Journal of Geochemical Exploration 231 (2021) 106872

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Knowledge discovery of geochemical patterns from a data-driven perspective

Bojun Yin [a], Renguang Zuo [a,*], Yihui Xiong [a,*], Yongsheng Li [b,c], Weigang Yang [d]

[a] State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan 430074, China
[b] Developement and Research Center, China Geological Survey, Beijing 100037, China
[c] Mineral Exploration Technical Guidance Center, Ministry of Natural Resources, Beijing 100083, China
[d] Gansu Institute of Geological Survey, Lanzhou 730000, China

ARTICLE INFO

Keywords:
Data-driven
Knowledge discovery
Data science
Geochemical exploration

ABSTRACT

We have entered the fourth research paradigm with the overwhelming availability of vast amounts of data. The processing and mining these data for a better understanding of earth systems and predicting mineral resources is challenging. This study discusses a data-driven knowledge discovery of geochemical patterns and presents a case study of geochemical data processing from a data-driven perspective. We employed local indicators of spatial association (LISA), principal component analysis (PCA), and deep autoencoder network (DAN) procedures to explore spatial association of geochemical patterns, extract elemental associations, and detect geochemical anomalies related to Au—Sb mineralization in the Daqiao district, Gansu Province, China. The results indicate the following: (1) both Au and Sb, and Pb and Zn have a close spatial correlation, indicating genetic connections among them; (2) the elemental association of Au, Sb, As, Hg and Ag can be adopted as a geochemical signature for the discovery of Au—Sb polymetallic mineralization in the study area; and (3) the geochemical anomalies identified by DAN exhibit a strong spatial relationship with locations of known mineral deposits and can provide a significant clue for further mineral exploration in this district. These findings indicate that data-driven procedures can help in the knowledge discovery of geochemical patterns in mineral exploration. Additional efforts are required for data-driven knowledge discovery in both geochemical prospecting and mineral exploration.

1. Introduction

The introduction of the data-intensive scientific discovery research paradigm (the fourth paradigm) (Tansley and Tolle, 2009; Chen and Zhang, 2014) marks the arrival of the era of big data. The new research paradigm utilizes a large amount of data from multiple sources and fields, and then analyzes them using techniques designed to discover complex patterns in high-dimensional data (Tansley and Tolle, 2009). Thus, new models, knowledge, and inferences derived from the data can be discovered and explored. Earth science studies have produced a vast number of datasets collected from satellite observations, ground sensor networks, and other sources (Reichstein et al., 2019). These are collectively called big earth data, which not only have common characteristics with scientific big data, such as complexity, comprehensiveness, and global coverage, but are also multi-source, heterogeneous, multi-temporal, multi-scale, high-dimensional, highly complex, and nonstationary. The availability of big earth data has promoted data-intensive research in geosciences (Guo et al., 2014; Zuo and Xiong, 2018). The large quantities of data make it possible to pursue a data-driven approach to gain knowledge from data rather than a knowledge-driven approach that extracts hypothesized patterns expected from the data (Bergen et al., 2019; Cheng et al., 2020; Zuo, 2020).

In geochemical exploration, high-quality, multi-element and multi-scale geochemical data have been accumulated in the past decades (e.g., Darnley et al., 1995; Xie et al., 1997; de Caritat et al., 2010; Reimann et al., 2012). These data not only make an important contribution to mineral prospecting (Zuo et al., 2016; Grunsky and de Caritat, 2020) and environmental assessment (Cohen et al., 2010; Galuszka and Migaszewski, 2011), but also lay an important foundation for global geochemical mapping and the establishment of a global geochemical reference network (Xie and Cheng, 2001; Wang and The CGB Sampling Team, 2015). In this context, it is particularly important to develop geochemical anomaly recognition and extraction technologies under complex geological settings for comprehensively evaluating

* Corresponding authors.
E-mail addresses: zrguang@cug.edu.cn (R. Zuo), xiongyh426@cug.edu.cn (Y. Xiong).

https://doi.org/10.1016/j.gexplo.2021.106872
Received 9 July 2021; Received in revised form 26 July 2021; Accepted 1 August 2021
Available online 8 August 2021

geochemical data for mineral exploration and environmental studies (Zuo and Xiong, 2018).

Data-driven approaches for geochemical pattern knowledge discovery have transformed from simple to complex. In addition to frequency-based methods, such as mean ± 2 × standard deviation (Hawkes and Webb, 1963), probability graphs (Sinclair, 1974), and gap statistics (Miesch, 1981), and spatial frequency-based methods, such as geostatistics (Matheron, 1962), and fractal/multifractal models (Cheng et al., 1994, 2000; Cheng, 2007), machine learning has attracted considerable attention as a potential tool for studying geochemical patterns (Zuo, 2017). Machine learning algorithms are suitable for quantifying complex and nonlinear data, and help overcome the conventional multivariate statistical methods that require the data to satisfy a hypothesis of a known multivariate probability distribution (Chen et al., 2014; Chen et al., 2019; Xiong and Zuo, 2016, 2020, 2021; Zuo et al., 2019; Parsa et al., 2018; Luo et al., 2020, 2021; Parsa, 2021; Parsa and Carranza, 2021; Zhang et al., 2021; Zhang and Zuo, 2021). Some supervised machine learning methods, such as neural networks (Ziaii et al., 2009; Yu et al., 2019), metric learning (Wang et al., 2019a, 2019b), support vector machines (Gonbadi et al., 2015), and random forests (Gonbadi et al., 2015; Sadr and Nazeri, 2018), as well as unsupervised machine learning methods, including continuous restricted Boltzmann machines (Chen et al., 2014), one-class support vector machines (Chen and Wu, 2017; Xiong and Zuo, 2020), and isolation forests (Wu and Chen, 2018), have been used for multivariate geochemical pattern quantification and anomaly recognition.

Peter Naur first defined the term data science (DS) in 1960 as "data processing" in view of computer science (Gibert et al., 2018), aiming at discovering new knowledge and extracting information from a number of datasets, thus guiding decision-making. Almost simultaneously, Tukey (1962) envisaged the existence of an as-yet unrecognized science based on statistics, that supported learning from data analysis (Tukey, 1962). With the development of computer science and information technology, especially the advent of artificial intelligence and machine learning algorithms, DS has come to the fore during the past decades. Based on developments in data-driven methods for geochemical pattern knowledge discovery, Zuo and Xiong (2020) proposed a systematic Geodata science (GDS) procedure for geochemical pattern recognition. GDS is a cross-discipline of geoscience and DS. Its purpose is to process and mine geoscience datasets to draw statistical inferences and predict geological processes or events (Zuo and Xiong, 2020). GDS comprises three procedures: data statistics, data mining, and data insight and prediction. These can be adopted to derive geoinformation and geoknowledge from the correlations revealed in big earth data (Zuo and Xiong, 2020; Zuo, 2020).

In this study, we applied data-driven scientific discovery methods, including statistical data analysis, data mining, and data insight and prediction to reveal geochemical patterns and support mineral exploration in the Daqiao district, Gansu province, China.

2. Study area and data

2.1. Geological background

The Daqiao district is located in the east domain of the West Qinling belt and the transition area of the northern south Qinling belt and middle Qinling belt, the West Qinling Orogen is one of the largest prospective gold regions in China (Chen and Santosh, 2014; Goldfarb et al., 2014; Liu et al., 2015). Previous studies have shown that Triassic and Devonian metasedimentary rocks host most of the gold mineralization in the West Qinling Orogen (Mao et al., 2002; Wu, 2019). The underlying Early Paleozoic metamorphosed sediments, such as the Silurian graphite schist, could have acted as the source of gold mineralization that efficiently released gold-bearing components during the greenschist-amphibolite facies metamorphism (Wu, 2019). Therefore, in places where regional deep faults were developed, we should focus on the favorable structural locations in the Triassic and Devonian metasediments. However, in other regions, it was more favorable to explore orogenic gold mineralization in underlying source rocks due to the lack of regional fluid pathways to carry gold away to depositional sites. The relationship between the strata and deposits was confirmed by the geological map of the Shixia area (Fig. 1). In addition, recent geochemical exploration in this area indicated that Au-Hg-As-Sb-Ag ca serve as a mineralization-related geochemical signature based on contour maps of regional geochemical data (Zhang et al., 2015).

