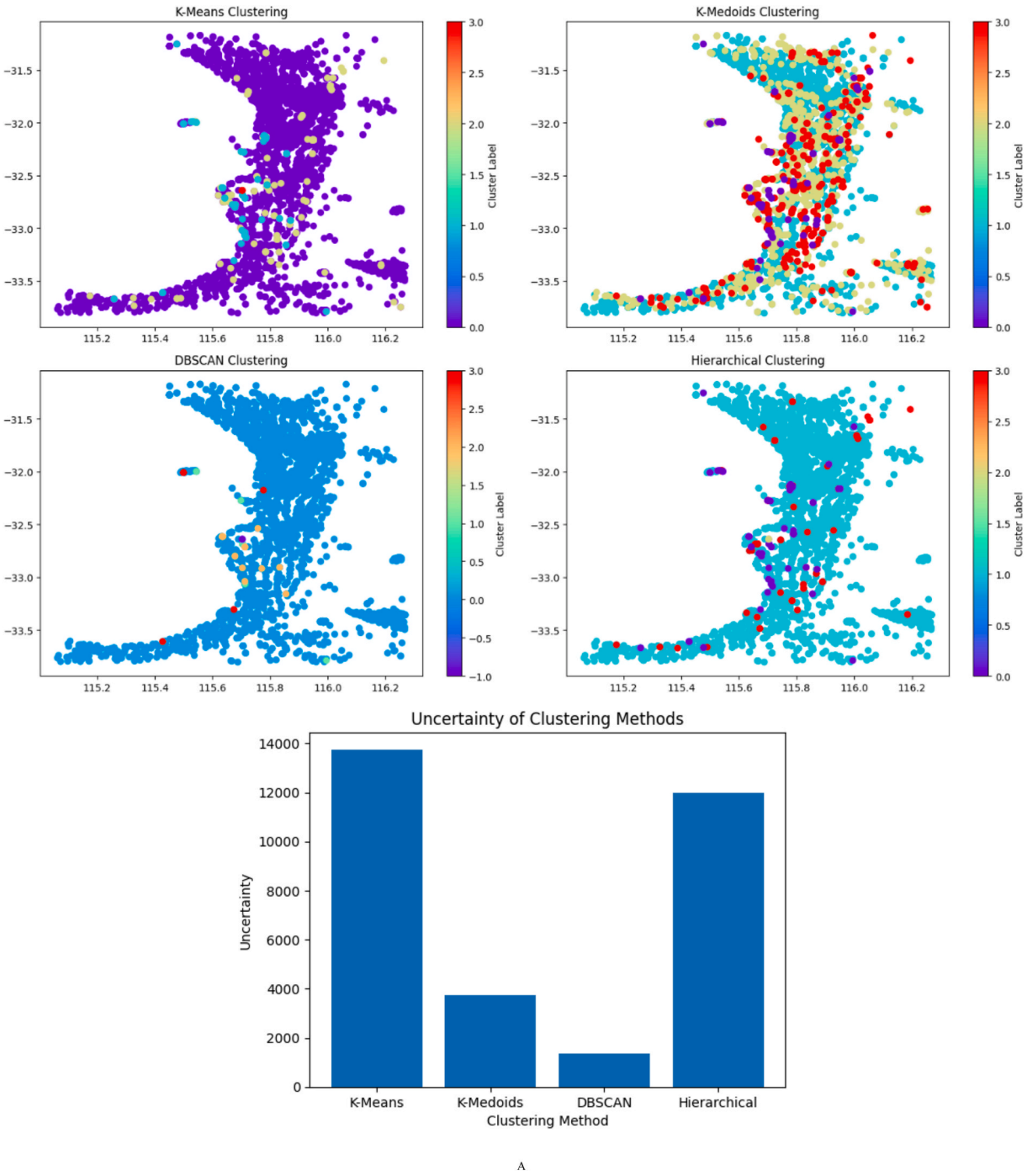
K-Means is among the most widely used clustering algorithms in unsupervised learning. It divides data into *k* clusters by minimizing the sum of squared distances between data points and the centroids of their respective clusters (Lloyd, 1982). The algorithm operates iteratively, assigning data points to the nearest centroid, recalculating centroids as the mean of the assigned points, and repeating this process until convergence. Its simplicity, speed, and scalability make K-Means well-suited for large datasets. However, it assumes spherical clusters, making it less effective for irregularly shaped data, and it is sensitive to outliers and the initial placement of centroids.

K-Means has been extensively applied in geosciences, particularly in areas such as exploratory data analysis, mineral exploration, geochemical anomaly detection, geophysical analysis, climatology, and satellite imagery (cf. Unal et al., 2003; Richards and Jia, 2006; Reimann et al., 2008; Lima et al., 2013; Di and Gao, 2016; Bigdeli et al., 2022; Sadeghi et al., 2024; Goncalves et al., 2024; Zhao et al., 2024). In mineral exploration, K-Means clustering is used to integrate various geospatial datasets, including geophysical, geochemical, and remote sensing data, to produce prospectivity maps. These maps identify areas with a high potential for hosting mineral deposits, thereby enhancing the efficiency of exploration efforts (Carranza, 2009). However, to maintain focus and brevity, this paper emphasizes its applications in geochemical analysis. Since K-Means is widely used for classifying geochemical data, it enables the identification of distinct compositional groups within multivariate datasets. This application is particularly important for detecting geochemical anomalies linked to mineralization processes (Reimann et al., 2008). For example, clustering can highlight areas with elevated concentrations of economically significant elements, providing valuable insights for mineral exploration.

While K-Means is a powerful tool, its assumptions of spherical clusters and equal cluster sizes often do not reflect the natural variability found in geoscientific data (Hajihosseinlou et al., 2024a). Furthermore, its sensitivity to the initial placement of centroids can result in suboptimal clustering outcomes, making it necessary to perform multiple runs or consider more robust alternatives, such as K-Medoids or hierarchical clustering.

#### 2.1.2. K-Medoids clustering method

K-Medoids, also known as Partitioning Around Medoids (PAM), is a clustering algorithm similar to K-Means but more robust to outliers and noise (Abdel-Basset et al., 2024; Heidari et al., 2024). Instead of computing centroids as the mean of data points, K-Medoids selects actual data points (medoids) as cluster centers. The algorithm minimizes the sum of dissimilarities between data points and their respective medoids. This approach makes K-Medoids less sensitive to extreme values and better suited for datasets with irregular structures or when using non-Euclidean distance measures (Kaufman and Rousseeuw, 1990).

