Front. Earth Sci. 2018, 12(3): 491–505
https://doi.org/10.1007/s11707-017-0682-8

RESEARCH ARTICLE

Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies

Shuguang ZHOU (✉)[1,2], Kefa ZHOU[1], Jinlin WANG[1], Genfang YANG[1,3], Shanshan WANG[1]

1 Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China
2 Xinjiang Laboratory of Mineral Resources and Digital Geology, Urumqi 830011, China
3 Laboratory of Desert Environment & Engineering, Urumqi 830011, China

Abstract Cluster analysis is a well-known technique that is used to analyze various types of data. In this study, cluster analysis is applied to geochemical data that describe 1444 stream sediment samples collected in northwestern Xinjiang with a sample spacing of approximately 2 km. Three algorithms (the hierarchical, k-means, and fuzzy c-means algorithms) and six data transformation methods (the *z*-score standardization, ZST; the logarithmic transformation, LT; the additive log-ratio transformation, ALT; the centered log-ratio transformation, CLT; the isometric log-ratio transformation, ILT; and no transformation, NT) are compared in terms of their effects on the cluster analysis of the geochemical compositional data. The study shows that, on the one hand, the ZST does not affect the results of column- or variable-based (R-type) cluster analysis, whereas the other methods, including the LT, the ALT, and the CLT, have substantial effects on the results. On the other hand, the results of the row- or observation-based (Q-type) cluster analysis obtained from the geochemical data after applying NT and the ZST are relatively poor. However, we derive some improved results from the geochemical data after applying the CLT, the ILT, the LT, and the ALT. Moreover, the k-means and fuzzy c-means clustering algorithms are more reliable than the hierarchical algorithm when they are used to cluster the geochemical data. We apply cluster analysis to the geochemical data to explore for Au deposits within the study area, and we obtain a good correlation between the results retrieved by combining the CLT or the ILT with the k-means or fuzzy c-means algorithms and the potential zones of Au mineralization. Therefore, we suggest that the combination of the CLT or the ILT with the k-means or fuzzy c-means algorithms is an effective tool to identify potential zones of mineralization from geochemical data.

Received April 13, 2017; accepted October 8, 2017

E-mail: zhoushuguang1010@163.com

Keywords cluster analysis, compositional data, geochemical anomaly, mineral exploration

1 Introduction

Cluster analysis mainly aims to classify variables or observations into meaningful multivariate homogeneous groups such that the members of individual groups are distinguishable from the members of other groups. The observations are mapped to clusters with centroids that summarize the cluster information, providing an overview of the structure of the data. In general, cluster analysis algorithms can be classified into hierarchical and non-hierarchical algorithms. In greater detail, these algorithms can be classified into groups that include partitioning, hierarchical, density-based, grid-based, model-based, constraint-based, and high-dimensional clustering algorithms (Han and Kamber, 2006; Meng et al., 2011).

Cluster analysis has been applied in many well-known fields (Kim et al., 2007; Lee and Song, 2007; Sahraei Parizi and Samani, 2013; Stück et al., 2013; Ghosh and Kanchan, 2014; Wang et al., 2014; Fattahi, 2016; Eilermann et al., 2017; Fatehi and Asadi, 2017; Kitzig et al., 2017), but controversies exist regarding the merits of this analytical method (Rock, 1988; Davis, 2002). For instance, different clustering algorithms yield different groupings from the same data (Templ et al., 2008). Furthermore, cluster analysis is often affected by other issues, such as the possibility that a different result will be obtained if even a single variable is added or removed (Templ et al., 2008). Additionally, the results of cluster analysis may differ with respect to the diverse parameters (the distance between or similarities among the variables or observations considered) used for the analysis (Abdel-Halim and Abdel-Aal, 1998). The distributional characteristics of the data and the method of data preparation may also affect the results of

cluster analysis, since some clustering algorithms are reliable only for normally distributed data (Reimann et al., 2002). Such algorithms are not appropriate for use with geochemical data (Reimann and Filzmoser, 2000; Zuo et al., 2013a). Furthermore, geochemical data are frequently made up of compositional information expressed in concentrations (i.e., wt.% or mg/kg) that sum up to a constant (Leite, 2016; Tolosana-Delgado and McKinley, 2016; Templ et al., 2017). Consequently, the multivariate statistical methods that are used to evaluate the data may produce biased results (Filzmoser et al., 2010; Buccianti, 2013). Therefore, appropriate data transformations must be applied to open the data prior to performing cluster analysis (Aitchison, 1982; Aitchison, 1999; Aitchison et al., 2000; Aitchison and Egozcue, 2005; Filzmoser et al., 2009; Zuo et al., 2013a). Some studies perform many repeated experiments using different clustering techniques and selections of variables until the results of the cluster analysis fit the preconceived ideas of the authors (Templ et al., 2008). However, all these previous studies indicate that cluster analysis can be applied as an exploratory data analysis tool for investigating the behavior of individual data sets and extracting different types of information from them.

Detecting outliers or anomalies is one of the main tasks in the statistical analysis of geochemical data. The data obtained from stream sediment reconnaissance surveys provide useful information on the regional controls on mineralization or the occurrences of deposits. Hence, such data can be used as a basis for mineral prospecting. In well-prospected areas of mineralization, the main objective is to identify new target areas, whereas areas without mineralization are examined to distinguish anomalies from the background (Carranza and Hale, 1997). Several methods can be applied to select thresholds for identifying outliers or anomalies. For example, 'mean±2 standard deviation (sdev)' is recommended as a threshold that separates anomalies from the background (Hawkes and Webb, 1962). The values within the range of 'mean±2 sdev' are defined as belonging to the background, whereas those values that lie outside of this range are considered to be anomalies. However, the threshold can be significantly influenced by extreme values. Thus, removing obvious outliers or anomalies from the data and carrying out a logarithmic transformation prior to calculating the 'mean±2 sdev' are two methods that are commonly used to avoid or mitigate inaccuracies. However, some studies do not recommend this method, because other methods that can perform such tasks more accurately are readily available (Reimann and Garrett, 2005).

Exploratory data analysis (EDA) (Tukey, 1977) is an unconventional and informal approach to analyze and interpret univariate data that do not follow a normal distribution (Carranza, 2009). It can be useful in the analysis and modeling of single-element geochemical anomalies (Yusta et al., 1998; Bounessah and Atkin, 2003; Reimann et al., 2005; Reimann and Garrett, 2005; Zumlot et al., 2009; Agharezaei and Hezarkhani, 2016). Both 'mean±2 sdev' and EDA are based on data values and only a single threshold is derived from each, which disregards the condition that the background may be spatially variable. Some studies have applied fractal and multi-fractal models to characterize the concentrations of geochemical elements in various environments, such as in soil and stream sediments (Cheng et al., 1996; Zuo and Cheng, 2008; Cheng and Agterberg, 2009; Afzal et al., 2010; Carranza, 2010, 2011; Pazand et al., 2011; Agterberg, 2012; Zuo, 2012, 2013a, b, 2015; Hassanpour and Afzal, 2013; Parsa et al., 2017). Fractal and multi-fractal models are suitable for analyzing geochemical data because geochemical distribution patterns (or geochemical landscapes) are regarded as plausibly consisting of fractals (background and anomaly patterns) (Bölviken et al., 1992). However, none of the methods mentioned above are especially appropriate for multivariate analysis. On the other hand, cluster analysis is designed for multivariate analysis; thus, it is considered to be a relevant method for identifying geochemical anomalies (Howarth, 1983).

In this study, cluster analysis is applied to analyze geochemical data and to identify possible anomalies from these data. Three different clustering algorithms, the hierarchical, k-means, and fuzzy c-means algorithms, are used to analyze geochemical data obtained from stream sediment collected in Karamay, northwestern Xinjiang, China. To compare the performance of these three algorithms, they are applied to the geochemical data to which six different transformations (the *z*-score standardization, ZST; the logarithmic transformation, LT; the additive log-ratio transformation, ALT; the centered log-ratio transformation, CLT; the isometric log-ratio transformation, ILT; and no transformation, NT) have been applied. In addition, the study also addresses the following critical questions related to cluster analysis: (i) How do the data transformation methods and clustering algorithms used affect the results of applying cluster analysis to geochemical data? (ii) How suitable are the clustering algorithms for extracting mineral-related information from geochemical data?