The Daqiao gold deposit (>105 t Au, 3–4 g/t) is a typical gold mineral deposit in this region (Zhang, 2016; Zhang et al., 2018). Two primary formations outcrop in this area: the turbidite sequences of the middle Triassic Huashiguan formation hosting orebodies and the underlying Carboniferous thick-bedded limestones. Silurian micaceous graphitic schists have also been identified through drilling (Wu et al., 2018, 2020). Zircon U—Pb dating of the granodiorite via LA-ICP-MS indicated that it was emplaced at 228–206 Ma (Shan et al., 2016) or 215–212 Ma, while the diorite porphyry was emplaced at 188 Ma (Wu et al., 2019). The $^{40}Ar/^{39}Ar$ plateau ages of the two types of sericites that had intimate textural relationships with auriferous pyrite were at 153.8–139.8 Ma and 133.9–126.6 Ma, respectively. This indicates that gold mineralization in this area occurred in two episodes starting from the Late Jurassic and terminating in the Early Cretaceous (Wu et al., 2018; Wu et al., 2019). These findings show that mineralization occurred after intrusion of the granodioritic and dioritic porphyry dykes. The Daqiao gold deposit was structurally controlled by the Zhouqu-Chengxian-Huixian Fault, northwestern flank of an inferred anticline, NE trending reverse faults and the Yaoshan-Shixia fault (You and Zhang, 2009; Wu et al., 2018; Wu, 2019). Numerous orebodies were cut by the NE-trending faults, indicating strong fault activity after gold mineralization (Wu et al., 2018).

In summary, Triassic and Devonian metasedimentary rocks combined with regional-scale faults and Paleozoic metamorphosed sediments, such as the Silurian graphite schist without regional scale faults, NE-trending small faults, and Au-Hg-As-Sb-Ag geochemical anomalies are important geological and geochemical prospecting factors (Zhang et al., 2015; Wu, 2019).

2.2. Data

A total of 2090 stream sediment geochemical samples, collected as a part of a 1:50,000 scale regional geological and mineral survey, were analyzed for 15 elements by the Gansu Geological Survey Institute of China (Fig. 2). Seven elements (Cu, Pb, Zn, Cd, W, Mo, and Au) were measured using inductively coupled plasma-mass spectrometry; two elements (Ag and Sn) were measured using emission spectrometry; two elements (Ba and Mn) were measured using inductively coupled plasma-atomic emission spectrometry, the remaining four elements (As, Sb, Bi, and Hg) were measured using atomic fluorescence spectroscopy. Details regarding about the detection limits and data quality control are available in Xie et al. (1997).

3. Data-driven approaches for geochemical patterns recognition

Generally, the discovery of geochemical patterns involves three procedures: revealing the spatial geochemical patterns, exploring geochemical element associations, and detecting geochemical anomalies linked to mineralization to support mineral exploration (Zuo and Xiong, 2020). In this study, three data-driven approaches, namely the local indicators of spatial association (LIAS), principal component analysis (PCA), and deep autoencoder network (DAN) were used as representative tools for geochemical spatial pattern recognition, geochemical elemental association extraction, and geochemical anomaly identification, respectively.

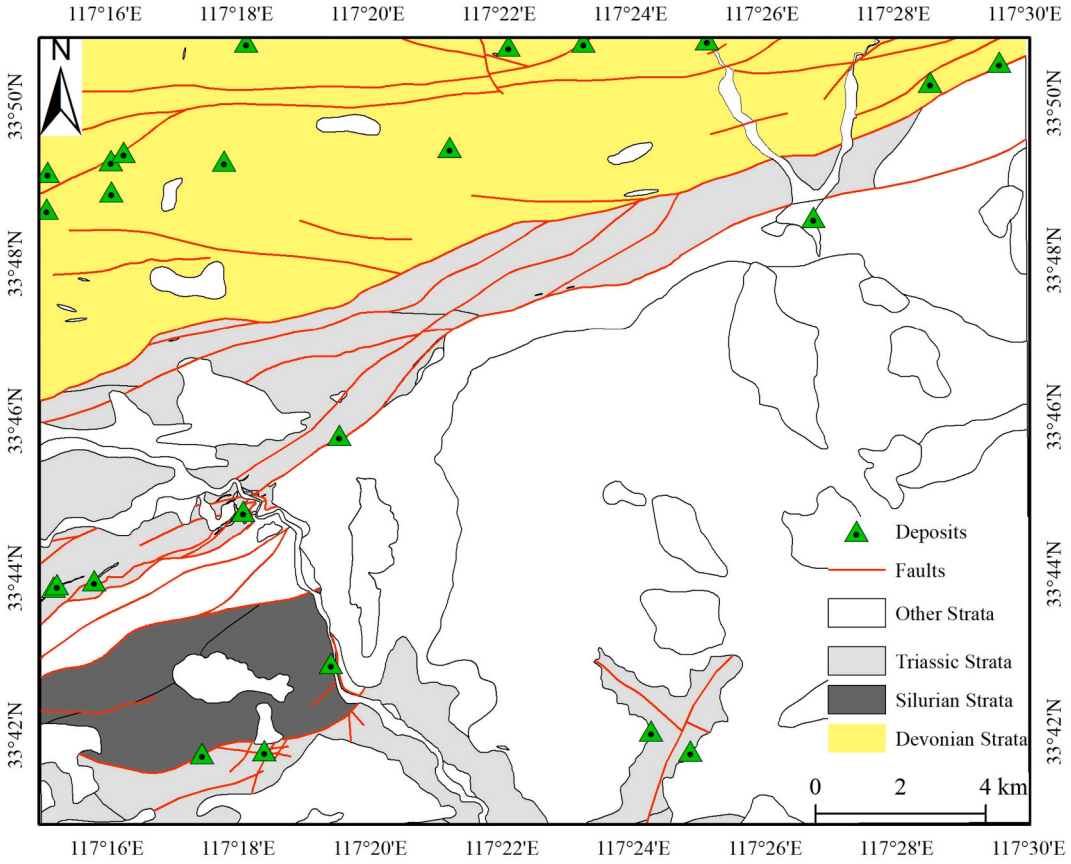

Fig. 1. A simplified geological map of Daqiao ore concentration area (Shixia area), Gansu Province, China.

3.1. Local indicators of spatial association

Moran (1950) formally introduced Moran's *I* statistics to assess the spatial distribution pattern of objects when studying the relationships between continuous and discrete processes that are distributed in two or more dimensional spaces. Tobler (1970) was the pioneer to reveal the interaction between objects with a spatial association: "*everything is related to everything else, but near things are more related than distant things*". With the expansion of spatial data, more tools are needed to evaluate it. Exploratory spatial data analysis (ESDA), as an extension of exploratory data analysis (Tukey, 1977), can quantify the spatial association in a dataset using spatial autocorrelation coefficients such as Moran's *I* and Geary's *C* (Anselin and Getis, 1993). These approaches focus on global statistics and cannot assess the local instabilities for each observation within the data.