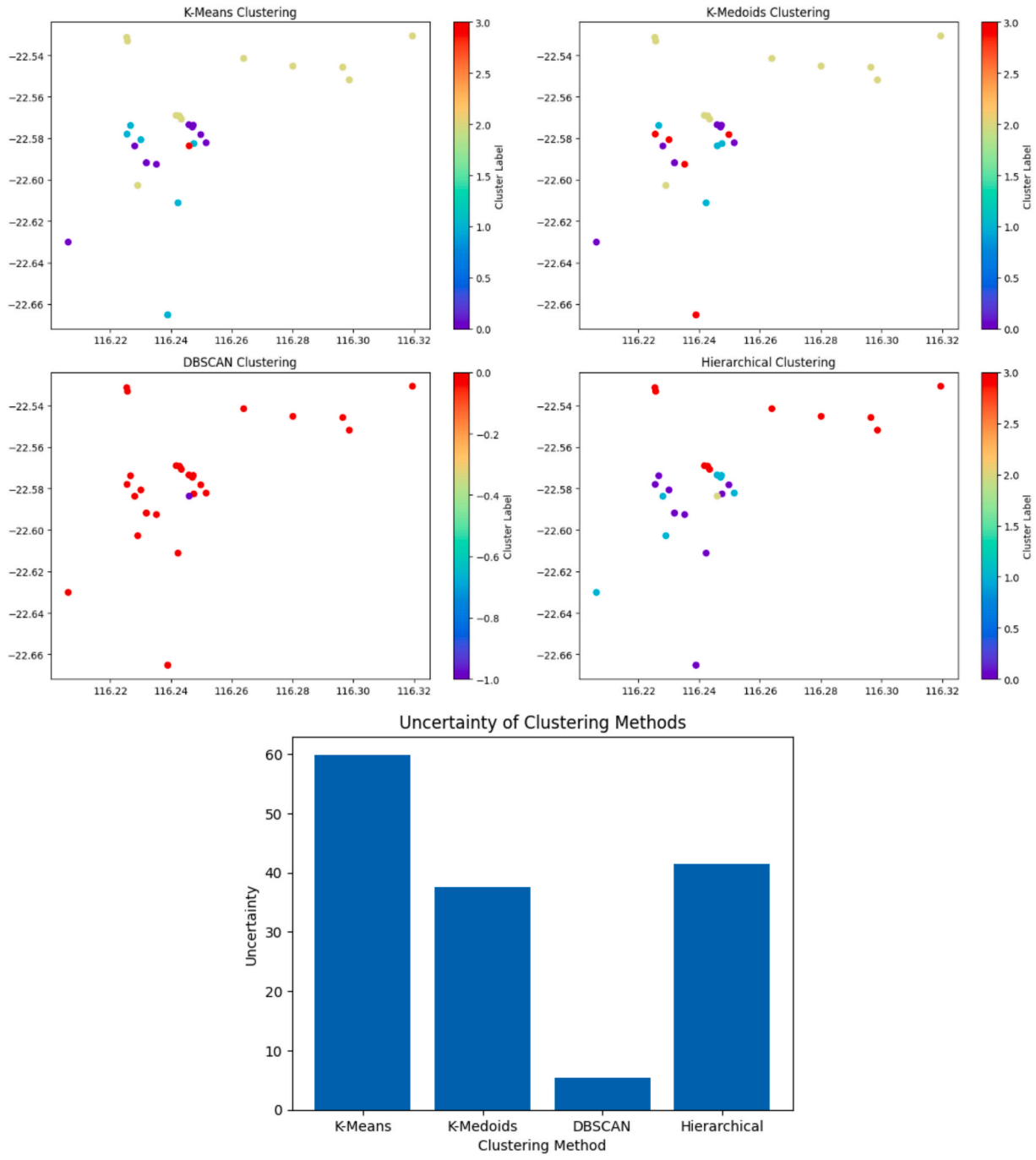



A

**Fig. 2.** The plots of the clustering methods applied to A) the Perth and B) Upper North Yilgarn hydrogeochemical data.

K-Medoids is particularly valuable in geosciences when dealing with datasets that include noise, outliers, or non-spherical clusters (cf. Mazor, 2004; Richards and Jia, 2006; Reimann et al., 2008; Carranza, 2009; Birks et al., 2012; Di and Gao, 2016; Sureja et al., 2022; Zhang and Shang, 2022; Li and Wang, 2023; Chen et al., 2023). Its robustness makes it well-suited for various applications. For instance, K-Medoids has been applied to detect geochemical anomalies by grouping data points into stable clusters, even in the presence of outliers. This capability is crucial for identifying anomalous geochemical signatures that may indicate mineralization (Reimann et al., 2008). Additionally, K-Medoids has been employed to integrate geospatial datasets for mineral prospectivity analysis. By clustering data from geochemical, geophysical, and remote sensing sources, it aids in pinpointing regions with potential mineral deposits (Carranza, 2009).

Although K-Medoids is more robust than K-Means, it is computationally more demanding, particularly when dealing with large datasets. The algorithm involves calculating pairwise distances for all data points, which can be a significant bottleneck in geosciences, where datasets are often extensive. This limitation, however, can be addressed using faster variations such as CLARA (Clustering Large Applications), which approximates the results of K-Medoids by working with subsets of the data (Martins et al., 2024).




B

**Fig. 2.** (*continued*).

### 2.1.3. DBSCAN (density-based spatial clustering of applications with Noise)

DBSCAN is a density-based clustering algorithm introduced by Ester et al. (1996). It identifies clusters by examining the density of points in a given region. The algorithm relies on two key parameters: epsilon ($\varepsilon$), which defines the maximum distance between points to be considered part of the same cluster, and *minPts*, the minimum number of points required to form a dense region. The algorithm categorizes points as: I) core points, that have at least *minPts* neighbors within the $\varepsilon$ radius; II) border points, which are within $\varepsilon$ of a core point but do not meet the density threshold themselves; and III) noise points, which are neither core nor border points (see Cheng et al., 2024; Hajihosseinlou et al., 2024b; Kim et al., 2024; Mardani et al., 2025).