2 Description of the study area

2.1 Geology

The study area is located in the western Junggar Basin, approximately 330 km to the northwest of Urumqi (Fig. 1). The major rock bodies in this area include ophiolitic mélange belts, volcanic-sedimentary rocks, and intermediate-acidic intrusive bodies that formed due to the amalgamation of island arcs and accretionary processes

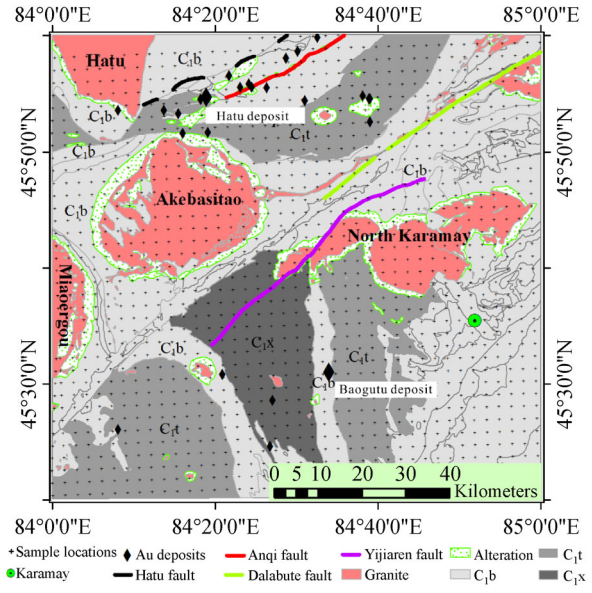

Fig. 1 Simplified geological map of Karamay District. C_1x, Xibeikula formation. C_1t, Tailegula formation. C_1b, Baogutu formation.

(He et al., 2004; Zhu et al., 2013b). Major plutonic rocks are represented in this area by the Miaoergou, Hatu, Akebasitao, Red Mountain and north Karamay granite batholiths, which have an age of 300 Ma, as determined from zircon LA-ICP-MS U-Pb techniques (Su et al., 2006). The distributions of intrusive rocks and Au deposits are clearly correlated with fault zones in this region. The major faults, including the Hatu, Anqi, Dalabute, and Yijiaren faults, strike NNE. The Dalabute ophiolitic mélange belt, which is approximately 50 km^2 in size and is extensively crosscut by imbricate structures with thrust faults, is distributed as a band along the Dalabute fault. Oceanic crust materials often occur in the terrigenous detrital sediments of the ancient continental margin, and these materials display the geochemical signatures of mantle-derived rocks (Zhang and Huang, 1992).

2.2 Au deposits

The Hatu and Baogutu Au deposits are located in the northern and southern parts of the study area, respectively, and they are representative deposits in this study area (Fig. 1). The distribution of the Hatu Au deposit is controlled by two NE-trending faults, specifically the Anqi extension fault and the Hatu compression and scissor fault. The metallogenic structures are associated with NW-, NE, and E-W-trending secondary faults, which are closely related to the NE-trending fault. The ore bodies are concentrated as groups and display en echelon and/or end-to-end alignment (Zhu et al., 2013a). The Hatu Au deposit is mainly composed of quartz vein-type and altered rock-type ore bodies, which are the products of homogeneous hydrothermal flow (Zhang, 2003). The gangue minerals of these two types of ore bodies are mainly quartz, albite, sericite, and carbonate minerals. Two types of native Au, encapsulated and fissure-filling Au, are found in these ore bodies, and arsenopyrite, pyrite, and quartz are the main Au-bearing minerals. The range of mineralization temperatures (200°C–280°C) is consistent between the main metallogenic stage of the quartz vein-type and altered rock-type ore bodies. Moreover, calcite veins were produced during the late stage of mineralization of these two types of ore bodies, and no Au is present in the calcite veins (Zhu et al., 2013a). Copper, Ag, As, and Sb appear in relatively high concentrations to varying degrees in altered Carboniferous tuffaceous shale.

The Baogutu Au deposit occurs in the Lower Carboniferous strata. There are two types of Au mineralizations in the Baogutu area. The first type is located in an area where intermediate and acidic dikes are concentrated and controlled by NE-trending faults, and it includes quartz stockwork ore bodies and quartz vein-type ore bodies. The second type is located in the contact zone between porphyry bodies and the surrounding host rock, and the Au is hosted in sulfide minerals. Andesite, tuff, and tuffaceous sandstone are the Au-bearing rocks, and the ore bodies are sulfide-bearing. Silicification, sericitization, carbonatation, pyritization, and arsenopyritization are the main modes of alteration of the host rock, and the primary mineral assemblage is arsenopyrite + pyrite + native Au + native arsenic + native antimony + stibnite. The ore-associated elements are Au, As, and Sb (Zhu et al., 2013a).

3 Materials and methods

3.1 Geochemical data

The geochemical data used in this study include 39 elements or variables that were determined from 1444 stream sediment samples, which were collected over an area of 5774 km^2 (Fig. 1). The data were acquired from the National Geochemical Mapping Project of China, also known as the Regional Geochemistry-National Reconnaissance Project (Xie et al., 1997, 2008). The stream sediment samples were mainly collected from natural gullies with a grid spacing of approximately one sample per km^2. Groups of four samples were then combined into single samples that each represent 4 km^2.

The concentrations of the 39 elements in the stream sediments were measured using various facilities (Wang et al., 2011), and more details are listed in Table 1. The maximum and minimum concentrations of the 39 elements and their standard deviations are shown in Table 2.

3.2 Distance measures

Measuring the distance among observations or variables is

Table 1 List of analytical equipment that was used to measure the concentrations of the 39 elements (modified from Xie et al., 2008)

Analytical method	No.	Elements determined
ICP-Ms [a)]	11	Bi, Cd, Co, Cu, La, Mo, Nb, Pb, Th, U, W
ICP-AES [b)]	11	Ba, Be, Ca, Li, Mg, Mn, Na, Ni, Sr, V, Zn
XRF [c)]	9	Al, Cr, Fe, K, P, Si, Ti, Y, Zr
ES [d)]	3	Ag, B, Sn
HG-AFS [e)]	2	As, Sb
GF-AAS [f)]	1	Au
CV-AFS [g)]	1	Hg
ISE [h)]	1	F
Total elements	39	

a) ICP-MS, inductively coupled plasma–mass spectrometry; b) ICP-AES, inductively coupled plasma–atomic emission spectrometry; c) XRF, X-ray fluorescence spectrometry; d) ES, emission spectrometry; e) HG-AFS, hydride generation–atomic fluorescence spectrometry; f) GF-AAS, graphite furnace–atomic absorption spectrometry; g) CV-AFS, cold vapor–atomic fluorescence spectrometry; h) ISE, ion selective electrode.

a key point in most types of cluster analysis. Note that the distance in cluster analysis does not reflect the geographical distance between two observations or variables; instead, it is a measure of the similarity among observations or variables. This distance is the basis of classifying different observations or variables into different groups. Small distances reflect similar observations, whereas large distances indicate possibly dissimilar observations. In this study, the distance between two variables is defined as one minus the correlation coefficient of each pair of variables, whereas the Euclidean, squared Euclidean, and correlation coefficient distances are used to classify the observations.

4 Clustering the geochemical compositional data

4.1 Clustering by variables (*R*-type cluster analysis)

To characterize the influence of the characteristics of the geochemical data (e.g., abnormal distributions and compositional properties) on the cluster analysis of the variables, five of the transformation methods (NT, the ZST, the LT, the ALT, and the CLT) are applied to the data describing the 39 elements or variables before the hierarchical cluster analysis algorithm is used. For this purpose, the Matlab R2012a software package is used for the ZST and the LT, whereas the R software package is used for the ALT, the CLT, and the ILT. Notably, the results of the ALT are affected by the ratio element used in this transformation (Templ et al., 2008); any element can be selected as the ratio element except the target element or

Table 2 The measured concentrations of the 39 elements in the stream sediments with their minimum and maximum values (modified from Wang et al., 2011)

Ele [a)]	25%	50%	75%	90%	95%	MAD	Ele [a)]	25%	50%	75%	90%	95%	MAD
Ag	70	80	90	100	100	10	Pb	14	17.3	20.7	24.7	27.3	3.3
As	9.3	11.2	13.7	18.2	22	2.1	Sb	0.7	0.8	1	1.2	1.4	0.11
Au	1	1.6	2.7	5.8	12	0.7	Sn	1.8	2.1	2.7	3.5	4.3	0.4
B	40.7	50.7	62.1	75	83.9	10.7	Sr	230	281	330	380	413	49
Ba	500	569	650	780	926	78.5	Th	6.7	8.1	9.6	11.5	12.6	1.4
Be	1.5	1.8	2.1	2.5	2.7	0.3	Ti	3570	3990	4670	5530	6241	507
Bi	0.2	0.2	0.3	0.3	0.4	0.03	U	1.8	2	2.3	2.9	3.4	0.3
Cd	200	230	300	400	400	70	V	80	88	100	115	130	10
Co	11.6	13.8	16.2	19.3	22.2	2.3	W	1.3	1.5	1.8	2.2	2.5	0.3
Cr	38.2	48	60	82	113	10	Y	23.3	28.3	34.8	42.9	52.5	5.45
Cu	33	38	43.8	49	53	5.1	Zn	84	94	106	125	140	11
F	505	560	620	706	790	55	Zr	159	190	235	316	405	37
Hg	17	24.9	34.8	45	51.5	8.15	Al_2O_3	12.8	13.4	14	14.5	14.8	0.6
La	25	29	34	40	46	4.4	CaO	3.53	4.9	6.4	7.69	8.4	1.41
Li	26	29	32	35.5	39	3	Fe_2O_3	4.66	5.06	5.48	6.04	6.48	0.41
Mn	849	978	1150	1300	1450	142	K_2O	2.59	2.74	2.86	3	3.1	0.15
Mo	1.3	1.6	1.9	2.4	2.9	0.3	MgO	2.07	2.42	2.81	3.47	4.01	0.36
Nb	9.5	11.5	13.7	16	17.3	2.1	Na_2O	2.2	2.44	2.74	3.03	3.29	0.28
Ni	22.1	25.8	31.4	44.4	69.8	4.45	SiO_2	55.8	58.2	60.4	62.5	64.3	2.31
P	890	1026	1182	1339	1440	152							

a) Ele represents Element, and the units of the elements are μg/g, apart from the oxides (%), Ag (ng/g), Au (ng/g) and Hg (ng/g). 25%, 50%, 75%, 90%, 95% are quantiles. MAD, median absolute deviation.