To further display the local patterns of spatial association, Anselin (1995) proposed the local Moran's *I* statistic as a corresponding local indicator of spatial association (LISA) to measure the spatial association for every observation with its neighbors. The Local Moran's *I* statistics of spatial association are given as (Anselin, 1995):

$$I_i = \frac{x_i - \overline{X}}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{i,j}\left(x_j - \overline{X}\right) \quad (1)$$

where x_i represents an attribute of a feature i, $\overline{X}$ is the mean of the corresponding attribute, and $w_{i,j}$ is the spatial weight between feature i and j. S_i^2 can be expressed as (Anselin, 1995):

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}\left(x_j - \overline{X}\right)^2}{n-1} \quad (2)$$

Local Moran's *I* is an effective way to identify local clusters (hot/cold spots) and local spatial outliers of a spatial dataset. A positive value of *I* indicates that a feature has neighboring features with similarly high or low attribute values; this feature is part of a cluster. A negative value of *I* indicates that a feature has neighboring features with dissimilar values; this feature is an outlier. The cluster/outlier type field distinguishes between a statistically significant cluster of high values (HH), cluster of low values (LL), outlier in which a high value is surrounded primarily by low values (HL), and outlier in which a low value is surrounded primarily by high values (LH).

3.2. Principal component analysis

PCA is an efficient tool for dimensionality reduction of large datasets to increase interpretability with minimal information loss (Jolliffe and Cadima, 2016; Zuo, 2011). Principal components (PCs) are presented as a linear combination of raw variables, and the coefficient (loading) of each variable in the linear combination represents the correlation between the variable and PC. The eigenvalue of each PC represents the share of the PC variance in the total variance of the original data. The larger the eigenvalue, the more important the corresponding PC. PCA has been used to identify pathfinder elements or elemental associations in geochemical exploration (Zuo, 2011; Xiong et al., 2018; Grunsky and Arne, 2020; Kadel-Harder et al., 2020; Zuo and Xiong, 2020; Parsa and Pour, 2021).

Geochemical exploration data are subject to the closure problem because all element contents in each sample sum up to a constant (e.g., 1 or 100%), which is the key characteristic of compositional data. They are parts of a whole that only contain relationship information (Aitchison, 1982, 1986; Buccianti and Pawlowsky-Glahn, 2005). The effects of the data closure problem on geochemical exploration have been demonstrated in various studies on geochemical exploration (e.g., Zuo et al., 2013; Zuo, 2014). The closure problem may result in spurious

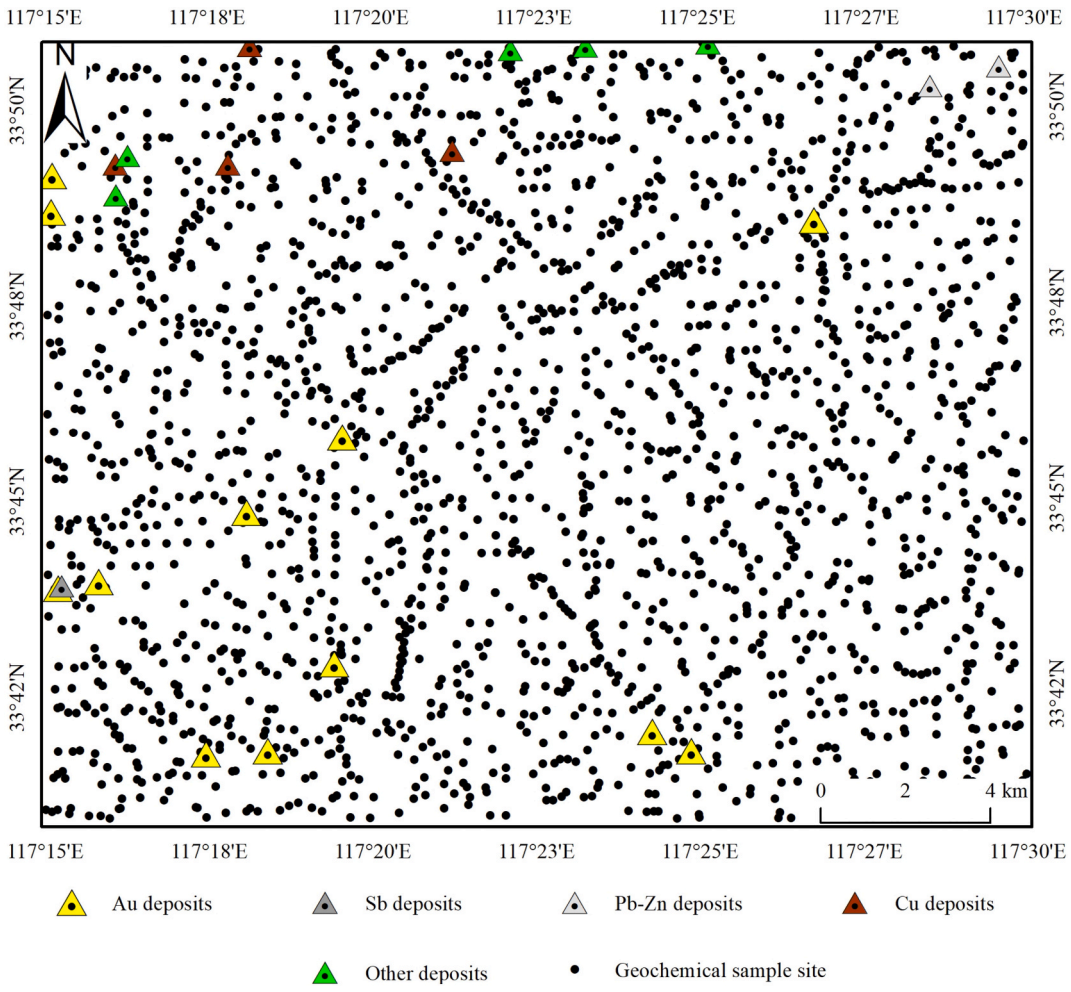

Fig. 2. A map showing stream sediment sampling sites.

correlations between geochemical variables and complicate the interpretation of different correlations between the same variables among different sub-compositions (Filzmoser et al., 2009). Standard multivariate statistical methods, such as PCA, are developed in Euclidean space (i.e., open system) and may not be appropriate for the analysis of compositional data that lie in a simplex space (i.e., closed system) (Aitchison, 1986). Aitchison (1982, 1986) proposed the additive logratio (alr) and centered logratio (clr) transformation to deal with the closure issue in compositional data. Egozcue et al. (2003) developed an isometric logratio (ilr) transformation that correctly represented compositional data in Euclidean space. However, the ilr-transformed variables lacked correspondence with the original variables and the ilr-transformed variables. Therefore, the analysis results generated from ilr-transformed variables are difficult to interpret (Reimann et al., 2008; Filzmoser et al., 2009). For interpreting results of ilr transformed data, we usually have to back-transform the results (e.g., loadings and scores of PCA) into the clr space (Filzmoser et al., 2009; Parsa et al., 2017).

3.3. Deep autoencoder network

DAN was developed as a variant of the deep belief network based on the restricted Boltzmann machine (RBM), which can model an ensemble of binary vectors (e.g., images) and Gaussian inputs (Hinton and Salakhutdinov, 2006). However, the RBM cannot handle non-Gaussian inputs with continuous values (Chen and Murray, 2003; Chen et al., 2014). Therefore, Chen and Murray (2003) introduced continuous RBM (CRBM), which can be used for modelling complex multivariate continuous data, making it suitable for handling geochemical exploration data characterized by a strongly (right) skewed, multi-modal data distribution. Detailed information, such as the optimal parameter selection process and weight initialization of the DAN, is available in Hinton and Salakhutdinov (2006) and Xiong and Zuo (2016).

