Computers and Geosciences 123 (2019) 10–19

Contents lists available at ScienceDirect

Computers and Geosciences

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

A novel hierarchical clustering analysis method based on Kullback–Leibler divergence and application on dalaimiao geochemical exploration data

Jie Yang[a,*], Eric Grunsky[b,c], Qiuming Cheng[b]

[a] Institute of Geosciences, China University of Geosciences (Beijing), Beijing, 100083, China
[b] State Key Lab of Geological Processes and Mineral Resources, China University of Geosciences (Beijing), Beijing, 100083, China
[c] Department of Earth and Environmental Sciences, University of Waterloo, N2L3G1, Canada

ARTICLE INFO

Keywords:
Kullback–Leibler divergence
Hierarchical cluster analysis
Geochemical exploration data
Geochemical pattern
Data mining

ABSTRACT

In this paper, we propose a new hierarchical clustering analysis method (HCA) that uses Kullback–Leibler divergence (D_{KLS}) of pairwise geochemical datasets of geo-objects (e.g., lithological units) as a measure of proximity. The method can reveal relationships among geo-objects based on geochemistry. This capability is verified through an application with geochemical exploration data from regolith that overlies the Dalaimiao region in China. D_{KLSM} and D_{KLSC}, two parts of D_{KLS}, respectively describe the differences on the mean and the differences on covariance and are also used as measures of proximity. D_{KLSM} characterizes rock type and D_{KLSC} describes spatial relationships and component similarities between geo-objects. This contribution not only provides a tool that can reveal relationships between geo-objects based on geochemical data, but also reveals that D_{KLS} and its two parts can characterize geochemical differences from different perspectives. These measures hold promise in the enhancement of methods for recognizing geochemical patterns.

1. Introduction

Hierarchical clustering analysis (HCA) is a method that builds a hierarchy of clusters of variables (R-mode) or observations (Q-mode) according to the proximity between pairwise variables or observations. This method is commonly used in geochemical data processing such as environmental assessment and mineralization exploration (Grunsky, 2010; Hernandez et al., 2004; Li et al., 2014; Mokhtari et al., 2014; Nezhad et al., 2015; O'Shea and Jankowski, 2006; Templ et al., 2008). However, in the instance of geochemical data most cases of the application of Q-mode HCA focus on the classifications of individual specimens but not on datasets or groups of samples. That is because measures of proximity such as the Manhattan Distance (Mumm et al., 2012; O'Shea and Jankowski, 2006), D-value (Kremer et al., 2012), and Euclidian distance (Fatehi and Asadi, 2017) are based on pairwise comparisons of specimens (data points). In regional geochemical data, many sites are sampled over a common geo-object (e.g., lithology unit, alteration zone, structural belt and other objects that occupy geographic space). When we focus on the relationships between those geo-objects, the HCA, based on the proximity between pairwise geochemical data points, may not perform well, as it can result in a large and complex dendrogram with many leaves, which is complicated and difficult to explain.

Based on the extent of the geo-objects the entire dataset can be divided into several sub-datasets, each containing sites collected over the same geo-objects, which characterize the geo-objects more precisely than at a single data point. Pairwise differences between sub-datasets as the measure of proximity in HCA, make it possible to design a new HCA algorithm that reveals relationships among the geo-objects. In this paper, the Kullback–Leibler divergence (KL-divergence) as a measure of proximity, is used to develop the HCA method, which is then applied to a geochemical data in the Dalaimiao district.

2. KL-divergence based HCA

2.1. Measure of proximity

HCA builds models based on proximity. For Q-type clustering, its proximities represent distances or dissimilarities between observations. When observations are datasets (groups or populations), it is necessary to measure the distance or dissimilarity between groups. For example, we can use a measure of dissimilarity (or metric) such as the Euclidian or Aitchison distances (Aitchison et al., 2000) between the centroids of the groups as a proximity measure for HCA, but it will lose the information on the "shape" of the groups in the variable space (feature space). The measure of dissimilarity considers the "centroid" and

* Corresponding author. Institute of Geosciences, China University of Geosciences, No. 29, Xueyuan Road, Haidian District, Beijing, China, 100083.
E-mail address: jieyang@cugb.edu.cn (J. Yang).

https://doi.org/10.1016/j.cageo.2018.11.003
Received 30 January 2018; Received in revised form 29 October 2018; Accepted 8 November 2018
Available online 11 November 2018
0098-3004/

Table 1
Some commonly used dissimilarities between two probability distributions P and Q.