DBSCAN is especially effective for datasets containing noise and clusters of varying densities. Its ability to identify clusters of arbitrary shapes and distinguish outliers makes it particularly well-suited for spatial data analysis. Unlike traditional methods like K-Means, DBSCAN excels in detecting non-convex clusters and managing datasets with significant noise, addressing challenges that many other clustering algorithms cannot handle effectively.

### 2.1.4. Hierarchical clustering

Hierarchical clustering is a method that creates a hierarchy of clusters using either a bottom-up (agglomerative) or top-down (divisive)

**Table 1**
The applied methods comparison table for the Perth and Upper North Yilgarn hydrogeochemical data.

| **Perth hydrogeochemical data** | | | | |
|---|---|---|---|---|
| Metric | K-Means | K-Medoids | DBSCAN | Hierarchical |
| SI | 0.9201 | 00.614 | 0.940 | 0.916 |
| DBI | 0.435 | 0.595 | 0.301 | 0.415 |
| CBI | 29171.982 | 7956.527 | 2862.606 | 25387.334 |
| Elbow | 663.906 | NaN | NaN | NaN |
| BIC | −12445.149 | NaN | NaN | NaN |
| **Upper North Yilgarn hydrogeochemical data** | | | | |
| Metric | K-Means | K-Medoids | DBSCAN | Hierarchical |
| SI | 0.604 | 0.528 | 0.586 | 0.561 |
| DBI | 0.382 | 0.594 | 0.2438 | 0.421 |
| CBI | 127.549 | 80.165 | 11.691 | 88.543 |
| Elbow | 0.849 | NaN | NaN | NaN |
| BIC | 89.318 | NaN | NaN | NaN |

**Table 2**
The uncertainty of all clustering methods for the Perth and Upper North Yilgarn hydrogeochemical data.

| Clustering method | Uncertainty | |
|---|---|---|
| | Perth | Upper North Yilgarn |
| K-Means | 13751.485 | 59.895 |
| K-Medoids | 3750.458 | 37.526 |
| DBSCAN | 1349.153 | 5.317 |
| Hierarchical | 11967.390 | 41.508 |

approach. In agglomerative clustering, each data point starts as its own cluster, and the algorithm iteratively merges the two closest clusters until only one cluster remains or the desired number of clusters is reached. Conversely, divisive clustering begins with all data points in a single cluster and recursively splits them into smaller clusters. The process relies on a distance or similarity measure, such as Euclidean distance, along with linkage criteria that determine how clusters are merged or divided (cf. Davis and Sampson, 2002; Güler et al., 2002; Cloutier et al., 2008; Yidana, 2010, Yidana et al., 2010; Mukherjee et al., 2012; Andrade and Stigter, 2013; Jiang et al., 2015; Eminagaoglu et al., 2022; Ebrahimi et al., 2022; Zhong et al., 2022; Ngombe et al., 2024). Common linkage methods include single linkage, which considers the minimum distance between points in different clusters; complete linkage, which uses the maximum distance; and average linkage, which calculates the average distance between all points in two clusters. This flexibility allows hierarchical clustering to adapt to diverse datasets and clustering objectives.

Hierarchical clustering produces a dendrogram, a tree-like diagram that visualizes the formation of clusters and allows users to identify the optimal number of clusters by "cutting" the tree at a chosen level. While it offers significant interpretability and flexibility, hierarchical clustering can face challenges when applied to large datasets due to its high computational complexity and sensitivity to noise, which can affect the quality of the results.

Despite their utility, clustering methods like K-Means and DBSCAN come with inherent limitations. K-Means assumes clusters are spherical and evenly distributed, which may not align with the irregular and anisotropic nature of geological data. Similarly, DBSCAN's performance depends on the careful selection of its parameters, epsilon and *minPts*, which can vary across datasets. To address these challenges, preprocessing techniques such as data transformation (e.g., log scaling for geochemical data) and parameter optimization are essential. By tailoring these adjustments to the data, the reliability of clustering results can be significantly enhanced.

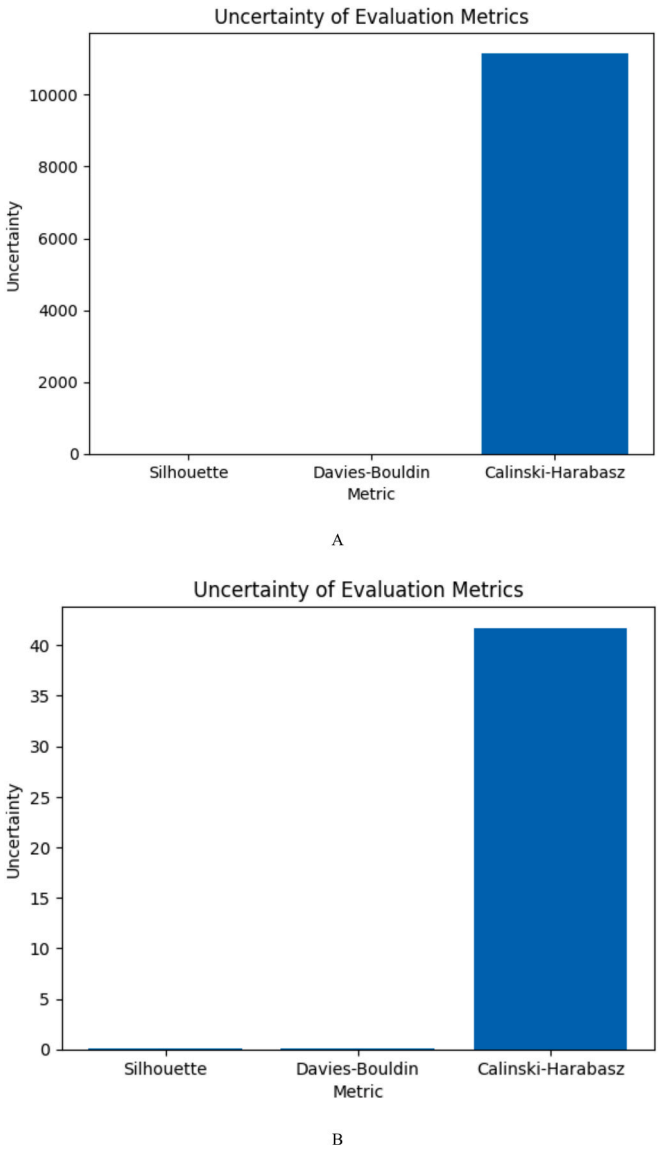



**Fig. 3.** Comparison plot of the uncertainty of the available evaluation metrics for A) the Perth and B) Upper North Yilgarn hydrogeochemical data.