the element of interest (i.e., those related to the Au deposit or hydrothermal or epithermal elements). In this study, SiO_2 is selected as the ratio element. The ILT is not suitable for R-type cluster analysis, considering that the relationship among the original variables will be lost and the newly produced variables will no longer be directly interpretable in terms of the originally entered variables (Templ et al., 2008); therefore, the ILT is not applied here.

The hierarchical algorithm is used to cluster the variables on the basis of the geochemical data preprocessing mentioned above. Building a dendrogram is one of the most important parts of the hierarchical algorithm, and various methods can be used to create dendrograms. In this study, the group average method, in which the distance among the variables is defined as one minus the correlation coefficient of each pair of variables, is selected to create the dendrogram because it is a moderate method of addressing the data space. The results of applying R-type cluster analysis are presented in Fig. 2.

R-type cluster analysis is also applied because it can provide valuable information for dimensional reduction. Such information is very important because the distribution of high-dimensional data is often complex and difficult to fully understand. In the present study, the Au deposit-related and the hydrothermal or epithermal elements are the target elements or the elements of interest, and Q-type cluster analysis (section 4.2) will be implemented based on these elements.

The results of applying R-type cluster analysis to the geochemical data after employing different data transformation methods are shown in Fig. 2. It can be seen that U-Mo-Au-Sb-B-Hg-W-As-Ag (Fig. 2(a)), U-Mo-Au-Sb-B-Hg-W-As-Ag (this figure is the same as that shown in Fig. 2(a)), U-Au-Hg-P-Cd-Th-F-Zn-Li-B-W-Mo-Sb-As (Fig. 2(b)), Au-Mo-Sb-As (Fig. 2(c)), and Mo-Au-As-Sb-Hg-B-Ag (Fig. 2(d)) belong to the relatively consistent group. These groups were determined and selected according to the following conventions:

- All or most of the Ag, As, Au, and Sb are included in the target cluster because they are the primary Au deposit-related elements and hydrothermal/epithermal elements in this study area.
- To reduce the dimensions of geochemical data, the number of elements in the target cluster should be as small as possible. To simplify the data analysis, we neglect the information that is uncorrelated or weakly correlated with the known Au deposits in the study area, which permits better interpretation of the results.

According to this study, R-type cluster analysis produces different results when various data transformation methods are used to analyze the present geochemical data, except when NT and the ZST are used. In this study, U-Mo-Au-Sb-B-Hg-W-As-Ag (i.e., the results of applying R-type cluster analysis to the output from NT) is selected for the following section, which is titled "Clustering by observations".

4.2 Clustering by observations (Q-type cluster analysis)

To classify the geochemical data, the hierarchical, k-means, and fuzzy c-means algorithms are applied to the U-Mo-Au-Sb-B-Hg-W-As-Ag group. However, the inverse distance weighted (IDW) interpolation method is applied to the selected elements before Q-type cluster analysis is employed to produce raster data for each element, which are helpful in mapping and interpreting the results. The spatial resolution of the raster data is determined to be approximately 2 km, according to the distance between each pair of adjacent sampling sites in this study. Second, the number of clusters used in Q-type cluster analysis needs to be defined in advance; However, the appropriate number of clusters is unknown prior to the calculation. The solutions for this issue are as follows.

For the hierarchical algorithm, 1) the results of applying Q-type cluster analysis to the results of NT, the ZST, the LT, the ALT, the CLT, and the ILT are all determined by increasing the inconsistent coefficient from zero until the number of clusters stabilizes; 2) the number of clusters obtained using the results of NT, the ZST, the LT, the ALT, the CLT, and the ILT should be nearly equal.

For the k-means and fuzzy c-means algorithms, each set of transformed geochemical data (i.e., the results of NT, the ZST, the LT, the ALT, the CLT, and the ILT) are classified into the same number of clusters to compare the results obtained using Q-type cluster analysis.

Based on the solutions mentioned above, the inconsistent coefficients of the hierarchical algorithm are determined to be 10, 9, 6, 6, 6, and 6 for NT, the ZST, the LT, the ALT, the CLT, and the ILT, respectively, and the level of inconsistency is determined to be 50. Thus, 12, 12, 13, 12, 11, and 11 clusters are produced for NT, the ZST, the LT, the ALT, the CLT, and the ILT, respectively. The cluster(s) that contain(s) a smaller number of observations is/are reclassified into a single cluster that is defined as an anomaly, whereas the cluster(s) that contain(s) most of the observations is/are defined as representing the background conditions.

The results of Q-type cluster analysis differ from each other, depending on whether the hierarchical (Fig. 3(a)), k-means (Fig. 3(c)), or fuzzy c-means (Fig. 3(e)) algorithms are applied to the output from NT. The known Au deposits are not associated with any of the clusters shown in Figs. 3(a) and 3(c). However, there is a relatively good relationship between the known Au deposits and cluster 1 in Fig. 3(e) (21 known Au deposits are located in cluster 1, representing 78% of the total number). Moreover, the results of applying Q-type cluster analysis to the output of the ZST data with the hierarchical (Fig. 3(b)), k-means (Fig. 3(d)), and fuzzy c-means (Fig. 3(f)) algorithms also differ from one another, and the known Au deposits are not associated with any of the clusters shown in Figs. 3(b), 3(d), or 3(f).