Dissimilarity	Definition	The dissimilarity between two multivariate normal distributions	Symmetry	Triangle Inequality	Scale Invariance	Separable
Wasserstein Metric	$W_l(P, Q) = \int_0^1 \lvert p^{-1}(x) - q^{-1}(x)\rvert^l dx,\ l \geq 1$	$W_2^2(N_0, N_1) = \lVert u_1 - u_2 \rVert_2^2 + tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$	True	True	False	True
Kullback–Leibler Divergence	$D_{KL}\left(P \lVert Q\right) = \int p(x) log\frac{p(x)}{q(x)} dx$	$D_{KL}(N_0 \lVert N_1) = \left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - k + \ln\left(\frac{det\Sigma_1}{det\Sigma_0}\right)\right)/2$	False	False	True	True
Bhattacharyya distance	$D_B(P, Q) = -\ln(\int \sqrt{p(x)q(x)}\, dx)$	$D_B(N_0, N_1) = \frac{1}{8}(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0) + \frac{1}{2}\ln\left(\frac{\det(\Sigma)}{\sqrt{\det(\Sigma_1)\det(\Sigma_0)}}\right)$, with $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$	True	False	True	True
Hellinger distance	$D_H^2(P, Q) = 1 - \sqrt{p(x)q(x)}dx$	$D_H^2(N_0, N_1) = 1 - \frac{\det(\Sigma_0)^{\frac{1}{4}}\det(\Sigma_1)^{\frac{1}{4}}}{\det(\Sigma)^{\frac{1}{2}}}\exp\left(-\frac{1}{8}(\mu_1 - \mu_0)^T\Sigma(\mu_1 - \mu_0)\right)$, with $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$	True	True	True	False

"shape" of the data, which is more informative.

There are many measures of dissimilarity such as the Kullback–Leibler divergence, Bhattacharyya distance, Hellinger distance and Wasserstein metric, which can characterize the dissimilarity (difference) between two probability distributions (the term "distance" does not mean that the measure is a metric in the strict sense). For distributions P and Q of continuous random variables, those measures and their properties are given in Table 1. Where $p(x)$ and $q(x)$ are the probability density functions (PDF) of P and Q, k is the number of variables, tr(Φ)and det(Φ) are the trace and the determinant of matrix Φ, $N_0(\mu_0, \Sigma_0)$ and $N_1(\mu_1, \Sigma_1)$ are two multivariate normal distributions, μ_0 and μ_1 are the corresponding mean vectors, Σ_0 and Σ_1 are the corresponding covariance matrices.

To get the values of dissimilarities between two datasets, a kernel density estimation can be used to obtain the PDFs of the two datasets and then calculate the dissimilarities according to the formulae in Table 1. However, based on knowledge about the distribution of geochemical data, those dissimilarities can be estimated more simply. Previously, many distributions were used to model element concentration data including the normal, log-normal, and power-law distributions, but none of them fit perfectly in practical data applications (Grunsky, 2010; Reimann and Filzmoser, 2000). Typically, a dataset consists of multiple populations from different sources and different processes (Reimann et al., 2002; Zhang et al., 2005, 2008). Therefore, hybrid models of the aforementioned distributions are proposed, and after some preprocessing steps including log transformation, log-ratio transformation (Pawlowsky-Glahn and Buccianti, 2011) and local singularity analysis (Cheng, 2007; Liu et al., 2014). Those models can approximate normal distributions or mixtures of normal distributions. For example, because power-law distribution can be explained as a mixture of lognormal distributions based on a geometric distribution (Allen et al., 2001; Mitzenmacher, 2004), a hybrid model of a log-normal body with a power-law tail proposed by Cheng and Agterberg (2009), followed by a log-transform the distribution will result in a mixture of normal distributions. For geochemical data, the simplest and most conservative assumption is that after proper pre-processing steps, groups of concentration values from the geo-objects approximate a family of normal distributions. Therefore, we can use the divergence between two multivariate normal distributions to roughly measure the dissimilarity between two datasets. The dissimilarities of measures for two multivariate normal distributions $N_0(\Sigma_0, \mu_0)$ and $N_1(\Sigma_1, \mu_1)$ are given in Table 1.

From Table 1 it can be observed that the Wasserstein Metric, KL-divergence, and Bhattacharyya distance between to multivariate normal distributions can be further divided into two parts, one describes the differences between means, and the other measures the differences between the corresponding covariances. For example, KL-divergence can be separated into $(tr(\Sigma_1^{-1}\Sigma_0) - k + \ln(det\Sigma_1) - \ln(det\Sigma_2))/2$ and $(\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0)/2$ (Kullback, 1978). In this way, the data can be observed from two perspectives. Besides, KL-divergence, the Bhattacharyya and Hellinger distances belongs to the family of f-Divergences that have the property of scale invariance (Basseville, 2013). This is an important property for geochemical data processing, because element concentrations (transformed or not) might be at a different scale, and this property removes the necessity of standardizing the data.

Table 1 shows that the Hellinger distance and the Wasserstein metric follow the triangle inequality (Clement and Desch, 2008; Steerneman, 1983), but the KL-divergence and Bhattacharyya distance violate this property (Kailath, 1967; Kullback, 1978). However, the triangle inequality is not a necessary property for the dissimilarities used in HCA (Jain et al., 1999). For example, HCA using a single, compete, average or weighted average linkage does not require a metric (Everitt et al., 2011). Therefore, although some measures in Table 1 are not metric, they can be still used in HCA with limited linkage methods.