**Table 3**
Uncertainty Scores for the available evaluation metrics for the Perth and Upper North Yilgarn hydrogeochemical data.

| Available evaluation metrics | Uncertainty | |
|---|---|---|
| | Perth | Upper North Yilgarn |
| SI | 0.135 | 0.029 |
| DBI | 0.105 | 0.125 |
| CHI | 11162.848 | 41.725 |

### 2.2. *Evaluation metrics*

#### 2.2.1. *Elbow method*

The Elbow Method is a popular heuristic for determining the optimal number of clusters ($k$) in clustering algorithms like K-Means and K-Medoids. The method involves plotting the within-cluster sum of squares (WCSS) or total intra-cluster variance against the number of clusters. The "elbow" point on the graph represents the stage where adding more clusters results in a diminishing reduction in WCSS, suggesting a balance between model complexity and performance (Ketchen

and Shook, 1996).

To apply the Elbow Method, the clustering algorithm (e.g., K-Means or K-Medoids) is executed for a range of cluster numbers ($k$). For each $k$, the within-cluster sum of squares (WCSS) is calculated, representing the compactness of the clusters. The next step involves plotting WCSS against $k$, and the "elbow" point, where the reduction in WCSS begins to plateau, is identified. This point indicates the optimal number of clusters.

The Elbow Method is widely applied in geosciences to determine the optimal number of clusters for various datasets (see Lima et al., 2013 for geology, Di and Gao, 2016 for seismic analysis, and Richards and Jia, 2006 for remote sensing using satellite images). In geochemistry, it is used to identify distinct geochemical populations or anomalies associated with mineralization (Reimann et al., 2008).

The Elbow Method is one of the most favored approaches in geo-data science due to its simplicity and ease of interpretation. It offers a clear visual representation of the trade-offs between model complexity and cluster compactness (Rylko et al. 2024). However, the method has its limitations. Identifying the exact location of the elbow can be subjective and ambiguous, especially when dealing with noisy or complex data. Additionally, for very large datasets, computing the WCSS for multiple $k$ values can become computationally intensive. When the elbow point is unclear, alternative methods such as the Silhouette Index (SI) and Davies-Bouldin Index (DBI) can provide valuable complementary insights into determining the optimal $k$.

#### 2.2.2. Silhouette index (SI)

The SI is a widely used metric for evaluating the quality of clustering results (Richards and Jia, 2006; Lima et al., 2013; Di and Gao, 2016; Bagirov et al., 2023; Du et al., 2024; Lenssen and Schubert, 2024). It assesses how well each data point is clustered by comparing its cohesion (similarity to other points within the same cluster) to its separation (dissimilarity to points in the nearest neighboring cluster). The Silhouette score ranges from $-1$ to $+1$, with values close to $+1$ indicating well-clustered data points, values near 0 suggesting overlapping clusters, and values near $-1$ implying incorrect clustering (Rousseeuw, 1987; Dudek, 2020).

For a given data point i:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{1}$$

where $a(i)$ represents the mean distance between $i$ and the other points in the same cluster (cohesion), and $b(i)$ represents the mean distance between $i$ and the points in the nearest cluster (separation). The overall Silhouette score is the average of $S(i)$ for all data points.

In geochemical studies, this method is used to evaluate the quality of clustering in geochemical datasets, ensuring that distinct geochemical populations associated with mineralization are well-separated and accurately identified (Reimann et al., 2008).

In mineral exploration, the SI is instrumental in validating the integration of geospatial datasets, ensuring that clusters effectively represent meaningful geological or geochemical anomalies (Carranza, 2009). This metric is versatile, as it can be applied to any clustering algorithm and provides an intuitive measure of how well clusters are defined. However, it has limitations. The computation of the SI is resource-intensive for large datasets, as it requires distance calculations for every point. Additionally, it may struggle to accurately evaluate non-convex clusters or clusters with varying densities.

#### 2.2.3. Davies-Bouldin index (DBI)

The Davies-Bouldin Index (DBI) is a clustering validation metric that measures the quality of clustering results by assessing the separation and compactness of clusters. It provides a single score, where lower values indicate better clustering performance (see Eqn. (2). The DBI is particularly valuable for comparing the effectiveness of different clustering algorithms or evaluating the performance of a single algorithm under varying parameter settings (cf. Davies and Bouldin, 1979; Mazor, 2004; Richards and Jia, 2006; Carranza, 2009; Di and Gao, 2016; Vergani and Binaghi, 2018).

For a clustering result with $k$ clusters:

$$DBI = \frac{1}{\mathrm{k}} \sum_{i=1}^{\mathrm{k}} max\left(\frac{s_i + s_j}{d_{ij}}\right), (j \neq i) \tag{2}$$

where $s_i$ is the average intra-cluster distance (compactness) for cluster $i$, $d_{ij}$ is the distance between the centroids of clusters $i$ and $j$ (separation), and $k$ is the total number of clusters. A lower DBI value indicates more compact and well-separated clusters.

The DBI is applied in geochemical studies to evaluate clustering results, ensuring that compositional anomalies are accurately grouped and that clusters effectively represent distinct geochemical populations (Reimann et al., 2008). In mineral exploration, DBI is used to validate the integration of geospatial datasets, ensuring that the resulting clusters represent meaningful geological anomalies or potential mineralization zones (Carranza, 2009).

This method combines compactness and separation into a single metric, making it applicable to any clustering algorithm and enabling comparisons of clustering results across methods or parameter settings. However, it assumes spherical and evenly sized clusters, which may not align with the true structure of geoscientific data. Additionally, the DBI is sensitive to the choice of distance metric used in the clustering algorithm. As a result, other metrics such as Calinski-Harabasz Index (CHI) can be used alongside it to provide more robust cluster validation.