The results of applying Q-type cluster analysis with the

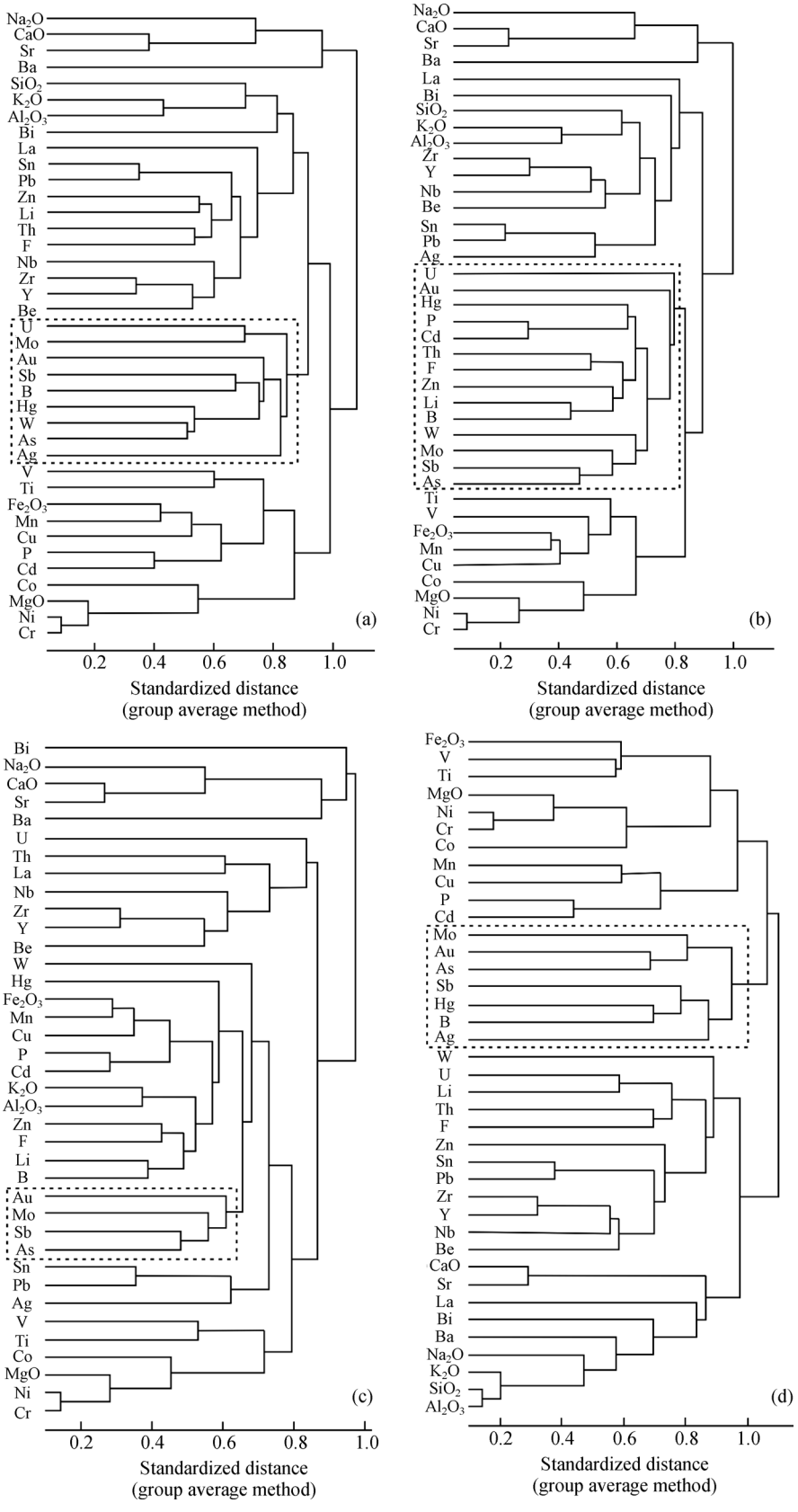

Fig. 2 Results of applying R-type cluster analysis to the geochemical data. ((a), (b), (c), and (d) correspond to the results from NT, the LT, the ALT and the CLT, respectively).

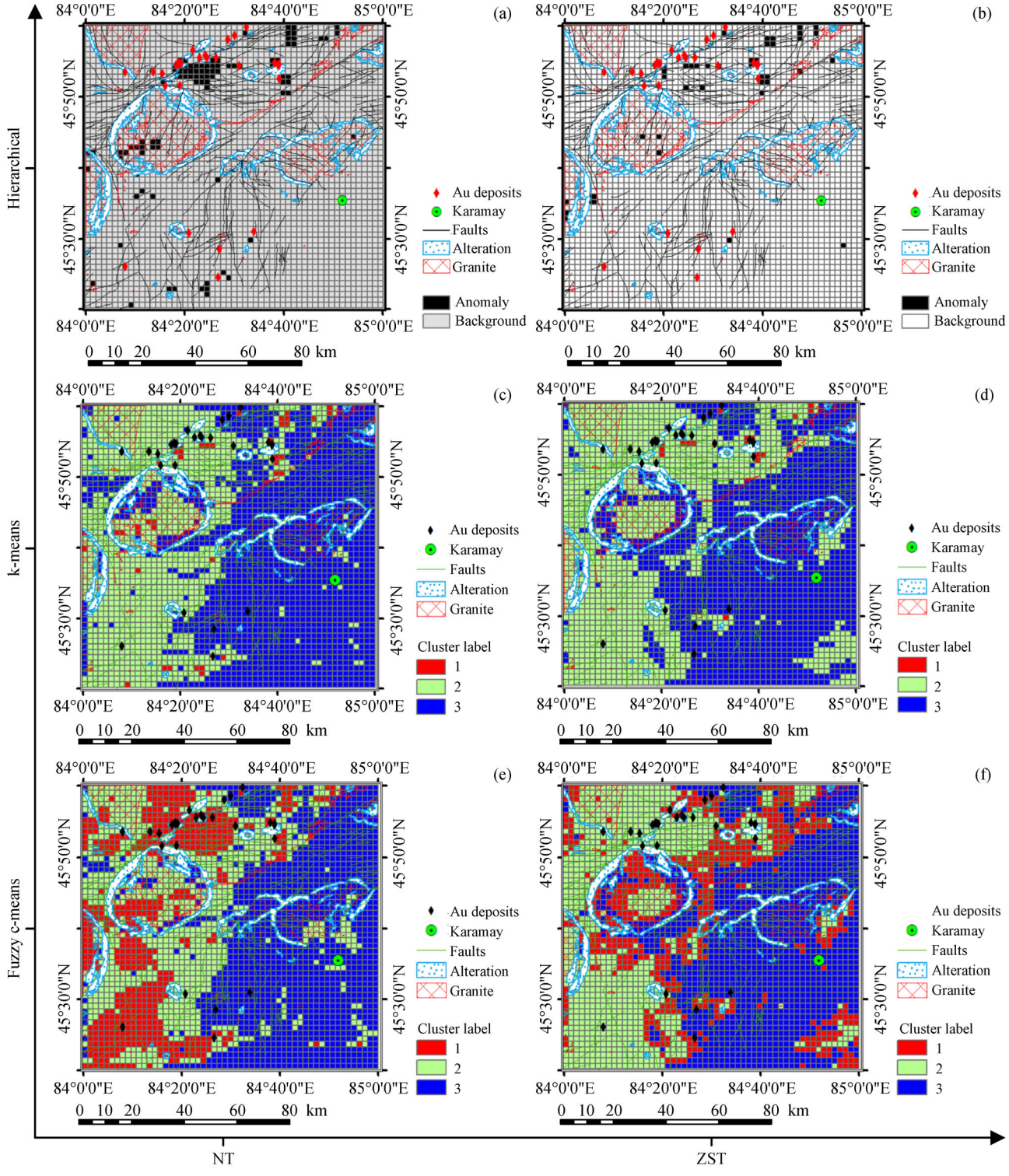

Fig. 3 The results of Q-type cluster analysis obtained by applying the hierarchical ((a) and (b)), k-means ((c) and (d)) and fuzzy c-means ((e) and (f)) algorithms to the output from NT ((a), (c), and (e)) and the ZST ((b), (d), and (f)).

hierarchical (Fig. 4(a)), k-means (Fig. 4(c)), and fuzzy c-means (Fig. 4(e)) algorithms to the output from the LT differ from each other. However, the known Au deposits are associated with the anomaly clusters in Fig. 4(a) or cluster 1 in Figs. 4(c) and 4(e), and the extents of cluster 1 in Fig. 4(c) and Fig. 4(e) are similar to each other (they are both mainly distributed in the north-central and south-western parts of the study area). In the same way, the