Among the measures listed in Table 1, the KL-divergence is a special one, because it belongs to many divergence families including f-,

Alpha-, Beta-, Gamma- and Bregman-divergences (Amari and Cichocki, 2010; Cichocki and Amari, 2010), and the KL-divergence benefits from the properties of those families such as affine invariance (Kanamori and Fujisawa, 2014). This means after a linear transform such as principal component analysis, the KL-divergence between two groups does not change. Moreover, the KL-divergence is one of the most widely known and used measures of dissimilarity in the field of machine learning. Therefore, it is reasonable to start with KL-divergence with a test using HCA and then extend to measures of other families.

The measure of proximity used in cluster analysis should be symmetrical. However, the KL-divergence is not. We can take the symmetric form of KL-divergence, $D_{KLS}(N_0, N_1) = D_{KL}(N_0||N_1) + D_{KL}(N_1||N_0)$, which also is called J-divergence (Jeffreys, 1946), that is:

$$D_{KLS}(N_0, N_1) = \frac{1}{2}\left(\left(\mu_1 - \mu_0\right)^T (\Sigma_1^{-1} + \Sigma_0^{-1})(\mu_1 - \mu_0) + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + tr(\Sigma_0^{-1}\Sigma_1) - 2k\right) \quad [1]$$

According to the separation of KL-divergence, D_{KLS} also can be divided into two parts:

$$D_{KLSM}(N_0, N_1) = \frac{1}{2}\left(\left(\mu_1 - \mu_0\right)^T \left(\Sigma_1^{-1} + \Sigma_0^{-1}\right)(\mu_1 - \mu_0)\right) \quad [2]$$

$$D_{KLSC}(N_0, N_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + tr(\Sigma_0^{-1}\Sigma_1) - 2k) \quad [3]$$

where D_{KLSM} measures the differences between means in terms of Mahalanobis distance, and D_{KLSC} characterizes the differences between the corresponding covariances. Because the measure D_{KLS} is not a metric, D_{KLS} and its two parts (D_{KLSM} and D_{KLSC}) will be called "dissimilarity" in the remainder of the manuscript. It must be emphasized again that the equations of symmetric KL divergence and its parts are based on an assumption that each dataset approximate a multivariate normal distribution. It is also possible to assume that the data follow two skew normal distributions to match the practice data more accurately, but it results in a significant cost of complexity (Contreras-Reyes and Arellano-Valle, 2012).

2.2. Hierarchical clustering

Strategies for hierarchical clustering methods are typicalyl agglomerative or divisive. The agglomerative method is the most popular and is used here. In the agglomerative strategy, each observation (or variable) starts in its own cluster (leaf node), and pairs of clusters that have the smallest distance are merged progressively in a hierarchical process. For example in a forest of observations (forest) or variables, when two clusters, s and t, are combined into a single cluster u, s and t are removed from the forest, and u is added to the forest. When only one cluster remains in the forest, the agglomerative algorithm stops, and this cluster becomes the root. A wide range of agglomerative hierarchical clustering algorithms have been proposed, and they generally fall into three categories: the linkage method, the center specified method, and others (Murtagh and Contreras, 2012). As proposed in the previous section, only linkage methods are available for the HCA that uses D_{KLS}, D_{KLSM} and D_{KLSC} as the measure of dissimilarity. In the linkage method, at each iteration, the dissimilarity (or distance) matrix is updated according to a linkage criterion to reflect the dissimilarity between a newly formed cluster u with the remaining clusters in the forest of observations (Murtagh, 1983). Four available linkage criteria for updating the dissimilarity $d(s, t)$ between two clusters s and t are given in Table 2. In the table, where cluster s and cluster t are combined to form cluster u, there are n original observations u_1, u_2…, u_n in cluster u, v is any remaining cluster in the forest that is not u, and there

Table 2
Available linkage criterions for KL-divergence based proximity measures, modified from Everitt et al. (2011).

Linkage	Alternative Name	Equation	Distance between clusters defined as:	Remarks
Single Linkage	Nearest Neighbor	$d(u, v) = min(d(u_i, v_j))$	The minimum distance (or dissimilarity) between observations of each cluster	Tends to produce unbalanced and straggly clusters (the 'chaining' effect), especially in large data sets.
Complete linkage	Further Neighbor	$d(u, v) = \max(d(u_i, v_j))$	The maximum distance (or dissimilarity) between observations of each cluster	Tends to find compact clusters with equal diameters. Does not take account of cluster structure.
Average Linkage	UPGMA	$d(u, v) = \sum \frac{d(u_i, v_j)}{(nm)}$	Average distance (or dissimilarity) between elements of each cluster	Tends to join clusters with small variances. Intermediate between single and complete linkage. Requires an assumption that the distances from the root to every branch tip are equal. Takes account of cluster structure. Relatively robust.
Weight Average linkage	WPGMA	$d(u, v) = \frac{d(s,v) + d(t,v)}{2}$	.Average distance (or dissimilarity) between elements of each cluster	Similar to average Linkage, but this method requires an additional assumption that branches are equal at every level of the tree

are m original objects v_1, v_2…, v_m in cluster v.