#### 2.2.4. Calinski-Harabasz index (CHI)

The CHI, also known as the Variance Ratio Criterion, is a clustering evaluation metric that measures clustering quality by comparing the dispersion within clusters to the dispersion between clusters (cf. Mazor, 2004; Richards and Jia, 2006; Reimann et al., 2008; Lima et al., 2013; Di and Gao, 2016). Higher CHI values indicate better-defined and more well-separated clusters, reflecting the effectiveness of the clustering (Calinski and Harabasz, 1974).

$$CH = \frac{trace(B_{\mathrm{k}})/(\mathrm{k} - 1)}{trace(W_{\mathrm{k}})/(n - \mathrm{k})} \tag{3}$$

where $B_k$ is the between-cluster scatter matrix (measures separation between clusters), $W_k$ is the within-cluster scatter matrix (measures compactness of clusters), $k$ is the number of clusters, and $n$ is the total number of the data points. The CHI balances compactness (low intra-cluster variance) and separation (high inter-cluster variance).

The CHI is applied to validate clustering results of geochemical datasets, ensuring that anomalies and distinct geochemical populations are well-separated (Reimann et al., 2008). The CHI promotes the formation of compact clusters with clear separation and is computationally efficient for medium to large datasets. Its versatility allows it to work with any clustering algorithm. However, the CHI assumes that clusters are spherical and evenly distributed, which may not align with the complexity of geoscientific datasets. Additionally, it tends to favor larger datasets and may not perform as effectively with smaller ones. To achieve a more comprehensive evaluation, the CHI can be complemented with other metrics such as SI and DBI.

#### 2.2.5. Bayesian information criterion (BIC)

The Bayesian Information Criterion (BIC) is a widely used model selection criterion for evaluating the quality of statistical models, including clustering algorithms. In clustering, the BIC helps determine the optimal number of clusters by balancing model fit with model complexity, ensuring that the selected model captures the underlying data structure without overfitting (cf. Richards and Jia, 2006; Reimann et al., 2008; Birks et al., 2012; Di and Gao, 2016). BIC addresses overfitting by incorporating a penalty term based on the number of parameters in the model. This penalty favors simpler models, particularly

when multiple models fit the data equally well, ensuring a balance between goodness-of-fit and model complexity (Schwarz, 1978):

$$BIC = -2.\ln(L) + k.\ln(n) \quad (4)$$

where $L$ is the maximized likelihood of the model, $k$ is the number of free parameters in the model, and $n$ is the number of the data points. Lower BIC values indicate a better trade-off between goodness-of-fit and model complexity.

This method is used to determine the optimal number of clusters in geochemical datasets, ensuring that the resulting clusters correspond to meaningful compositional groups. For example, it helps identify geochemical anomalies linked to mineral deposits. In mineral exploration, BIC is also applied to evaluate clustering models that integrate geospatial datasets, ensuring that clusters accurately reflect realistic geological anomalies and potential mineralization zones (Carranza, 2009).

BIC balances model fit with complexity by penalizing overfitting, making it useful for comparing models with different numbers of clusters. It is versatile, being applicable to a wide range of clustering models and likelihood-based approaches. However, BIC assumes a Gaussian distribution of data, which may not always align with the nature of geoscientific datasets. Additionally, computing likelihoods can be resource-intensive for large datasets, and BIC performs best when clusters have regular shapes, such as spherical or elliptical forms. While effective, BIC can be complemented with other metrics like the Akaike Information Criterion (AIC) (Akaike, 1974), SI, or Gap Statistic (Tibshirani et al., 2001) for a more comprehensive evaluation of clustering results.

## 3. Discussion

This study highlights the necessity of selecting the appropriate clustering and evaluation methods based on specific dataset characteristics, rather than relying on a one-size-fits-all approach. Using hydrogeochemical data (here Cl (mg/litre)) as a case study, two datasets (from Western Australia) were analyzed—one from Perth, representing dense sampling with 10,064 samples, and the other from Upper North Yilgarn, featuring sparse sampling with only 30 samples. These contrasting datasets demonstrate how clustering performance and evaluation depend on data distribution and density.

In this research, the emphasis is placed on the uncertainty associated with clustering methods and their evaluation metrics, rather than on the geochemical characteristics or procedures related to the elements themselves. The uncertainty here was calculated by analyzing the variability in the performance of the applied clustering methods across the evaluation metrics. For each metric, the uncertainty is quantified as the absolute difference in scores between the two clustering methods. This difference tells us how consistent (or robust) the clustering solution is across different methods. Additionally, the average uncertainty across these metrics is computed to evaluate the overall stability of the methods. By normalizing these values, the uncertainty is scaled, enabling a clear comparison between metrics, which helps determine the most robust metric for assessing clustering performance. If the scores are similar, the clustering is considered stable (i.e., low uncertainty); if they differ widely, it indicates higher uncertainty or sensitivity to the method used. This approach highlights the broader applicability and reliability of the methods, irrespective of the elements under investigation.

To address these challenges, the custom Python package *pyClusterWise* was developed, capable of applying multiple clustering algorithms—including K-Means, K-Medoids, DBSCAN, and Hierarchical Clustering—and evaluating their results using metrics such as the SI, DBI, CHI, the Elbow Method, and Bayesian Information Criterion (BIC). The package quantifies uncertainties associated with each method, ensuring reliable recommendations for optimal clustering and evaluation metrics.

For both datasets, *pyClusterWise* systematically calculated the optimal number of clusters ($k$) using the evaluation metrics, applied the clustering algorithms, and assessed the results (Fig. 1). Across the examples, DBSCAN consistently emerged as the most reliable clustering method due to its ability to handle irregularly shaped clusters and noise, particularly in the sparse dataset where other methods faltered (Fig. 2; Tables 1 and 2).

In terms of evaluation metrics, DBI proved to be the most reliable metric for the dense Perth dataset, effectively balancing cluster compactness and separation. Meanwhile, SI demonstrated its robustness in the sparse Upper North Yilgarn dataset, providing clear differentiation among clusters despite the limited sample size (Fig. 3 and Table 3).

In the figures, each bar represents the value of one of these metrics for either K-Means or K-Medoids. For example, the Silhouette Score reflects how well-separated the clusters are, while the Davies-Bouldin Index shows the average similarity between clusters (with lower values being better). The Calinski-Harabasz Score indicates the ratio of between-cluster to within-cluster dispersion, and the Elbow method provides the relative change in inertia, which helps determine the optimal number of clusters. The BIC score balances model fit and complexity.

The reliability of evaluation metrics such as DBI and SI is context-dependent, influenced by the clustering algorithm and the underlying data characteristics. For example, DBI excels in datasets with compact, well-separated clusters but may perform inconsistently in cases of high overlap or irregular shapes. Similarly, while SI provides robust insights for sparse datasets, its reliance on distance calculations makes it less effective for dense datasets with subtle variations. This underscores the importance of multi-metric evaluation, as relying on a single metric may obscure critical nuances in clustering quality.