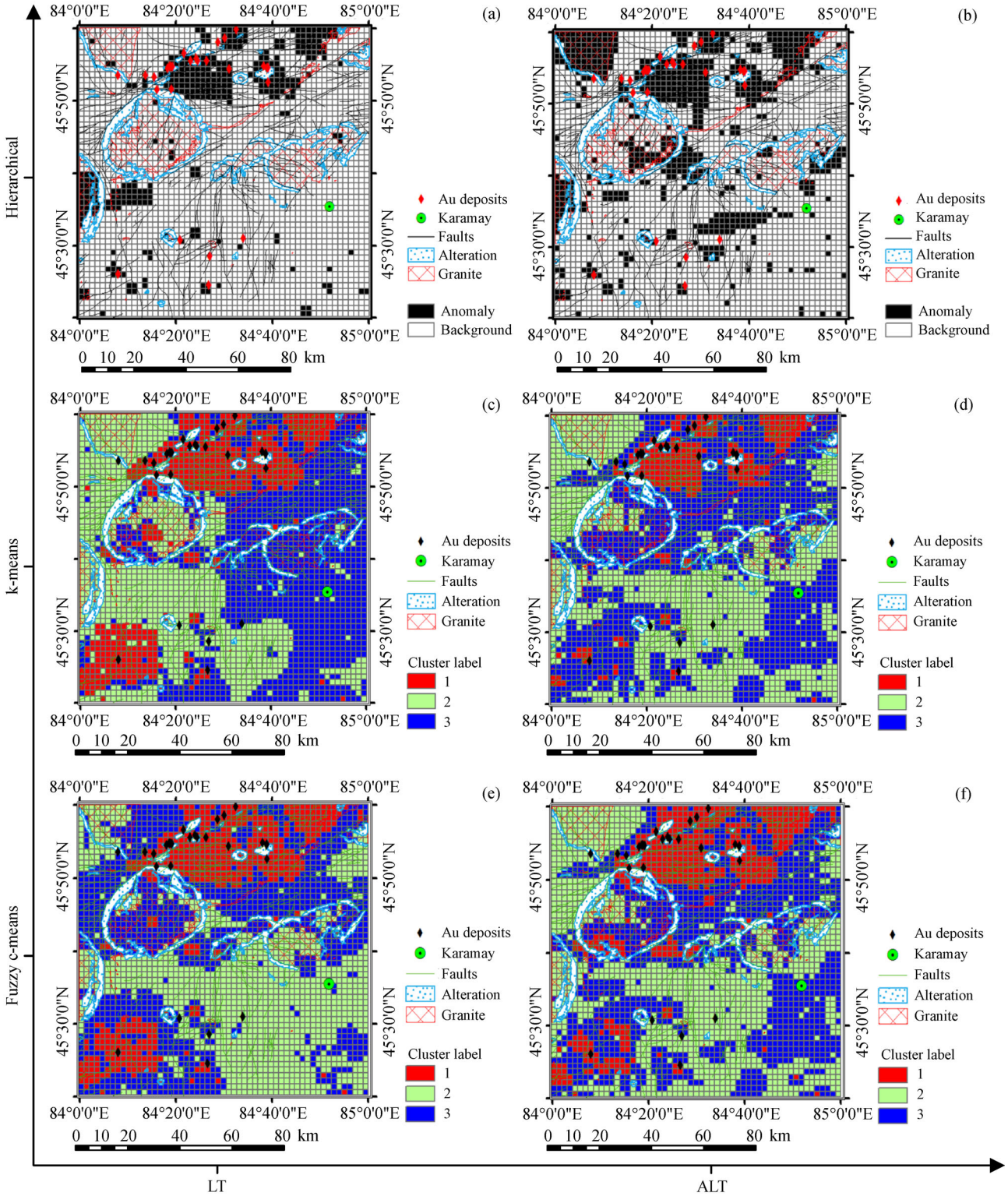

Fig. 4 The results of applying Q-type cluster analysis with the hierarchical ((a) and (b)), k-means ((c) and (d)) and fuzzy c-means ((e) and (f)) algorithms to the output from the LT ((a), (c) and (e)) and the ALT ((b), (d) and (f)).

anomaly clusters (specifically, the anomaly cluster in Fig. 4(b) and cluster 1 in Fig. 4(d) and Fig. 4(f)) of Q-type cluster analysis of ALT data with hierarchical (Fig. 4(b)), k-means (Fig. 4(d)), and fuzzy c-means (Fig. 4(f)) algorithms are also correlated with the known Au deposits. The distribution of cluster 1 in Figs. 4(d) and 4(f) are

similar to each other to some extent. The results of applying Q-type cluster analysis to the output from the LT and the ALT differ from each other, even when the same clustering algorithm is used (Fig. 4). However, notably, there is always a cluster that is clearly associated with the known Au deposits within the study area, regardless of which clustering algorithm is applied to the output from the LT or the ALT.

Most of the known Au deposits are located within the anomaly clusters shown in Figs. 5(a) and 5(b) or within cluster 1 in Figs. 5(c), 5(e), 5(d), and 5(f). The results of applying Q-type cluster analysis to the output from the CLT and the ILT correlate well with each other, regardless of which clustering algorithm is used (Fig. 5). In particular, similar results (Figs. 5(c), 5(e), 5(d), and 5(f)) are produced when the k-means and fuzzy c-means algorithms are applied to the output of the CLT and the ILT.

The observations of the geochemical data are grouped into three clusters when the k-means and fuzzy c-means algorithms are applied to the geochemical data, and this number of clusters may be subjectively chosen. To improve the reliability of the results of Q-type cluster analysis, the number of clusters employed in Q-type cluster analysis must be discussed further in the following section.

Considering the length restrictions of this paper, the present study only compares the results of applying the k-means and fuzzy c-means algorithms to the output from NT and the ILT, with the goal of carrying out a further test of the k-means and fuzzy c-means algorithms. The observations are classified into four, six, or eight clusters, as shown in Figs. 6 and 7, respectively. The results of applying Q-type cluster analysis differ from each other when the k-means algorithm is applied to the output of NT (Figs. 6(a), 6(c) and 6(e)). Similar results are also shown in Figs. 6(b), 6(d), and 6(f) when the k-means algorithm is applied to the output from the ILT. However, one or two cluster(s) (the cluster(s) are outlined in solid black and/or white lines) are present in Fig. 6(b), Fig. 6(d), and Fig. 6 (f)). The spatial pattern(s) of these clusters are similar, and most of the known Au deposits are located within these cluster(s).

A cluster (shown within the solid white line) is always observed when Q-type cluster analysis with the fuzzy c-means algorithm is applied to the output from NT and the ILT, and this cluster is associated with the known Au deposits to some extent. Additionally, one or two cluster(s) (shown within the solid black and/or white lines) is/are always present in Figs. 7(b), 7(d) and 7(f), and their spatial patterns are similar, and the known Au deposits are associated with the cluster(s).

The anomaly cluster(s) (shown within the solid black and/or white lines) in Figs. 6(b), 6(d), 6(f), 7(b), 7(d), and 7(f) are associated with the known Au deposits, and the distributions of the cluster(s) are more or less similar (they are mainly located in the north-central and southwestern portions of the study area), regardless of whether the k-means or fuzzy c-means algorithms are applied to the output from the ILT or whether the observations are divided into four, six, or eight clusters.

5 Discussion

Some interesting results have been produced from the geochemical data using cluster analysis algorithms accompanied by compositional data transformation techniques, and the results are valuable for further mineral exploration.

The present study first classifies the 39 elements or variables of the geochemical data into several groups using the hierarchical algorithm, and one of these groups represents the hydrothermal and epithermal activity that has occurred within the study area. The study reveals that the results of applying R-type cluster analysis to the output from NT and the ZST are the same. On the other hand, the results obtained with the output from LT, the ALT, and the CLT differ significantly from each other and are different from the results obtained using the output of NT and the ZST. These results suggest that the results of R-type cluster analysis are affected by LT, the ALT, and the CLT, rather than by the ZST. Although ALT, the CLT, and the ILT are considered to be data transformation techniques that are effective in opening compositional data (Egozcue et al., 2003; Templ et al., 2008), their effects on R-type cluster analysis are not clear because the results of applying R-type cluster analysis to the output from the ALT and the CLT differ from each other and are difficult to interpret. In general, one of the variables of geochemical data must be selected as a ratio variable to open the compositional data when the ALT method is used, and the selected variable then cannot be used in further analyses. The CLT method results in collinear data (Templ et al., 2008) when it is applied to compositional data. Although the ILT method avoids collinearity (Egozcue et al., 2003), it is not a reliable choice for R-type cluster analysis, as it can erase the direct relationships among the original variables, and the results will become fuzzy. Selecting an appropriate data transformation method before applying R-type cluster analysis is a challenging task. Fortunately, some knowledge and previous studies in the present study area are helpful for the selection of the most suitable method (e.g., expert knowledge of the target deposit and the geological background of the study area can help determine which elements are the elements of interest that should be targeted).

Second, the number of clusters should be pre-defined when the hierarchical algorithm is used to perform Q-type cluster analysis; however, this assignment is difficult. In this study, the inconsistent coefficient is used as a criterion to produce the different results from the hierarchical algorithm. If the inconsistent coefficient is too small, then the number of clusters will be too large to explain the meaning of each cluster. Fortunately, the results of