The single linkage method has the drawback of the "chaining" effect which may mask geochemical structures, and the average linkage or weighted average linkage requires additional assumptions. However, the complete linkage method overcomes these limitations and is applied in the following data processes.

In general, the proposed method has the following steps:

(1) Partition the geochemical data based on the unique classes of geo-objects. That is, extract the data points falling in the same geo-object class (e.g., the "Join" operation for point feature or the "Extract by Mask" operation for raster data in ArcGIS);
(2) Calculate the mean vectors and covariance matrices of every geo-object class sub-dataset and then according to the results, calculate the D_{KLS} (or D_{KLSC}, D_{KLSM}) matrix of sub-datasets
(3) Combine the two clusters that have the minimum dissimilarity into a new cluster.
(4) Update the dissimilarities between the new cluster and the other clusters according to the linkage method (complete linkage is used here).
(5) Return to Step 3 until all observations are assigned to a cluster.
(6) Draw dendrogram.

In this paper, the Python module NumPy (Walt et al., 2011) is used in Step 2, SciPy (Jones et al., 2001) is used to compute Steps 3–5 and Matplotlib (Hunter, 2007) is used in Step 6.

3. Geology and geochemical data

3.1. Geology

The study area of 2952 km^2 is in the Inner Mongolia Autonomous Region of China. The area overlies a Neopaleozoic accretion complex, the Uliastai active continent margin, the subduction zone of the Siberian plate and the North China platform. Most of the area is covered by a thin layer of wind-transported sand and soil. The underlying lithology is recognized by saprolite (rock debris) and a few outcrops scattered over the surface. Quaternary regolith sediments cover 41% of the area (Fig. 1). The geological structure and faults are inferred from geophysical and remote sensing data.

Most intrusions, strata, and faults have been influenced due to subduction during the Paleozoic and exhibit northwestern trending structures. The volcano-sedimentary strata are divided into seven formations: Ordovician Bayanhushu Formation of shale, siltite, sandstone and limestone mainly exposed in the middle belt of the study area. Resting unconformably on the Ordovician Bayanhushu Formation is the Devonian Niqiuhe Formation consisting of siltite, clastic rocks, limestone and tuff exposed in the southeastern corner of the study area. The Carboniferous-Permian Baoligaomiao Formation is the most widely exposed strata and has an unconformable contact with the underlying Niqiuhe Formation. The Baoligaomiao Formation consists of andesite, tuff, quartz sandstone, shale, siltite, and glutenite. The Jurassic Baiyingaolao Formation is comprised of tuff and rhyolite and lies unconformably on the Baoligaomiao Formation and is commonly exposed in the southwestern part of the area. The Cretaceous Damoguaihe Formation is composed of glutenite and mudstone and rests conformably on the Baiyingaolao Formation and occurs primarily in the southeastern part of the region. Outcrops of the Paleogene Yierdingmanha Formation comprised of mudstone and sandstone occur in the western part of the study area. A few outcrops of the Quaternary Abaga Formation comprised of basalt occur in the southern and northern parts of the area. The weathering process in the Dalaimiao district is primarily mechanical through intensive wind erosion, transportation, and deposition. As a result, the soil profile is poorly developed, and the near-surface soils are mixtures of wind-deposited sand, silt with some saprolite from nearby outcrops.

Quartz-rich felsic intrusions are exposed extensively in the Dalaimiao district. Most of these intrusions were emplaced in the Carboniferous to Permian periods, and a few are Cretaceous in age. Carboniferous-Permian magmatic activities formed batholiths or plutons, whereas the Jurassic magmatic events formed small dikes or veins. The Carboniferous-Permian intrusions consist of monzogranite, alkali feldspar granite, biotite monzogranite, biotite granite, granodiorite, quartz diorite, and diorite. The Jurassic intrusions consist of quartz porphyry, granite porphyry, granite, alkali-feldspar granite and biotite granite. Most of the Jurassic intrusions located on the boundaries of Carboniferous-Permian intrusions are more acid and finer grained. In the study area, five Mo ore deposits occur within, inside or on the boundary of Carboniferous-Permian intrusions. Four of these are associated with Jurassic magmatic activities, and the fifth is related to Carboniferous-Permian magmatic activities.