These findings underscore the importance of tailoring clustering approaches to the specific characteristics of each dataset. While dense datasets like the Perth example benefit from the scalability and simplicity of algorithms such as K-Means, sparse datasets demand methods that can manage noise and varying densities, as exemplified by DBSCAN's performance. Moreover, the comparison of evaluation metrics highlights the limitations of relying solely on conventional methods like the Elbow Method, which may not adequately capture the nuances of complex geoscientific data. The results further emphasize that the choice of evaluation metric can be as critical as the choice of clustering algorithm, with DBI and SI each excelling in different scenarios.

The two examples presented here illustrate the foundational capability of the *pyClusterWise* package in handling varying dataset densities. While these case studies focus on 2D geochemical datasets (considering the availability of the data), future versions of *pyClusterWise* will incorporate additional methods and metrics to extend its utility to more complex scenarios, such as multivariate data, 3D spatial data or datasets with anisotropic distributions.

The *pyClusterWise* package is not designed to outperform existing clustering tools in computational speed or complexity but to simplify workflows and emphasize the importance of uncertainty quantification and its impact on the final clustering-based decisions. By integrating multiple clustering algorithms and metrics in a single interface, it provides a practical framework for reducing uncertainty and improving the reliability of decisions.

By integrating uncertainty analysis into clustering workflows, this study demonstrates a path toward more informed and reliable decision-making in geo-data science. The results emphasize that selecting the right combination of clustering and evaluation methods can significantly reduce uncertainties, ensuring meaningful and interpretable results that align with the dataset's inherent variability.

While this study highlights the role of uncertainty quantification in clustering-based decision-making, the theoretical and methodological details have been extensively covered in the author's previous works (see Sadeghi, 2020; Sadeghi and Cohen, 2023; Sadeghi, 2024a,b-including a variety of relevant references). Readers are encouraged to

consult these resources for a comprehensive understanding of uncertainty quantification and its practical application in mineral exploration.

The results of this study underline the importance of method selection in real-world scenarios, particularly in mineral exploration, where decisions often rely on sparse and noisy datasets. By tailoring clustering methods and metrics to the specific dataset, exploration strategies can be optimized, leading to more reliable identification of potential mineralization zones. For example, the demonstrated robustness of DBSCAN in sparse datasets makes it invaluable for early-stage exploration in undersampled regions, while metrics like DBI ensure clear separation of geochemical anomalies, reducing ambiguity in anomaly detection.

## 4. Conclusions

This study highlights the importance of selecting appropriate clustering methods and evaluation metrics tailored to the characteristics of datasets in geo-data science. Using contrasting hydrogeochemical datasets, it demonstrates how clustering performance varies with data density, distribution, and noise.

The *pyClusterWise* package, developed for this research, enables robust analysis by supporting multiple clustering algorithms (e.g., K-Means, DBSCAN) and evaluation metrics (e.g., SI and DBI). DBSCAN proved most reliable for handling irregular clusters and noise, particularly in sparse datasets, while DBI and SI excelled as evaluation metrics for dense and sparse datasets, respectively.

By integrating uncertainty analysis, *pyClusterWise* provides a practical framework for more reliable clustering and decision-making, reducing uncertainties in mineral exploration and similar fields. Future efforts will expand the package with additional algorithms and metrics to address evolving needs in geo-data science, especially because while the statistical metrics described in this study—such as the Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, the Elbow method (relative change in inertia), and BIC—provide a quantitative measure of model uncertainty, geological validity extends beyond these numerical assessments. Geological validity reflects how well the derived clusters correspond to actual geological features—such as known mineralization zones, lithological boundaries, or spatial patterns in geochemical anomalies. To quantify geological validity, one may integrate independent geoscientific data (e.g., geological maps or field observations), employ spatial statistical techniques (like spatial autocorrelation), or incorporate expert geological judgment. In practice, this means comparing the spatial distribution of clusters with established geological structures or verifying that clusters delineate regions with distinct geochemical signatures that have been validated in the field.

Ultimately, the selection of an optimal clustering method in geo-data science should be based on a careful trade-off between model uncertainty and geological validity. For example, a clustering method that exhibits higher statistical uncertainty might still be preferred if its clusters are in excellent agreement with known geological features, whereas a method with lower uncertainty but poor geological correspondence would be less useful in practice. Thus, a multi-criteria evaluation framework is essential—one that considers both the quantitative metrics of clustering uncertainty and the qualitative aspects of geological interpretability. This integrated approach ensures that the final clustering outcome is not only mathematically robust but also geologically meaningful, thereby improving decision-making in applications such as mineral exploration.

## 5. Supplementary

To install the package, run the following command: *pip install pyClusterWise*. To request access to the package repository, please reach out to the author using the provided email addresses.

## 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.

## Acknowledgements

The author sincerely thanks Dr. Ahmed Eleish from Rensselaer Polytechnic Institute (USA) for the invaluable feedback on the draft of this paper and their insightful suggestions that helped refine and improve the ideas presented here. The author also acknowledges the reviewers for their constructive and helpful comments.

## Data availability

The data that has been used is confidential.

## References

Abdel-Basset, M., Mohamed, R., El-Shahat, D., Sallam, K.M., Hezam, I.M., AbdelAziz, N. M., 2024. Energy-efficient trajectory optimization algorithm based on K-medoids clustering and gradient-based optimizer for multi-UAV-assisted mobile edge computing systems. Sustainable Comput. Inf. Syst. 44, 101045. https://doi.org/10.1016/j.suscom.2024.101045.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Andrade, A., Stigter, T., 2013. The distribution of arsenic in shallow alluvial groundwater under agricultural land in central Portugal: insights from multivariate geostatistical modeling. Sci. Total Environ. 449, 37–51.

Bagirov, A.M., Aliguliyev, R.M., Sultanova, N., 2023. Finding compact and well-separated clusters: clustering using silhouette coefficients. Pattern Recogn. 135, 109144. https://doi.org/10.1016/j.patcog.2022.109144.

Bigdeli, A., Maghsoudi, A., Ghezelbash, R., 2022. Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran. J. Geochem. Explor. 233, 106923. https://doi.org/10.1016/j.gexplo.2021.106923.