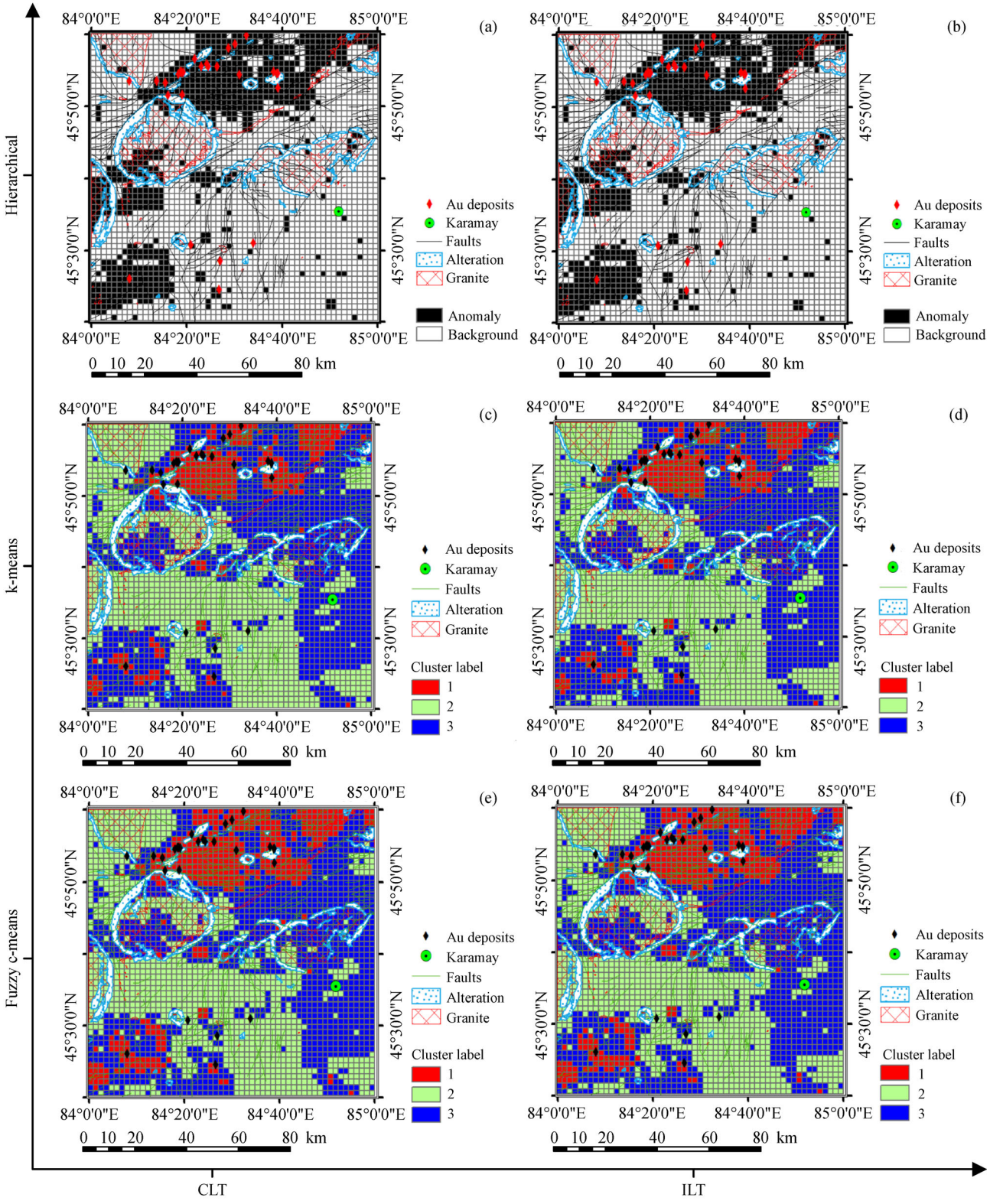

Fig. 5 The results of applying Q-type cluster analysis with the hierarchical ((a) and (b)), k-means ((c) and (d)) and fuzzy c-means ((e) and (f)) algorithms to the output from the CLT ((a), (c) and (e)) and the ILT ((b), (d) and (f)).

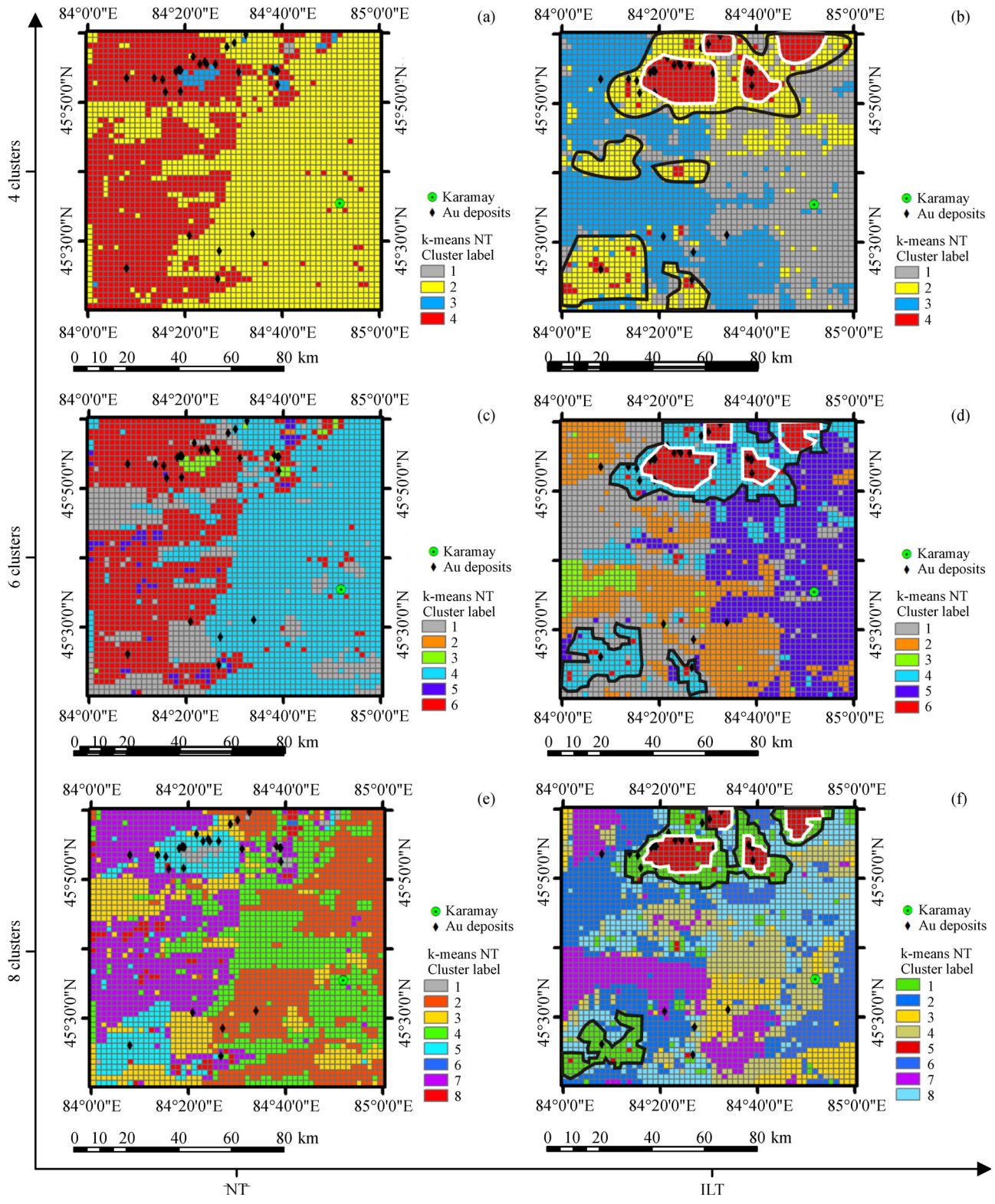

Fig. 6 The results of applying Q-type cluster analysis to the output from NT ((a), (c), and (e)) and the ILT ((b), (d), and (f)) when the k-means algorithm is used. The observations of the geochemical data are divided into 4 clusters ((a) and (b)), 6 clusters ((c) and (d)) and 8 clusters ((e) and (f)).