3.2. Geochemical data

Soil samples were collected from 17,610 sites in the Dalaimiao area (Fig. 2). The following elements (lower limits of detection in brackets) were determined: Ag (50 ppb), As (1 ppm), Au (1 ppb), Bi (0.3 ppm), Cu (2 ppm), Co (1 ppm), Hg (50 ppb), Mo (1 ppm), Ni (1 ppm), Pb (10 ppm), Sb (0.3 ppm), Sn (2 ppm), W (1 ppm), Zn (20 ppm). Concentration values of these elements were determined by inductively coupled plasma-mass spectrometry (ICP-MS) or atomic absorption spectroscopy (AAS) according to Chinese Geochemical Survey Specification (DZ/T 0011–91). All element concentration values of samples were reported to be higher than the lower limits of detection.

3.3. Geo-objects and their geochemical datasets

There are 20 lithology units (geo-objects) of significant areal extent, which are fully covered by geochemical samples (Fig. 3 and Table 3). However, within, or on the peripheries, of the geo-objects, there are several deposits, mineralization occurrences, alteration zones and contact zones. Because they are related to mineralization events, their element concentration values follow power-law distributions (Cheng et al., 1994). We can select areas that are further from mineralization and select elements that are not significantly associated with mineralization events to make the concentration values more log-normally distribution. Thus, 18 smaller geo-objects (selected areas in Fig. 3) were extracted from the 20 geo-objects. Each of these objects has the same lithology and contain large numbers of sample sites far away from potential mineralization. Moreover, some areas (geo-objects) are located in same lithology units, for example SOC1, SOC2, and SOC3 are three geo-objects in the Bayanhushu Formation. We can use them can to test the performance of HCA method as they should form clusters first, followed by a merge with other clusters.

A logarithmic transformation was applied to the concentration values to overcome the non-normal distributions and large positive skewness. HCA was also applied to the data after a centered log-ratio transform, it yielded proximity matrices that only have tiny differences to proximity matrices generated from log-transformed data, but the dendrograms from those two different data sources are nearly identical. The similarity of the results is because the log-ratio transformation is a linear transform on log-transformed variables and the KL divergence has the affine invariance property. The proximity matrices are not identical because a dimension is lost in the linear transform. Therefore, for the remainder of this study only a log-transform was used.

The Kolmogorov–Smirnov test was used to test the normality of individual elements on the log-transformed datasets of the 18 geo-objects. p-values are given in Table 4 p-values that higher than 0.05 are highlighted in bold, and generally means that the null hypothesis cannot be rejected and the log-transformed concentration values follow a normal distribution. The column "NO" and row "NE" shows the number of cases that the null hypothesis is rejected for the geo-objects

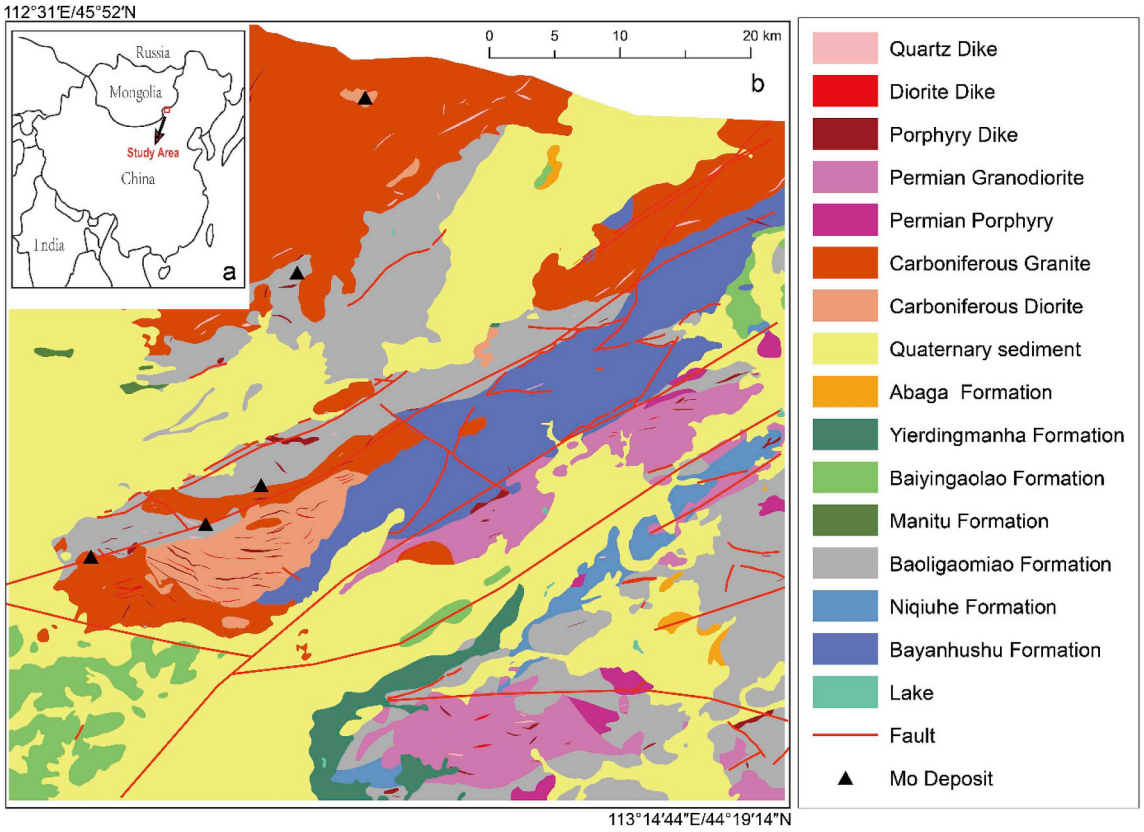

Fig. 1. Locations and Geology of Dalaimiao district. (a) The location of the study area; (b) Geological map of the study area. The map is modified from Tao (2009).