Training DAN involves three phases: pre-training, unrolling and fine-tuning. In the first phase, each CRBM is pre-trained to initialize weights by minimizing the contrastive divergence learning rule and greedy layer-wise learning (Hinton, 2002; Hinton and Salakhutdinov, 2006). Once pre-trained, CRBMs are unrolled to build a DAN where one CRBM is "stacked" on top of the former pre-trained CRBM, using the output of the former as its input. Finally, the parameters of the entire DAN are fine-tuned and adjusted based on a back-propagation algorithm. The primary aim of the DAN is to guarantee that the reconstructed output is infinitely close to the input by minimizing reconstruction errors. The DAN-based geochemical anomaly recognition relies on the higher construction error of geochemical anomaly samples than geochemical background samples because small sample sizes are generally linked to a low probability of detection (Xiong and Zuo, 2016).

4. Results and discussion

4.1. Spatial geochemical patterns exploration

The spatial autocorrelation of the major ore-forming elements Au, Sb, Pb, and Zn was first analyzed using Anselin Local Moran's *I* (Anselin, 1995) with ArcGIS@™ 10.7. There are a total of 2090 points (Fig. 2). For

Au, 1676 points were neither outliers nor clusters. Fifty-two HH clusters and three HL outliers were identified (Fig. 3a). Six out of twelve Au deposits were located around HH clusters, while only three of twelve other kinds of mineralization were associated with HH clusters of Au. For Sb, 1502 points were classified as neither outliers nor clusters. A total of 110 HH clusters, 4 HL outliers, and 57 LH outliers were identified (Fig. 3b). Nine of twelve Au deposits were spatially related to Sb HH outliers, while only four out of twelve other types of mineralization were related to Sb HH clusters. Most places with Au HH clusters also contained Sb HH clusters; both Sb and Au HH clusters were visibly associated with faults, while the latter also displayed a close relationship with Triassic strata in the area, suggesting a spatial correlation and the same genetic connection between Au and Sb.

For Pb, 1607 points belonged to neither outliers nor clusters. Ninety HH clusters and 58 LH outliers were identified (Fig. 3c). Most of the known Pb mineralization is spatially related to Pb HH clusters. For Zn, 1615 points belonged to neither outliers nor clusters. Seventy-three HH clusters, two HL outliers, and 63 LH outliers were identified (Fig. 3d). Most of the known Zn mineralization occurrences are spatially related to Zn HH clusters. The cluster and outlier patterns of Pb were similar to those of Zn because more than 80% of Pb HH clusters also belonged to Zn HH clusters, implying that Pb and Zn have a close spatial autocorrelation.

LH outliers are typically located close to HH clusters of the four elements (Fig. 3), indicating that the area with elemental concentration often accompanies elemental depletion. Therefore, we can use negative anomalies (i.e., LH outliers) to guide mineral exploration.

4.2. Geochemical element associations selection

Before conducting PCA, the frequency distribution of the raw data, log-transformed data, and ilr-transformed data of the ore-forming elements (Cu, Pb, Zn, Au, and Sb) were inspected using boxplots. The boxplots of the raw (Fig. 4a) and logarithmic data (Fig. 4b) for Cu, Pb, Zn, Au, and Sb show that the median is closer to the bottom of the box, and the whisker is shorter at the lower end of the boxes, suggesting that the geochemical data distributions are right-skewed. The medians of ilr-transformed Cu, Pb, Zn, Au, and Sb (Fig. 4c) are exhibited in the middle of the boxes, and the whiskers are approximately the same on both sides, suggesting that the distributions are symmetric. The histograms of Au (Fig. 5) show that the raw Au has a right-skewed distribution, and both log-transformed and ilr-transformed Au show symmetrical distributions.

The elemental associations related to the Au polymetallic mineralization were then visualized using PCA. The results for the ilr-transformed data (where the ilr transformation method was implemented to "open" the data) for the 15 elements studied are displayed in Fig. 6. The plot of PC1 versus PC2 based on ilr-transformed data showed three different assemblage compositions (Fig. 6). The first comprises Au, As, and Sb positive PC2 loadings, which may represent Au mineralization associated with NE-trending faults (Zhang et al., 2015). The second assemblage, which consists of Zn, Pb, Cd, Ag, and Hg with negative PC1 loadings, possibly represents base metal polymetallic mineralization in

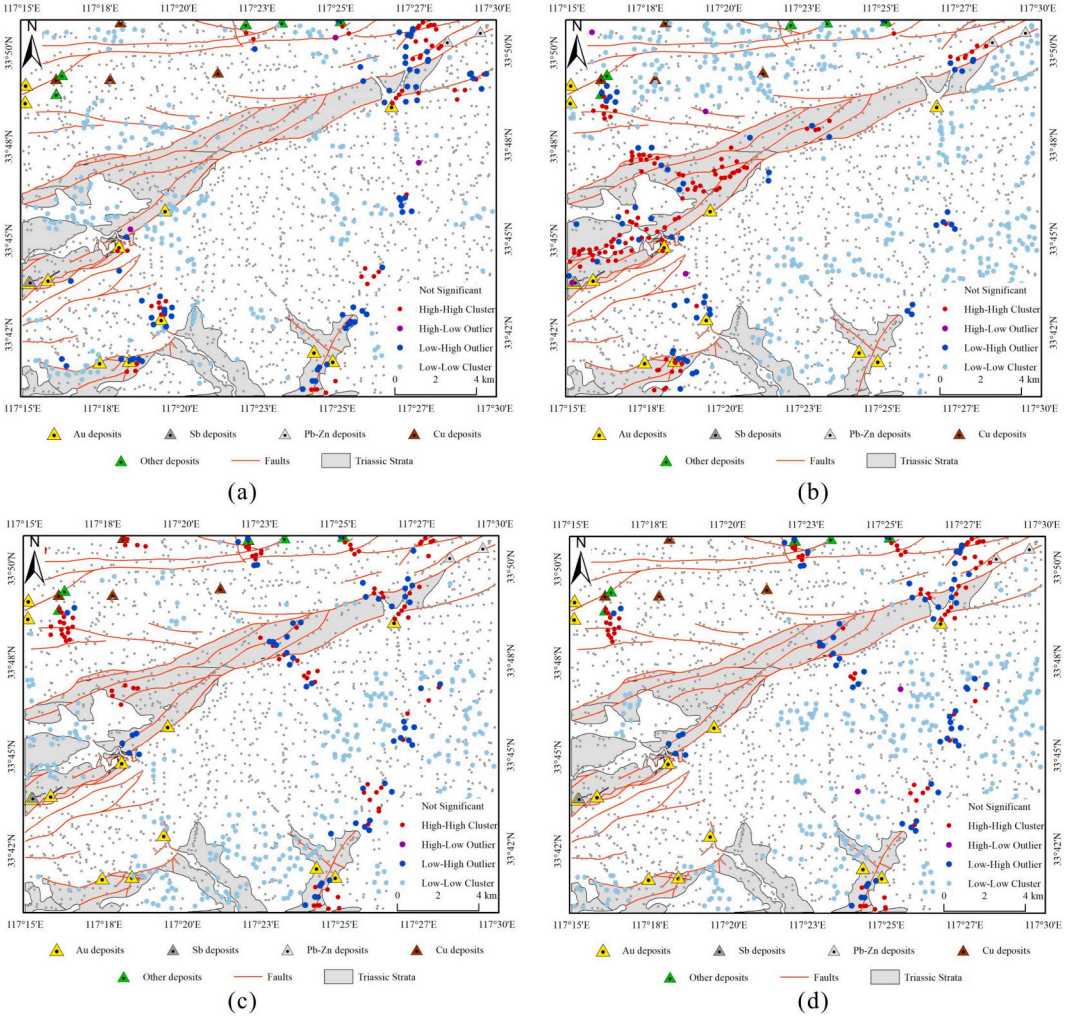

Fig. 3. Cluster and outlier analysis of (a) Au, (b) Sb, (c) Pb, and (d) Zn.