Birks, H.J.B., Heiri, O., Seppä, H., Bjune, A.E., 2012. Strengths and weaknesses of quantitative climate reconstructions based on late-Quaternary biological proxies. The Open Ecology Journal 3 (1), 68–110.

Caers, J., 2011. Modeling Uncertainty in the Earth Sciences. Wiley-Blackwell.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. Statist. Simulation Comput. 3 (1), 1–27.

Cardoso-Fernandes, J., Lima, J., Lima, A., Roda-Robles, E., Kohler, M., Schaefer, S., Barth, A., Knobloch, A., Goncalves, M.A., Goncalves, F., Teodoro, A.C., 2022. Stream sediment analysis for Lithium (Li) exploration in the Douro region (Portugal): a comparative study of the spatial interpolation and catchment basin approaches. Journal of Geochemical Exploration, Volume 236, May 2022, 106978.

Carranza, E.J.M., 2009. Geochemical anomaly and mineral prospectivity mapping in GIS. Handbook of Exploration and Environmental Geochemistry 11, 197–226.

Chen, Y., Li, H., Lin, H., Wang, Y., Zhao, Y., Chen, Y., 2023. Critical slip line recognition and extraction method of slope based on modified k-medoid clustering algorithm. Comput. Geotech. 154, 105125. https://doi.org/10.1016/j.compgeo.2022.105125.

Cheng, D., Zhang, C., Li, Y., Xia, S., Wang, G., Huang, J., Zhang, S., Xie, J., 2024. GB-DBSCAN: a fast granular-ball based DBSCAN clustering algorithm. Inf. Sci. 674, 120731. https://doi.org/10.1016/j.ins.2024.120731.

Cloutier, V., Lefebvre, R., Therrien, R., Savard, M.M., 2008. Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. J. Hydrol. 353 (3), 294–313.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2), 224–227.

Davis, J.C., Sampson, R.J., 2002. Statistics and Data Analysis in Geology, vol. 3. Wiley, New York.

Di, H., Gao, D., 2016. Seismic facies classification and reservoir prediction using K-means cluster analysis in a case study from the northern Gulf of Mexico. J. Pet. Sci. Eng. 139, 245–256.

Du, P., Li, F., Shao, J., 2024. Multi-agent reinforcement learning clustering algorithm based on silhouette coefficient. Neurocomputing 596, 127901. https://doi.org/10.1016/j.neucom.2024.127901.

Dudek, A., 2020. Silhouette index as clustering evaluation tool. Classification and Data Analysis: Theory and Applications 28, 19-33, Springer International Publishing.

Ebrahimi, P., Guarino, A., Allocca, V., Caliro, S., Avino, R., Bagnato, E., Capecchiacci, F., Carandente, A., Minopoli, C., Santi, A., Albanese, S., 2022. Hierarchical clustering and compositional data analysis for interpreting groundwater hydrogeochemistry: the application to Campi Flegrei volcanic aquifer (south Italy). J. Geochem. Explor. 233, 106922. https://doi.org/10.1016/j.gexplo.2021.106922.