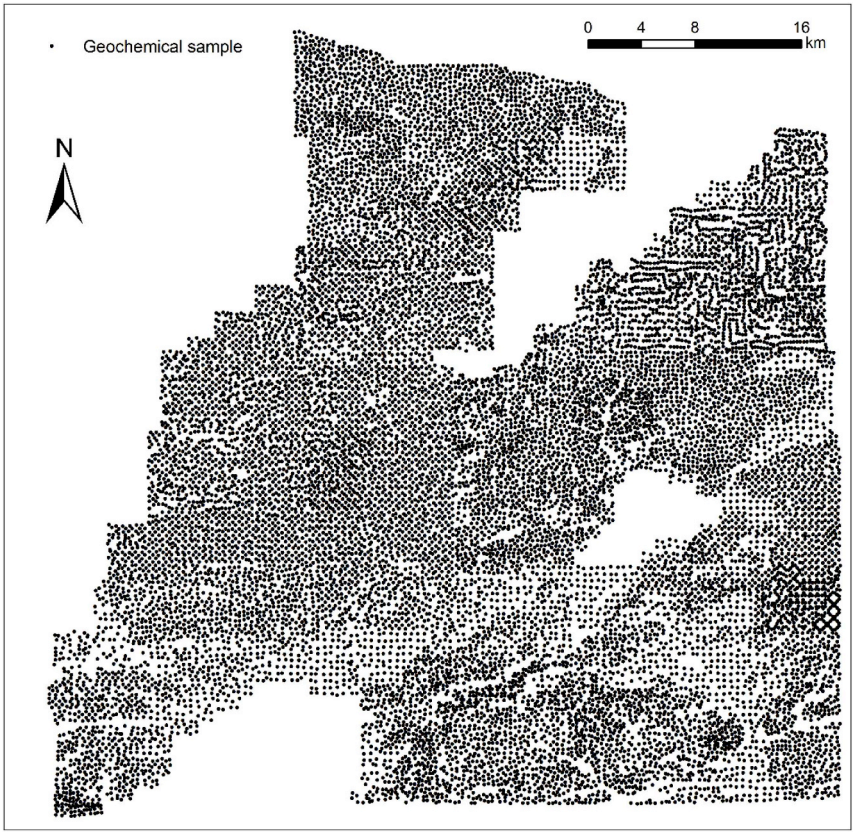

Fig. 2. Locations of soil geochemical samples (Tao, 2009). 17,610 samples collected from nine map sheets at 1:50,000 scale, the average sampling density is four samples per square kilometer.