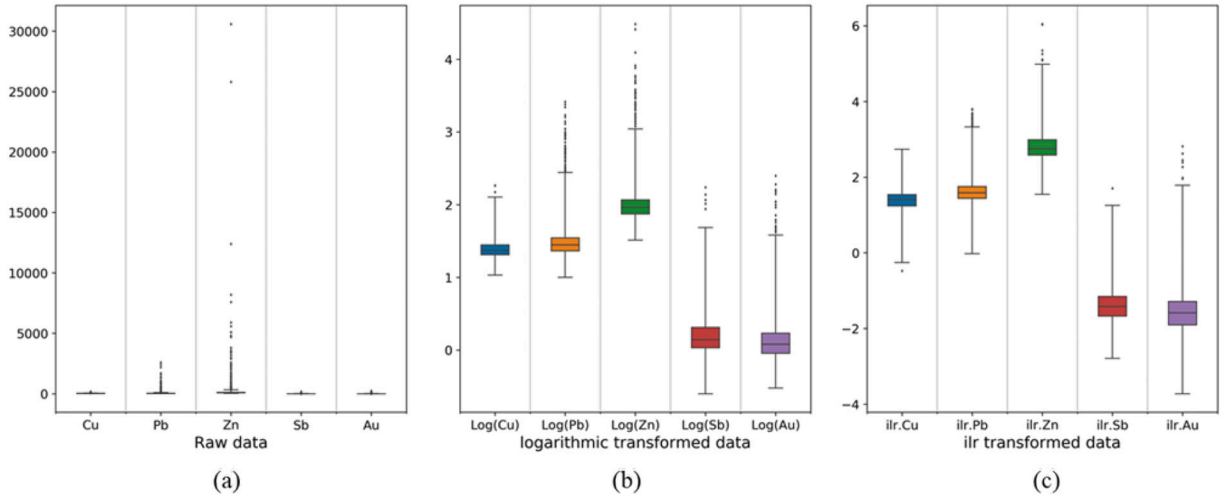

Fig. 4. Boxplots of (a) raw, (b) log-transformed, and (c) ilr-transformed Cu, Pb, Zn, Sb, and Au.

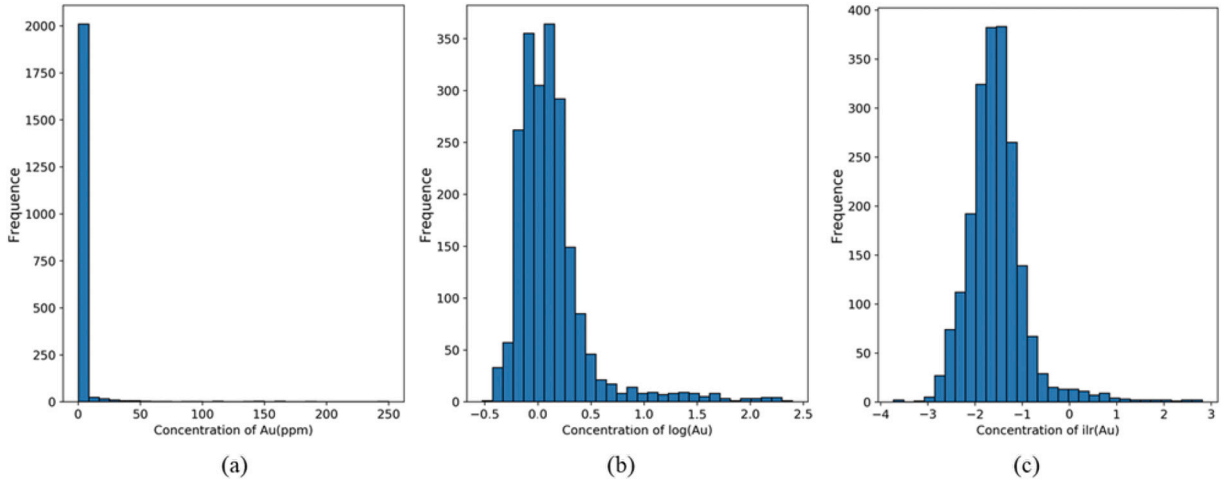

Fig. 5. Histograms for (a) raw, (b) log-transformed, and (c) ilr-transformed Au data.

this district. The remaining Ba, Mn, Cu, Bi, W, Sn, and Mo elements likely reflect background lithological variation, with the latter three elements being associated with granodioritic intrusions. The areas with high PC2 scores were spatially associated with the locations of known Au polymetallic mineralization and showed a close relationship with NE-trending faults (Fig. 7). Therefore, the assemblage of Au-Sb-As-Hg-Ag can be regarded as the geochemical signature for mineral exploration in this area and should be further studied.

4.3. Geochemical anomalies recognition

Five geochemical elements (Au, Sb, As, Hg, and Ag), derived from the positive PC2 loadings (Table 1), were selected to further recognize geochemical anomalies related to Au mineralization using DAN. These five elements in point format were interpolated to a raster grid of 500 m × 500 m. Reconstruction error is the anomaly index of the DAN, which is based on the principle that large samples have a lower reconstruction error than small ones when differentiating geochemical anomalies from the background. Based on several experiments, the number of units in each hidden layer were determined to be 5, 10, 20, and 40. The iteration times and learning rate of the network were determined to be 100 and 0.3, respectively. The network optimization and parameter setting process are detailed in Xiong and Zuo (2016).

After parameter selection, the reconstruction errors of each cell, regarded as the geochemical anomaly scores, were estimated using the DAN. The high anomaly areas detected by DAN identified nearly all known Au deposits and 9 out of 12 other kinds of mineralization (Fig. 8). The success-rate curve was drawn based on a comparison of the grid cells with known mineral deposits with the extracted geochemical anomalies to assess the performance of DAN and PCA. The success-rate measures how the geochemical anomaly areas fit known mineral deposits in the study area. Specifically, a geochemical pattern map is divided into various classes, ranging from those with the highest probability values to those with the lowest probability values. Then, the number of grid cells containing known mineral deposits in each class was calculated, and a cumulative curve was plotted. The curves, depicted in Fig. 9, show that the 23.2% and 29.8% of anomalous areas identified by DAN delineated 50% and 62.5% of the known mineral deposits, whereas 46.3% and 49.5% anomalous areas identified by PCA delineated 50% and 62.5% of the known mineral deposits, respectively. These findings demonstrate that DAN can accommodate complex nonlinear problems, enhancing the identification of geochemical anomalies related to mineralization and the recognition of hidden geochemical patterns. More importantly, the DAN reveals hidden geochemical anomaly patterns in areas with different types of mineral deposits, which is superior to the anomaly map based on PC2 scores.

From a geological perspective, the geochemical anomalies related to Au mineralization identified by DAN are located in or near Triassic

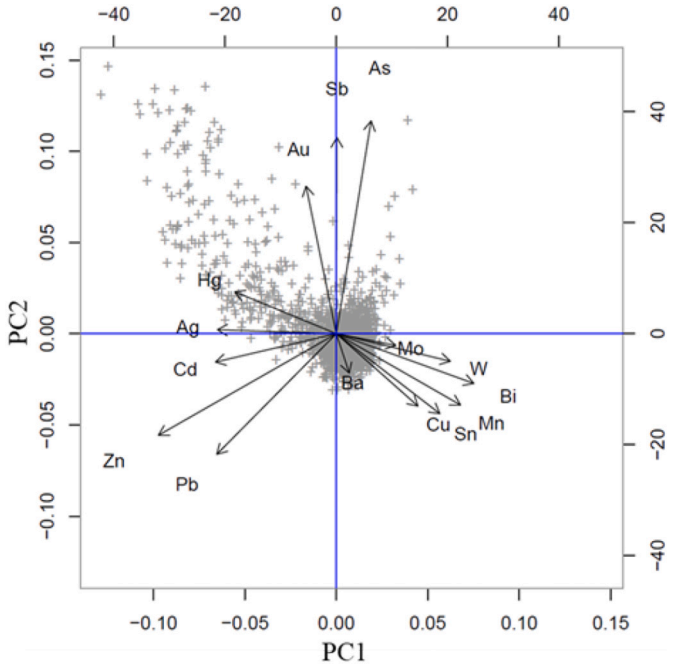

Fig. 6. Biplot of PC1 and PC2 for ilr-transformed Au data.

formation (Fig. 8), which hosts most of the gold mineralization in the West Qinling Orogen. This suggests that the resulting geochemical anomalies are credible and can provide a significant indicator for further mineral exploration in this district.