Eminagaoglu, M., Oskay, R.G., Karayigit, A.I., 2022. Evaluation of elemental affinities in coal using agglomerative hierarchical clustering algorithm: a case study in a thick and mineable coal seam (kM2) from Soma Basin (W. Turkey). Int. J. Coal Geol. 259, 104045. https://doi.org/10.1016/j.coal.2022.104045.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), 226-231.
Goncalves, M., Rasteiro da Silva, D., Duuring, P., Gonzalez-Alvarez, I., Ibrahimi, T., 2024. Mineral exploration and regional surface geochemical datasets: an anomaly detection and K-means clustering exercise applied on laterite in Western Australia. J. Geochem. Explor. 258, 107400.
Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochemistry: Exploration Environment, Analysis 10, 27–74.
Güler, C., Thyne, G.D., McCray, J.E., Turner, K.A., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. 10 (4), 455–474.
Hajihosseinlou, M., Maghsoudi, A., Ghezelbash, R., 2024a. A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins. Geochemistry 84 (2), 126094. https://doi.org/10.1016/j.chemer.2024.126094.
Hajihosseinlou, M., Maghsoudi, A., Ghezelbash, R., 2024b. Intelligent mapping of geochemical anomalies: adaptation of DBSCAN and mean-shift clustering approaches. J. Geochem. Explor. 258, 107393. https://doi.org/10.1016/j.gexplo.2024.107393.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
Heidari, J., Daneshpour, N., Zangeneh, A., 2024. A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers. Pattern Recogn. 155, 110639. https://doi.org/10.1016/j.patcog.2024.110639.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31 (8), 651–666.
Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.
Jansson, N.F., Allen, R.L., 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. 233, 106909.
Jiang, Y., Guo, H., Jia, Y., Cao, Y., Hu, C., 2015. Principal component analysis and hierarchical cluster analyses of arsenic groundwater geochemistry in the Hetao basin Inner Mongolia. Geochemistry 75 (2), 197–205.
Kaiser, H.F., 1992. On Cliff's formula, the Kaiser-Guttman rule, and the number of factors. Percept. Mot. Skills 74, 595–598.
Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data: an introduction to cluster analysis. Wiley.
Ketchen, D.J., Shook, C.L., 1996. The application of cluster analysis in strategic management research: an analysis and critique. Strateg. Manag. J. 17 (6), 441–458.
Kim, J., Lee, H., Ko, Y.M., 2024. Constrained Density-Based Spatial Clustering of Applications with Noise (DBSCAN) using hyperparameter optimization. Knowledge-Based Systems, 303, 112436, doi: 10.1016/j.knosys.2024.112436.
Lenssen, L., Schubert, E., 2024. Medoid Silhouette clustering with automatic cluster number selection. Inf. Syst. 120, 102290. https://doi.org/10.1016/j.is.2023.102290.
Li, F., Wang, C., 2023. Develop a multi-linear-trend fuzzy information granule based short-term time series forecasting model with k-medoids clustering. Inf. Sci. 629, 358–375.
Lima, A., De Vivo, B., Cicchella, D., 2013. Multivariate analysis for geochemical data interpretation. Springer Handbook of Geochemistry 809–832.
Lindsay, M., Jessell, M.W., Aillères, L., et al., 2019. Stochastic geological modeling for uncertainty analysis. Geosci. Front. 10 (1), 229–241.
Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In L.M. Le Cam and Neyman (Eds.) proceeding of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281-297). California: university of California Press.
Mardani, K., Maghooli, K., Farokhi, F., 2025. Segmentation of coronary arteries from X-ray angiographic images using density based spatial clustering of applications with noise (DBSCAN). Biomed. Signal Process. Control 101, 107175. https://doi.org/10.1016/j.bspc.2024.107175.
Martins, A.S.C., de Araujo, L.R., Penido, D.R.R., 2024. K-Medoids clustering applications for high-dimensionality multiphase probabilistic power flow. Int. J. Electr. Power Energy Syst. 157, 109861. https://doi.org/10.1016/j.ijepes.2024.109861.
Mazor, E., 2004. Chemical and isotopic groundwater hydrology: The applied approach. CRC Press.
Mazor, E., 2004. Chemical and isotopic groundwater hydrology: The applied approach. CRC Press.
Mazor, E., 2004. Chemical and isotopic groundwater hydrology: The applied approach. CRC Press.
McKinley, J., Hron, K., Filzmoser, P., Reimann, C., 2016. Robust statistical analysis of geochemical data. Chem. Geol. 447, 150–162.
Mukherjee, A., Scanlon, B.R., Fryar, A.E., Saha, D., Ghosh, A., Chowdhuri, S., Mishra, R., 2012. Solute chemistry and arsenic fate in aquifers between the Himalayan foothills and Indian craton (including central Gangetic plain): influence of geology and geomorphology. Geochim. Cosmochim. Acta 90, 283–302.
Ngombe, O.G.D., Walter, J., Chesnaux, R., Molson, J., 2024. Application of hierarchical cluster analysis and principal component analysis to identify compositional trends of brine in crystalline basements and sedimentary basins. Appl. Geochem. 169, 106030. https://doi.org/10.1016/j.apgeochem.2024.106030.
Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical data analysis explained: Applied environmental statistics with R. Wiley.
Richards, J.A., Jia, X., 2006. Remote sensing digital image analysis: an introduction. Springer.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (1), 53–65.
Rylko, N., Stawiarz, M., Kurtyka, P., Mityushev, V., 2024. Study of anisotropy in polydispersed 2D micro and nano-composites by Elbow and K-Means clustering methods. Acta Mater. 276, 120116. https://doi.org/10.1016/j.actamat.2024.120116.
Sadeghi, B., 2020. Quantification of Uncertainty in Geochemical Anomalies in Mineral Exploration. University of New South Wales.
Sadeghi, B., 2024. Fractals and Multifractals in the Geosciences, ELSEVIER, 302 pages (doi: 10.1016/C2020-0-03441-9).
Sadeghi, B., 2024. Fractals and Multifractals in the Geosciences. Elsevier, 302 pages.
Sadeghi, B., Cohen, D.R., 2023. Decision-making within geochemical exploration data based on spatial uncertainty–a new insight and a futuristic review. Ore Geol. Rev. 161, 105660.
Sadeghi, B., Grunsky E., Pawlowsky-Glahn V., 2023 Uncertainty Quantification. In: Daya Sagar B., Cheng Q., McKinley J., Agterberg F. (eds) Encyclopedia of Mathematical Geosciences. Encyclopedia of Earth Sciences Series. Springer, Cham., 1583–1589. doi: 10.1007/978-3-030-26050-7_334-1.
Sadeghi, M., Casey, P., Carranza, E.J.M., Lynch, E.P., 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 Geology Reviews 167, 106002. https://doi.org/10.1016/j.oregeorev.2024.106002.
Scheidt, C., Li, L., Caers, J., 2018. Quantifying Uncertainty in Subsurface Systems. American Geophysical Union.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Soltani, Z., Imamalipour, A., 2022. An improved classification of mineralized zones using particle swarm optimization: a case study from Dagh-Dali ZnPb (±Au) prospect Northwest Iran. Geochemistry 82 (1), 125850.
Sureja, N., Chawda, B., Vasant, A., 2022. An improved K-medoids clustering approach based on the crow search algorithm. Journal of Computational Mathematics and Data Science 3, 100034. https://doi.org/10.1016/j.jcmds.2022.100034.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 23 (8), 2198–2213.
Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18 (4), 267–276. https://doi.org/10.1007/BF02289263.
Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 63 (2), 411–423. https://doi.org/10.1111/1467-9868.00293.
Unal, Y., Kindap, T., Karaca, M., 2003. Redefining the climate zones of Turkey using cluster analysis. Int. J. Climatol. 23 (9), 1045–1055.
Vergani, A.A., Binaghi, E., 2018. A Soft Davies-Bouldin Separation Measure. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Rio de Janeiro, Brazil, 2018, pp. 1-8, doi: 10.1109/FUZZ-IEEE.2018.8491581.
Wang, L., Yin, D.Z., Caers, J., 2023. Data Science for the Geosciences. Cambridge University Press.
Xu, R., Tian, Y., 2015. A comprehensive survey of clustering algorithms. Ann. Data Sci. 2 (2), 165–193.
Yidana, S.M., 2010. Groundwater classification using multivariate statistical methods: Southern Ghana. J. Afr. Earth Sci. 57 (5), 455–469.
Yidana, S.M., Banoeng-Yakubo, B., Akabzaa, T.M., 2010. Analysis of groundwater quality using multivariate and spatial analyses in the Keta basin Ghana. J. Afr. Earth Sci. 58 (2), 220–234.
Zhang, Y., Shang, P., 2022. KM-MIC: An improved maximum information coefficient based on K-Medoids clustering. Commun. Nonlinear Sci. Numer. Simul. 111 (2022), 106418. https://doi.org/10.1016/j.cnsns.2022.106418.
Zhao, Y., Gao, Y., Wei, Q., Tang, X., Zhao, Y., 2024. Identification of geochemical anomalies in multibackground areas using the combined k-means clustering and residual contrast value method: a case study in a district in Hunan. China. Journal of Geochemical Exploration 261, 107451. https://doi.org/10.1016/j.gexplo.2024.107451.
Zhong, C., Wang, H., Yang, Q., 2022. Hydrochemical interpretation of groundwater in Yinchuan basin using self-organizing maps and hierarchical clustering. Chemosphere 309 (2), 136787. https://doi.org/10.1016/j.chemosphere.2022.136787.