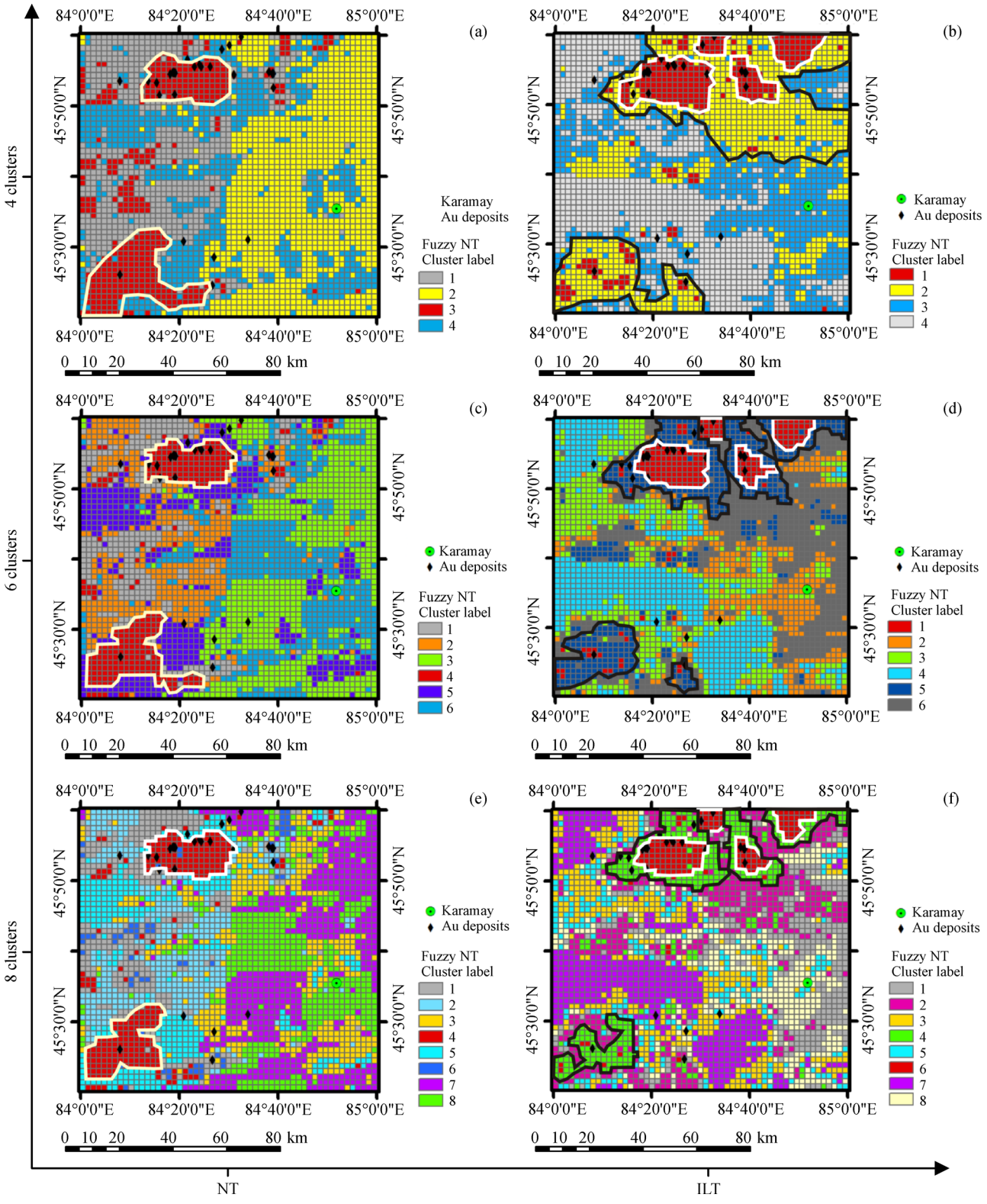

Fig. 7 The results of applying Q-type cluster analysis to the output from NT data ((a), (c), and (e)) and the ILT ((b), (d), and (f)) when the fuzzy c-means algorithm is used. The observations of geochemical data are clustered into 4 clusters ((a) and (b)), 6 clusters ((c) and (d)) and 8 clusters ((e) and (f)).

applying Q-type cluster analysis tend to stabilize when the inconsistent coefficient changes with an appropriate step from a small number to a larger one. Therefore, the relatively stable results can represent an appropriate solution. All of the clusters that contain a small number of observations are reclassified into a new cluster, which is defined as the anomaly cluster. This solution is reasonable because the process of mineralization is an event with small probability, and the geochemical anomaly should have a lower proportion. The clusters which contain the largest number of observations can be regarded as representing the background. However, the solution above is subjective and can only be used as an exploratory data analysis tool. The k-means and fuzzy c-means algorithms are more user-friendly for Q-type cluster analysis, given that they can easily produce the pre-defined number of clusters. Furthermore, there is/are always a/some cluster(s) that can be considered to be anomaly cluster(s), and the anomaly cluster(s) is/are closely associated with the known Au deposits within the study area.

The results of applying Q-type cluster analysis to the output from NT and ZST are unstable, regardless of what clustering algorithm is used. Hence, it is difficult to interpret the results. The known Au deposits are not associated with any of the clusters obtained from NT and ZST data (with the exception of Fig. 3(e)), and the reason is not clear. However, the known Au deposits are in good correlation with one of the clusters identified by applying Q-type cluster analysis to the output from the LT and the ALT. The most reliable and interpretable results are produced using the output from the CLT and the ILT when they are compared with NT, the ZST, the LT, and the ALT data. Moreover, the known Au deposits are associated with the anomaly cluster(s) of CLT and the ILT data. Furthermore, the spatial distributions of the cluster(s) are very similar when the same clustering algorithm is applied to the output from the CLT and the ILT. To clarify the reason why the results of applying cluster analysis to the output from the CLT and the ILT are more stable and interpretable and can provide more valuable information for mineral exploration than other data transformation methods, we compare the structure of the data after different data transformation methods are used. The quantile-quantile plot of the variables in each transformed data set show that the LT, the ALT, the CLT, and the ILT improve the structure of geochemical data (i.e., most of the transformed variables resemble normal distributions more closely); however, the reason why the CLT and the ILT are more suitable for use in the cluster analysis of geochemical data than the LT and the ALT is not very clear, and this point needs to be discussed in future studies.

The known Au deposits are always associated with one (contained within the solid white lines shown in Figs. 6 and 7) or two (contained within the solid white and black lines shown in Figs. 6 and 7) clusters when the observations of the output of the ILT are classified into four, six, or eight clusters, regardless of whether the k-means algorithm or the fuzzy c-means algorithm is used. Also, the solid white lines are nested within the solid black lines, and they display a halo feature together (Figs. 6 and 7). This pattern may represent the geochemical dispersion halo. Thus, the cluster(s) within the solid white and black lines shown in Figs. 6 and 7 is/are considered as strong and moderate anomalies, respectively.

6 Conclusions

1) The result of R-type cluster analysis is not affected by the use of the ZST; however, it is evidently affected by use of the LT, the ALT, and the CLT.

2) The k-means and fuzzy c-means algorithms are more user-friendly for Q-type cluster analysis of geochemical data than the hierarchical algorithm, but they are not suitable for application to the output from the NT and the ZST.

3) The results of applying Q-type cluster analysis to the output from the CLT and the ILT are very similar, regardless of whether the k-means or the fuzzy c-means algorithm is used. This result suggests that the use of the CLT and the ILT can lead to more stable results than the ALT.

4) The use of different distance metrics with the same clustering algorithm can produce different Q-type cluster analysis results (not presented in this study). This statement is especially true for applying Q-type cluster analysis to the output from NT, the ZST, the LT, and the ALT. However, the different distance metrics do not strongly affect the results of applying Q-type cluster analysis to the output from the CLT and the ILT.

5) The hierarchical algorithm is not recommended for use in Q-type cluster analysis because it is subjective, and it is difficult to determine the number of clusters in advance. NT and the ZST are also not recommended for use before performing Q-type cluster analysis of geochemical data.

6) In combination with the k-means or fuzzy c-means algorithm, the CLT or the ILT yields more reliable and interpretable results when Q-type cluster analysis is applied to geochemical data. According to the results of applying Q-type cluster analysis to the output from the CLT and the ILT, the northeastern and southwestern parts (i.e., the areas within the solid white and black lines) of the study area are promising areas for further geological exploration.

Acknowledgements The authors thank Ratheesh Kumar R.T, Rustam Orozbaev for their assistance to revise the language before we submit the manuscript and the authors are grateful for the anonymous reviewers'

constructive comments and suggestions. This study was funded by the National Natural Science Foundation of China (Grant Nos. U1503291 and 41402296), and a Major Project in Xinjiang Uygur Autonomous Region (201330121-3).

References

Abdel-Halim R E, Abdel-Aal R E (1998). Classification of urinary stones by cluster analysis of ionic composition data. Comput Methods Programs Biomed, 58(1): 69–81

Afzal P, Khakzad A, Moarefvand P, Omran N R, Esfandiari B, Alghalandis Y F (2010). Geochemical anomaly separation by multifractal modeling in Kahang (Gor Gor) porphyry system, Central Iran. J Geochem Explor, 104(1–2): 34–46

Agharezaei M, Hezarkhani A (2016). Delineation of geochemical anomalies based on Cu by the boxplot as an exploratory data analysis (EDA) method and concentration-volume (C-V) fractal modeling in Mesgaran mining area, Eastern Iran. Open Journal of Geology, 6(10): 1269–1278

Agterberg F P (2012). Multifractals and geostatistics. J Geochem Explor, 122: 113–122

Aitchison J (1982). The statistical analysis of compositional data. J R Stat Soc B, 44(2): 139–177

Aitchison J (1999). Logratios and natural laws in compositional data analysis. Math Geol, 31(5): 563–580

Aitchison J, Barcelo-Vidal C, Martin-Fernandez J A, Pawlowsky-Glahn V (2000). Logratio analysis and compositional distance. Math Geol, 32(3): 271–275

Aitchison J, Egozcue J J (2005). Compositional data analysis: where are we and where should we be heading? Math Geol, 37(7): 829–850

Bölviken B, Stokke P R, Feder J, Jossang T (1992). The fractal nature of geochemical landscapes. J Geochem Explor, 43(2): 91–109

Bounessah M, Atkin B P (2003). An application of exploratory data analysis (EDA) as a robust non-parametric technique for geochernical mapping in a semi-arid climate. Appl Geochem, 18(8): 1185–1195

Buccianti A (2013). Is compositional data analysis a way to see beyond the illusion? Comput Geosci, 50: 165–173

Carranza E J M (2009). Geochemical anomaly and mineral prospectivity mapping in GIS. Handbook of exploration and environmental geochemistry, 11. Elsevier Science

Carranza E J M (2010). Catchment basin modelling of stream sediment anomalies revisited: incorporation of EDA and fractal analysis. Geochem Explor Environ Anal, 10(4): 365–381

Carranza E J M (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. J Geochem Explor, 110(2): 167–185