5. Conclusions

In this study, we focused on data-driven knowledge discovery of geochemical patterns in support of mineral exploration and applied geoscience data processing to mine geochemical exploration data. The following conclusions were drawn:

(1) Au and Sb, as well as Pb and Zn are separately spatially correlated, indicating genetic connections; this aspect should be further explored via mineral deposit studies.

(2) The ilr transformation could overcome the closure problem in geochemical exploration data, and the elemental associations of Au, Sb, As, Hg, and Ag led to their identification as pathfinder elements for the discovery of Au—Sb polymetallic mineralization while the assemblage of Zn, Pb, Cd, Ag and Hg with negative PC1 loadings possibly represents base metal polymetallic mineralization in this district, and the remaining Ba, Mn, Cu, Bi, W, Sn, and Mo elements likely reflect lithological variation, with the latter three elements being associated with granodioritic intrusions.

(3) The geochemical anomalies obtained by the DAN were strongly spatially associated with the locations of known mineralization and, therefore, can provide critical information for further mineral exploration in the study area.

Table 1
Loadings of principal component analysis.

Elements	PC1	PC2	PC3
Cu	0.207	−0.184	−0.130
Pb	−0.303	−0.307	−0.181
Zn	−0.451	−0.258	−0.039
Ag	−0.301	0.010	0.185
Mo	0.151	−0.030	−0.134
Sn	0.262	−0.203	−0.094
W	0.289	−0.070	−0.163
Mn	0.315	−0.181	0.174
Ba	0.033	−0.100	0.738
As	0.088	0.541	0.008
Sb	0.001	0.499	−0.267
Bi	0.349	−0.126	−0.105
Cd	−0.306	−0.072	−0.024
Au	−0.077	0.375	0.334
Hg	−0.258	0.106	−0.302

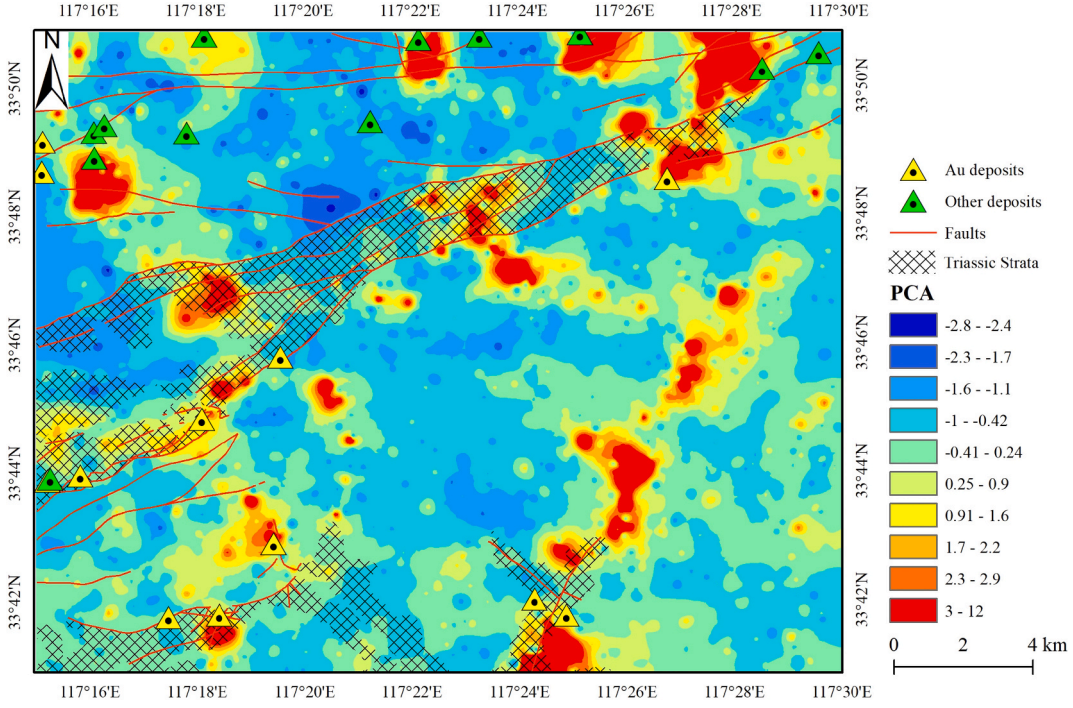

Fig. 7. The spatial distribution of PC2 scores of the ilr-transformed data.

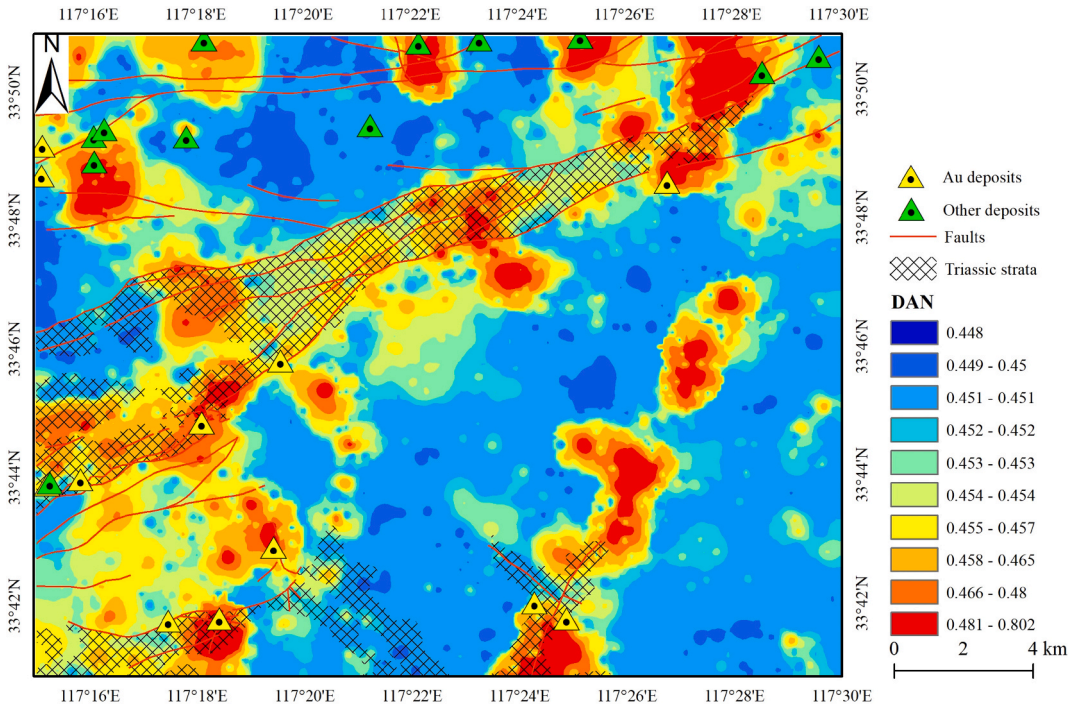

Fig. 8. Geochemical anomalies identified by the deep autoencoder network.