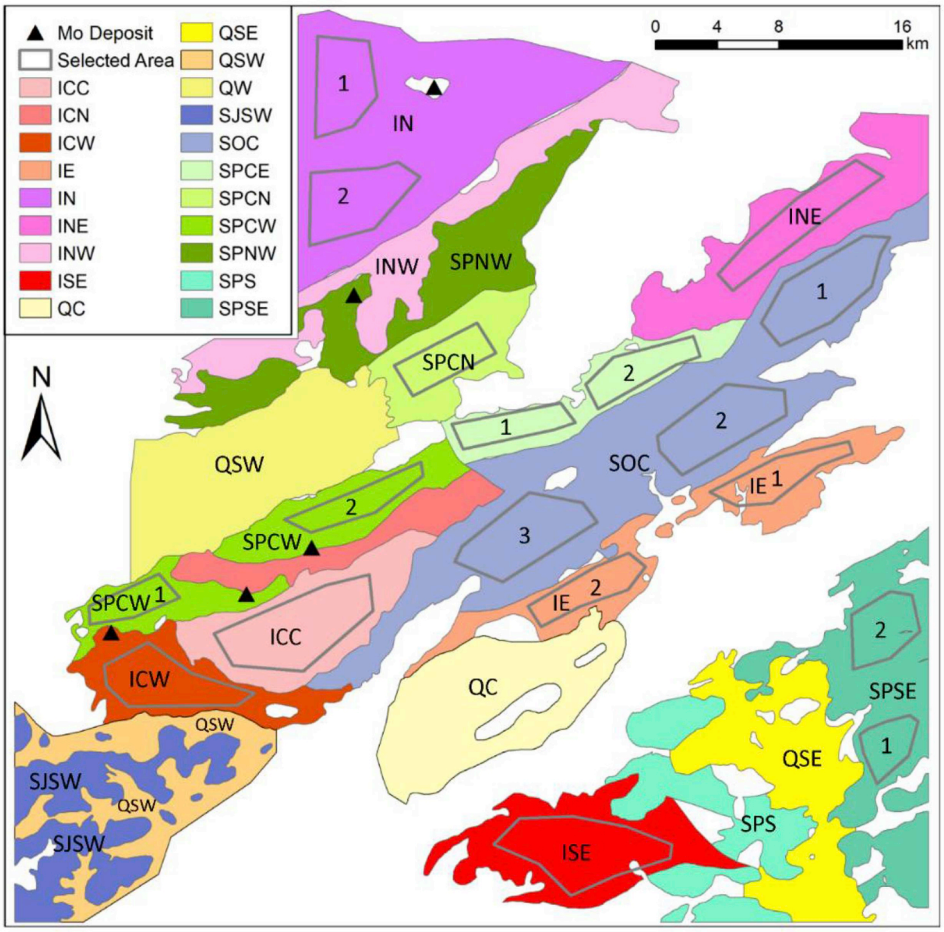

Fig. 3. Map of geo-objects. Color patterns stand for lithology units (the 20 geo-objects), symbols for mineral deposits, hollow polygons for selected areas (the 18 smaller geo-objects). Geo-objects are named in the form of code C—Y—P-X, where C represents lithology, Y represents age, P represents Position, X represents the number of geo-object: For lithology code, I = Intrusion, S = Strata, Q = Quaternary regolith; For age code, P = Permian, O = Ordovician, J = Jurassic. The ages of intrusions and regolith are Carboniferous-Permian and Quaternary, the age codes for intrusion and regolith are omitted; for position code, C = center, CW = West of Center, N = North, NE = Northeast, etc. For the number of geo-object, the code only exists in the case of eighteen-geo-objects and when there are several geo-objects in a same lithological unit. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and elements respectively. It shows that igneous rock geo-objects generally have higher "NO" values (the average value is 7.6) than strata-based geo-objects (the average value is 4.2). This implies that element concentration values in the igneous rock tend to not follow log-normal distributions. The occurrence of hydrothermal events likely causes the non-log-normality happened in the intrusions. Additionally, the row "NE" shows that Mo, W, Sb, Bi, Au, and Sn concentration values do not follow log-normal distributions (NE > 8). This is due to the fact Mo,

Table 3
Details about the geo-objects.

TWENTY GEO-OBJECTS	DESCRIPTIONS	SELECTED EIGHTEEN GEO-OBJECTS
IN	A batholith of coarse grain syenogranite at the northwest corner, with several Mo mineralization occurrences.	IN1 and IN2
INW	A northeast-trending intrusion with a Mo deposit and a belt-shape Mo—Cu geochemical anomaly on its edge.	None
ICN	An intrusion of fine grain biotite adamellite with Mo deposits on its edge	ICN
ICW	An intrusion of coarse grain biotite adamellite with Mo deposits on its edge	ICW
ICC	An intrusion of biotite granodiorite	ICC
INE	An intrusion of biotite adamellite	INE
IE	A northeast-trending intrusion of monzogranite	IE1 and IE2
ISE	A complex intrusion of syenogranite and biotite adamellite	ISE
SPNW	A geo-object of Baoligaomiao Formation I consists of quartz sandstone, shale, siltite, and glutenite, with a Mo deposit and a belt-shape Mo—Cu geochemical anomaly on its edge	None
SPCE	A geo-object of Baoligaomiao Formation I consists of quartz sandstone, shale, siltite, and glutenite	SPCE1, SPCE2
SPCN	A geo-object of Baoligaomiao Formation consists of quartz sandstone, shale, siltite, and glutenite	SPCN
SPCW	A geo-object of Baoligaomiao Formation I consists of quartz sandstone, shale, siltite, and glutenite, with two Mo deposits on its edge	SPCW1, SPCW2
SOC	A geo-object of Bayanhushu Formation consists of shale, siltite, sandstone and limestone	SOC1, SOC2, and SOC3
SPSE	A geo-object of Baoligaomiao Formation II consists of andesite and tuff	SPSE1, SPSE2
SPS	A geo-object of Baoligaomiao Formation II consists of andesite and tuff	None
SJSW	Baiyingaolao Formation consists of andesite, tuff, quartz sandstone, shale, siltite, and glutenite	None
QW	Quaternary regolith, some outcrops of Baoligaomiao Formation occur in QW	None
QSE	Quaternary regolith, located between SPS and SPSE, some outcrops of Baoligaomiao Formation and biotite adamellite occur inside.	None
QC	Quaternary regolith	None
QSW	Quaternary regolith	None

Table 4
p-values of log-transformed element concentration of 18 geo-objects.