Carranza E J M, Hale M (1997). A catchment basin approach to the analysis of reconnaissance geochemical-geological data from Albay Province, Philippines. J Geochem Explor, 60(2): 157–171

Cheng Q, Agterberg F P (2009). Singularity analysis of ore-mineral and toxic trace elements in stream sediments. Comput Geosci, 35(2): 234–244

Cheng Q, Agterberg F P, Bonham-Carter G F (1996). A spatial analysis method for geochemical anomaly separation. J Geochem Explor, 56 (3): 183–195

Davis J C (2002). Statistics and Data Analysis in Geology (3rd ed). New York, Chichester, Brisbane, Toronto, Singapore: John Wiley and Sons

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003). Isometric logratio transformations for compositional data analysis. Math Geol, 35(3): 279–300

Eilermann M, Post C, Schwarz D, Leufke S, Schembecker G, Bramsiepe C (2017). Generation of an equipment module database for heat exchangers by cluster analysis of industrial applications. Chem Eng Sci, 167: 278–287

Fatehi M, Asadi H H (2017). Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data, a case study from the Dalli Cu-Au porphyry deposit in central Iran. Ore Geol Rev, 81: 245–255

Fattahi H (2016). Indirect estimation of deformation modulus of an in situ rock mass: an ANFIS model based on grid partitioning, fuzzy c-means clustering and subtractive clustering. Geosci J, 20(5): 681–690

Filzmoser P, Hron K, Reimann C (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ, 407(23): 6100–6108

Filzmoser P, Hron K, Reimann C (2010). The bivariate statistical analysis of environmental (compositional) data. Sci Total Environ, 408(19): 4230–4238

Ghosh T, Kanchan R (2014). Geoenvironmental appraisal of groundwater quality in Bengal alluvial tract, India: a geochemical and statistical approach. Environ Earth Sci, 72(7): 2475–2488

Han J, Kamber M (2006). Data Minning: Concepts and Techniques (2nd ed). Beijing: China Machine Press

Hassanpour S, Afzal P (2013). Application of concentration–number (C–N) multifractal modeling for geochemical anomaly separation in Haftcheshmeh porphyry system, NW Iran. Arab J Geosci, 6(3): 957–970

Hawkes H E, Webb J S (1962). Geochemistry in Mineral Exploration. New York: Harper

He G Q, Chen S D, Xu X, Li J Y, Hao J (2004). An Introduction to the Explanatory Text of the Map of Tectonics of Xinjiang and Its Neighbouring Area (1:250000). Beijing: Geological Publishing House (in Chinese)

Howarth R J (1983). Statistics and Data Analysis in Geochemical Prospecting. Handbook of Exploration Geochemistry, 2. Amsterdam-Oxford-New York Elsevier

Kim T, Moon D C, Park W B, Park K H, Ko G W (2007). Classification of springs of Jeju Island using cluster analysis of annual fluctuations in discharge variables: investigation of the regional groundwater system. Geosci J, 11(4): 397–413

Kitzig M C, Kepic A, Kieu D T (2017). Testing cluster analysis on combined petrophysical and geochemical data for rock mass classification. Explor Geophys, 48(3): 344–352

Lee J Y, Song S H (2007). Groundwater chemistry and ionic ratios in a western coastal aquifer of Buan, Korea: implication for seawater intrusion. Geosci J, 11(3): 259–270

Leite M L C (2016). Applying compositional data methodology to nutritional epidemiology. Stat Methods Med Res, 25(6): 3057–3065

Meng H, Song Y, Song F, Shen H (2011). Research and application of cluster and association analysis in geochemical data processing.

Computat Geosci, 15(1): 87–98

Sahraei Parizi H, Samani N (2013). Geochemical evolution and quality assessment of water resources in the Sarcheshmeh copper mine area (Iran) using multivariate statistical techniques. Environ Earth Sci, 69 (5): 1699–1718

Parsa M, Maghsoudi A, Yousefi M, Carranza E J M (2017). Multifractal interpolation and spectrum–area fractal modeling of stream sediment geochemical data: implications for mapping exploration targets. J Afr Earth Sci, 128: 5–15

Pazand K, Hezarkhani A, Ataei M, Ghanbari Y (2011). Application of multifractal modeling technique in systematic geochemical stream sediment survey to identify copper anomalies: a case study from Ahar, Azarbaijan, Northwest Iran. Chemie der Erde- Geochemistry, 71(4): 397–402

Reimann C, Filzmoser P (2000). Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39(9): 1001–1014

Reimann C, Filzmoser P, Garrett R G (2002). Factor analysis applied to regional geochemical data: problems and possibilities. Appl Geochem, 17(3): 185–206

Reimann C, Filzmoser P, Garrett R G (2005). Background and threshold: critical comparison of methods of determination. Sci Total Environ, 346(1–3): 1–16

Reimann C, Garrett R G (2005). Geochemical background- concept and reality. Sci Total Environ, 350(1–3): 12–27

Rock N M S (1988). Numerical Geology. Lecture Notes in Earth Sciences, 18. New York, Berlin, Heidelberg: Springer-Verlag

Stück H, Koch R, Siegesmund S (2013). Petrographical and petrophysical properties of sandstones: statistical analysis as an approach to predict material behaviour and construction suitability. Environ Earth Sci, 69(4): 1299–1332

Su Y, Tang H, Hou G, Liu C (2006). Geochemistry of aluminous A-type granites along Darabut teconic belt in west Junggar, Xinjiang. Geochimica, 35(1): 55–67 (in Chinese)

Templ M, Filzmoser P, Reimann C (2008). Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem, 23(8): 2198–2213

Templ M, Hron K, Filzmoser P (2017). Exploratory tools for outlier detection in compositional data with structural zeros. J Appl Stat, 44 (4): 734–752

Tolosana-Delgado R, McKinley J (2016). Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). Appl Geochem, 75: 263–276

Tukey J W (1977). Exploratory Data Analysis. Reading: Addison-Wesley

Wang L, Wang Y, Zhang W, Xu C, An Z (2014). Multivariate statistical techniques for evaluating and identifying the environmental significance of heavy metal contamination in sediments of the Yangtze River, China. Environ Earth Sci, 71(3): 1183–1193

Wang X Q, Xie X J, Zhang B R, Hou Q Y (2011). Geochemical probe into China's continental crust. Acta Geoscientica Sinica, 32: 65–83 (in Chinese)

Xie X, Mu X, Ren T (1997). Geochemical mapping in China. J Geochem Explor, 60(1): 99–113

Xie X, Wang X, Zhang Q, Zhou G, Cheng H, Liu D, Cheng Z, Xu S (2008). Multi-scale geochemical mapping in China. Geochem Explor Environ Anal, 8(3–4): 333–341

Yusta I, Velasco F, Herrero J M (1998). Anomaly threshold estimation and data normalization using EDA statistics: application to lithogeochemical exploration in lower Cretaceous Zn-Pb carbonate-hosted deposits, northern Spain. Appl Geochem, 13(4): 421–439

Zhang C, Huang X (1992). The ages and tectonic settings of ophiolites in West Junggar, Xinjiang. Geological Review, 38(6): 509–524 (in Chinese)

Zhang F (2003). The study of geological characteristics of the gold associated minerals and gold vine of Hatu gold deposit. Journal of Xinjiang Nonferrous Metals, 26(3): 5–6 (in Chinese)

Zhu Y, An F, Xu C, Guo H, Xia F, Xiao F, Zhang F, Lin C, Qiu T, Wei S (2013a). Geology and Au-Cu Deposits in the Hatu and its Adjacent Region (Xinjiang): Evolution and Prospecting Model. Beijing: Geological Publishing House

Zhu Y, Chen B, Xu X, Qiu T, An F (2013b). A new geological map of the western Junggar, north Xinjiang (NW China): implications for Paleoenvironmental reconstruction. Episodes, 36(3): 205–220

Zumlot T, Goodell P, Howari F (2009). Geochemical mapping of New Mexico, USA, using stream sediment data. Environmental Geology, 58(7): 1479–1497

Zuo R (2012). Exploring the effects of cell size in geochemical mapping. J Geochem Explor, 112: 357–367

Zuo R, Cheng Q (2008). Mapping singularities- a technique to identify potential Cu mineral deposits using sediment geochemical data, an example for Tibet, west China. Mineral Mag, 72(1): 531–534

Zuo R, Wang J, Chen G, Yang M (2015). Identification of weak anomalies: a multifractal perspective. J Geochem Explor, 148: 12–24

Zuo R, Xia Q, Wang H (2013a). Compositional data analysis in the study of integrated geochemical anomalies associated with mineralization. Appl Geochem, 28: 202–211

Zuo R, Xia Q, Zhang D (2013b). A comparison study of the C-A and S-A models with singularity analysis to identify geochemical anomalies in covered areas. Appl Geochem, 33: 165–172