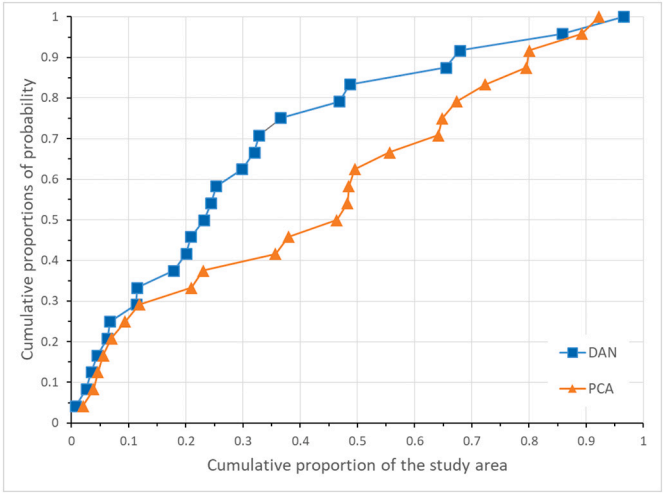

Fig. 9. The success-rate curves for geochemical patterns identified by DAN and PCA.

Declaration of competing interest

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

Acknowledgments

We thank Dr. Mohammad Parsa and an anonymous reviewer's comments and suggestions which help us improve this study. This study was supported by National Natural Science Foundation of China (No. 41972303), and MOST Special Fund from the State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences (MSFGPMR03-3).

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B 44, 139–177.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.
Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27, 93–115.
Anselin, L., Getis, A., 1993. Spatial statistical analysis and geographic information systems. In: Geographic Information Systems, Spatial Modelling and Policy Evaluation. Springer, pp. 35–49.
eaau0323 Bergen, K.J., Johnson, P.A., Maarten, V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363 (6433).
Buccianti, A., Pawlowsky-Glahn, V., 2005. New perspectives on water chemistry and compositional data analysis. Math. Geol. 37 (7), 703–727.
Chen, C.P., Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. 275, 314–347.
Chen, H., Murray, A.F., 2003. Continuous restricted Boltzmann machine with an implementable training algorithm. IEE Proc. Vision Image Sig. Process 150, 153–158.
Chen, L., Guan, Q., Xiong, Y., Liang, J., Wang, Y., Xu, Y., 2019. A Spatially Constrained Multi-Autoencoder approach for multivariate geochemical anomaly recognition. Comput. Geosci. 125, 43–54.
Chen, Y., Santosh, M., 2014. Triassic tectonics and mineral systems in the Qinling Orogen, central China. Geol. J. 49, 338–358.
Chen, Y., Wu, W., 2017. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. Geochem. Explor. Environ. Anal. 17, 231–238.
Chen, Y., Lu, L., Li, X., 2014. Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly. J. Geochem. Explor. 140, 56–63.
Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32, 314–324.
Cheng, Q., Agterberg, F.P., Ballantyne, S.B., 1994. The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51, 109–130.
Cheng, Q., Xu, Y., Grunsky, E., 2000. Integrated spatial and spectrum method for geochemical anomaly separation. Nat. Resour. Res. 9, 43–52.
Cheng, Q., Oberhänsli, R., Zhao, M., 2020. A new international initiative for facilitating data-driven Earth science transformation. Geol. Soc. Lond., Spec. Publ. 499, 225–240.
Cohen, D.R., Kelley, D.L., Anand, R., Coker, W.B., 2010. Major advances in exploration geochemistry, 1998–2007. Geochem. Explor. Environ. Anal. 10, 3–16.
Darnley, A.G., Björklund, A., Bölviken, B., Gustavsson, N., Koval, P.V., Plant, J.A., Steenfelt, A., Tauchid, M., Xie, X., Garrett, R.G., Hall, G.E.M., 1995. A Global Geochemical Database for Environmental and Resource Management Final Report of IGCP Project 259, Earth Sciences, 19. UNESCO Publishing, Paris.
de Caritat, P., Cooper, M., Pappas, W., Thun, C., Webber, E., 2010. National geochemical survey of Australia: analytical methods manual. In: Geoscience Australia Record (2010/15 (22 pp.)).

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.
Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics 20, 621–632.
Galuszka, A., Migaszewski, Z., 2011. Geochemical background - an environmental perspective. Mineralogia 42, 7–17.
Gibert, K., Horsburgh, J.S., Athanasiadis, I.N., Holmes, G., 2018. Environmental data science. Environ. Model Softw. 106, 4–12.
Goldfarb, R.J., Taylor, R.D., Collins, G.S., Goryachev, N.A., Orlandini, O.F., 2014. Phanerozoic continental growth and gold metallogeny of Asia. Gondwana Res. 25, 48–102.
Gonbadi, A.M., Tabatabaei, S.H., Carranza, E.J.M., 2015. Supervised geochemical anomaly detection by pattern recognition. J. Geochem. Explor. 157, 81–91.
Grunsky, E.C., Arne, D., 2020. Mineral-resource prediction using advanced data analytics and machine learning of the QUEST-South stream-sediment geochemical data, southwestern British Columbia, Canada. Geochem. Explor. Environ. Anal. 21, m2020–m2054.
Grunsky, E.C., de Caritat, P., 2020. State-of-the-art analysis of geochemical data for mineral exploration. Geochem. Explor. Environ. Anal. 20, 217–232.
Guo, H., Wang, L., Chen, F., Liang, D., 2014. Scientific big data and digital earth. Chin. Sci. Bull. 59, 5066–5073.
Hawkes, H.E., Webb, J.S., 1963. Geochemistry in mineral exploration. Soil Sci. 95, 283.
Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Comput. 14, 1771–1800.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences 374, 201502022065.
Kadel-Harder, I.M., Spry, P.G., Mccombs, A.L., Zhang, H., 2020. Identifying pathfinder elements for gold in bulk-rock geochemical data from the Cripple Creek Au-Te deposit: a statistical approach. Geochem. Explor. Environ. Anal. 21, m2020–m2048.
Liu, J., Dai, H., Zhai, D., Wang, J., Wang, Y., Yang, L., Mao, G., Liu, X., Liao, Y., Yu, C., Li, Q., 2015. Geological and geochemical characteristics and formation mechanisms of the Zhaishang Carlin-like type gold deposit, western Qinling Mountains, China. Ore Geol. Rev. 64, 273–298.
Luo, Z., Xiong, Y., Zuo, R., 2020. Recognition of geochemical anomalies using a deep variational autoencoder network. Appl. Geochem. 122, 104710.
Luo, Z., Zuo, R., Xiong, Y., Wang, X., 2021. Detection of geochemical anomalies related to mineralization using the GANomaly network. Appl. Geochem. 131, 105043.
Mao, J.W., Qiu, Y.M., Goldfarb, R.J., Zhang, Z.C., Garwin, S., Ren, F.S., 2002. Geology, distribution, and classification of gold deposits in the western Qinling belt, Central China. Mineral. Deposita 37, 352–377.
Matheron, G., 1962. Traité de géostatistique appliquée, 1. Editions Technip.
Miesch, A.T., 1981. Estimation of the geochemical threshold and its statistical significance. J. Geochem. Explor. 16, 49–76.
Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.
Parsa, M., 2021. A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted Zn-Pb mineral systems of Western Iran. J. Geochem. Explor. 228, 106811.
Parsa, M., Carranza, E.J.M., 2021. Modulating the impacts of stochastic uncertainties linked to deposit locations in data-driven predictive mapping of mineral prospectivity. Nat. Resour. Res. https://doi.org/10.1007/s11053-021-09891-9.
Parsa, M., Pour, A.B., 2021. A simulation-based framework for modulating the effects of subjectivity in greenfield mineral prospectivity mapping with geochemical and geological data. J. Geochem. Explor. 229, 106838.
Parsa, M., Maghsoudi, A., Carranza, E.J.M., Yousefi, M., 2017. Enhancement and mapping of weak multivariate stream sediment geochemical anomalies in Ahar Area, NW Iran. Nat. Resour. Res. 26, 443–455.
Parsa, M., Maghsoudi, A., Yousefi, M., 2018. Spatial analyses of exploration evidence data to model skarn-type copper prospectivity in the Varzaghan district, NW Iran. Ore Geol. Rev. 92, 97–112.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley.
Reimann, C., de Caritat, P., GEMAS Project Team, NGSA Project Team, 2012. New soil composition data for Europe and Australia: demonstrating comparability, identifying continental–scale processes and learning lessons for global geochemical mapping. Sci. Total Environ. 416, 239–252.
Sadr, M.P., Nazeri, M., 2018. Random forests algorithm in podiform chromite prospectivity mapping in Dolatabad area, SE Iran. J. Mining Environ. 9, 403–416.
Shan, L., Zhang, D., Pang, C., Liu, J., Zhang, W., Zhao, X., Zhang, Z., 2016. Late Triassic magmatic activity in the Daqiao gold deposit of West Qinling belt: zircon U-Pb chronology and Lu-Hf isotope evidence. Geol. Bull. China 35, 2045–2057.
Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs. J. Geochem. Explor. 3, 129–149.
Tansley, S., Tolle, K.M., 2009. In: Hey, A.J. (Ed.), The Fourth Paradigm: Data-intensive Scientific Discovery, vol. 1. Microsoft Research, Redmond, WA.
Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.
Tukey, J.W., 1962. The future of data analysis. Ann. Math. Stat. 33, 1–67.
Tukey, J.W., 1977. Exploratory data analysis. J. Am. Stat. Assoc. 28, 1.
Wang, X., The CGB Sampling Team, 2015. China geochemical baselines: sampling methodology. J. Geochem. Explor. 148, 25–39.
Wang, Z., Dong, Y., Zuo, R., 2019a. Mapping geochemical anomalies related to Fe–polymetallic mineralization using the maximum margin metric learning method. Ore Geol. Rev. 107, 258–265.
Wang, Z., Zuo, R., Dong, Y., 2019b. Mapping geochemical anomalies through integrating random forest and metric learning methods. Nat. Resour. Res. 28, 1285–1298.
Wu, W., Chen, Y., 2018. Application of isolation forest to extract multivariate anomalies from geochemical exploration data. Glob. Geol. 21, 36–47.
Wu, Y., 2019. Genesis of the World-class Daqiao Gold Deposit, West Qinling Orogen, China. Doctoral thesis. China University of Geosciences, Wuhan.
Wu, Y., Li, J., Evans, K., Koenig, A.E., Li, Z., O'Brien, H., Lahaye, Y., Rempel, K., Hu, S., Zhang, Z., Yu, J., 2018. Ore-forming processes of the Daqiao Epizonal Orogenic Gold Deposit, West Qinling Orogen, China: constraints from textures, trace elements, and sulfur isotopes of pyrite and marcasite, and raman spectroscopy of carbonaceous material. Econ. Geol. 113, 1093–1132.
Wu, Y., Li, J., Evans, K., Fougerouse, D., Rempel, K., 2019. Source and possible tectonic driver for Jurassic-cretaceous gold deposits in the West Qinling Orogen, China. Geosci. Front. 10, 107–117.
Wu, Y., Evans, K., Fisher, L.A., Zhou, M., Hu, S., Fougerouse, D., Large, R.R., Li, J., 2020. Distribution of trace elements between carbonaceous matter and sulfides in a sediment-hosted orogenic gold system. Geochim. Cosmochim. Acta 276, 345–362.
Xie, X., Cheng, H., 2001. Global geochemical mapping and its implementation in the Asia-Pacific region. Appl. Geochem. 16, 1309–1321.
Xie, X., Mu, X., Ren, T., 1997. Geochemical mapping in China. J. Geochem. Explor. 60, 99–113.
Xiong, Y., Zuo, R., 2016. Recognition of geochemical anomalies using a deep autoencoder network. Comput. Geosci. 86, 75–82.
Xiong, Y., Zuo, R., 2020. Recognizing multivariate geochemical anomalies for mineral exploration by combining deep learning and one-class support vector machine. Comput. Geosci. 140, 104484.
Xiong, Y., Zuo, R., 2021. Robust feature extraction for geochemical anomaly recognition using a stacked convolutional denoising autoencoder. Math. Geosci. https://doi.org/10.1007/s11004-021-09935-z.
Xiong, Y., Zuo, R., Wang, K., Wang, J., 2018. Identification of geochemical anomalies via local RX anomaly detector. J. Geochem. Explor. 189, 64–71.
You, G., Zhang, Z., 2009. Geological characteristics of Daqiao Gold Deposit in Gansu Province and its significance in prospecting for gold deposit. Gansu Geol. 18, 1–8.
Yu, X., Xiao, F., Zhou, Y., Wang, Y., Wang, K., 2019. Application of hierarchical clustering, singularity mapping, and Kohonen neural network to identify Ag-Au-Pb-Zn polymetallic mineralization associated geochemical anomaly in Pangxidong district. J. Geochem. Explor. 203, 87–95.
Zhang, C., Zuo, R., 2021. Recognition of multivariate geochemical anomalies associated with mineralization using an improved generative adversarial network. Ore Geol. Rev. 136, 104264.
Zhang, C., Zuo, R., Xiong, Y., 2021. Detection of the multivariate geochemical anomalies associated with mineralization using a deep convolutional neural network and a pixel-pair feature method. Appl. Geochem. 130, 104994.
Zhang, D., 2016. Geological and Geochemical Characteristics and Genesis of the Daqiao Gold Deposit in Gansu Province. Master thesis. China University of Geosciences, Beijing (In Chinese with English abstract).
Zhang, F., Wu, Y., Zhang, Y., Liu, Y., 2015. Geochemical anomaly characteristics of Daqiao gold deposit in Gansu Province. Gansu Geol. 24, 36–41.
Zhang, Z., Wu, Y., Li, J., 2018. Characteristics and genesis of the Silicified Breccias in the Daqiao gold deposit, West Qinling Orogen. Geol. Sci. Technol. Inf. 37, 79–88.
Ziaii, M., Pouyan, A.A., Ziaei, M., 2009. Neuro-fuzzy modelling in mining geochemistry: identification of geochemical anomalies. J. Geochem. Explor. 100, 25–36.
Zuo, R., 2011. Identifying geochemical anomalies associated with Cu and Pb-Zn skarn mineralization using principal component analysis and spectrum-area fractal modeling in the Gangdese Belt, Tibet (China). J. Geochem. Explor. 111, 13–22.
Zuo, R., 2014. Identification of geochemical anomalies associated with mineralization in the Fanshan district, Fujian, China. J. Geochem. Explor. 139, 170–176.
Zuo, R., 2017. Machine learning of mineralization-related geochemical anomalies: a review of potential methods. Nat. Resour. Res. 26, 457–464.
Zuo, R., 2020. Geodata science-based mineral prospectivity mapping: a review. Nat. Resour. Res. 29, 3415–3424.
Zuo, R., Xiong, Y., 2018. Big data analytics of identifying geochemical anomalies supported by machine learning methods. Nat. Resour. Res. 27, 5–13.
Zuo, R., Xiong, Y., 2020. Geodata science and geochemical mapping. J. Geochem. Explor. 209, 106431.
Zuo, R., Xia, Q., Wang, H., 2013. Compositional data analysis in the study of integrated geochemical anomalies associated with mineralization. Appl. Geochem. 28, 202–211.
Zuo, R., Carranza, E.J.M., Wang, J., 2016. Spatial analysis and visualization of exploration geochemical data. Earth Sci. Rev. 158, 9–18.
Zuo, R., Xiong, Y., Wang, J., Carranza, E.J.M., 2019. Deep learning and its application in geochemical mapping. Earth Sci. Rev. 192, 1–14.