	Cu	Zn	Ni	Pb	Ag	Co	Mo	W	As	Sb	Bi	Hg	Au	Sn	NO
ICC	**0.671**	**0.058**	**0.713**	**0.235**	0.000	**0.444**	0.007	0.000	**0.125**	0.003	**0.051**	0.002	0.000	0.024	7
ICW	**0.415**	0.007	**0.493**	**0.721**	**0.096**	**0.258**	0.000	0.000	**0.162**	0.000	0.023	**0.618**	0.000	**0.153**	6
IE1	**0.091**	0.029	**0.722**	**0.145**	**0.268**	**0.083**	0.000	0.000	0.044	0.035	**0.133**	**0.317**	**0.206**	0.000	6
IE2	**0.062**	**0.116**	0.006	**0.443**	**0.779**	0.045	0.000	0.000	0.042	**0.109**	**0.071**	0.012	**0.253**	**0.081**	6
IN1	0.007	**0.454**	0.004	**0.776**	0.001	**0.691**	0.000	0.000	**0.438**	0.004	0.000	**0.083**	0.014	0.001	9
IN2	0.005	**0.162**	**0.213**	**0.210**	0.001	**0.823**	0.000	0.014	0.011	0.012	0.030	0.015	0.034	**0.959**	9
INE	**0.183**	**0.259**	**0.492**	**0.448**	0.035	**0.368**	0.000	0.000	**0.313**	0.002	**0.338**	0.022	0.000	0.018	7
IS	0.035	0.007	**0.099**	0.001	0.001	0.012	0.000	0.000	**0.590**	**0.079**	0.006	0.002	0.000	0.001	11
SOC1	**0.528**	0.000	**0.056**	0.010	**0.974**	**0.138**	0.013	**0.071**	**0.528**	0.027	0.000	**0.185**	0.000	0.000	7
SOC2	**0.093**	**0.184**	**0.848**	**0.686**	**0.196**	**0.626**	0.004	**0.251**	**0.734**	**0.189**	0.000	**0.772**	0.007	**0.378**	3
SOC3	**0.468**	**0.333**	**0.347**	**0.322**	**0.513**	0.004	0.017	**0.052**	**0.635**	**0.052**	**0.202**	**0.524**	0.000	**0.298**	3
SPCE1	**0.321**	**0.353**	**0.065**	**0.113**	**0.386**	**0.947**	**0.065**	**0.431**	0.003	**0.350**	**0.428**	**0.974**	0.042	**0.747**	2
SPCE2	**0.547**	**0.501**	**0.475**	**0.886**	**0.961**	**0.457**	**0.072**	**0.362**	**0.773**	**0.235**	0.009	**0.140**	0.033	**0.107**	2
SPCN	**0.164**	**0.685**	**0.102**	**0.669**	**0.812**	0.029	**0.537**	0.005	0.014	0.029	**0.254**	**0.498**	0.007	**0.223**	5
SPCW1	**0.906**	**0.525**	**0.887**	**0.133**	**0.414**	**0.828**	**0.617**	**0.699**	**0.996**	**0.823**	**0.603**	**0.131**	**0.241**	0.002	1
SPCW2	0.024	**0.164**	**0.366**	0.004	**0.513**	0.034	**0.263**	**0.690**	0.019	**0.857**	0.002	**0.661**	0.021	0.033	7
SPSE1	**0.728**	0.037	0.033	**0.053**	0.007	**0.801**	0.007	**0.911**	**0.396**	**0.931**	**0.153**	**0.954**	**0.064**	**0.058**	4
SPSE2	0.030	0.001	0.004	**0.305**	0.011	0.000	**0.091**	0.008	**0.355**	**0.429**	**0.465**	**0.543**	0.000	0.001	8
NE	5	6	4	3	7	6	12	10	6	8	8	5	14	9	

W, Bi, Sn, and Sb are involved in the ore-forming process of Mo deposits; Au tends to aggregate in exogenic processes. Those processes tend to generate concentration values that follow power-law distributions but log-normal distributions (Cheng, 2012).

The table shows that eight elements of Cu, Zn, Ni, Pb, Ag, Co, As and Hg pass the log-normal test, therefore we can use datasets of 18 geo-objects with those 8 elements to test the method. It is important to note that they only pass a single variable normal test but do not pass any multivariate normal distribution test. The number of elements can be extended to 14 along with 20 geo-objects to test the method under situations where the data are non-normal. Thus, there are four available datasets for assessing the KL-Divergence method: 18 geo-objects with eight elements (G18E8), 18 geo-objects with 14 elements (G18E14), 20 geo-objects with eight elements (G20E8), 20 geo-objects with 14 elements (G20E14). The datasets in G18E8 approximate normal distributions, the datasets in G18E14 and G20E8 are moderately normal, and the datasets in G20E14 are highly non-normal.

4. Data analysis

4.1. Results using dataset G18E8

The results of the new method using dataset G18E8 are shown in Fig. 4. Fig. 4a shows the results of HCA based on D_{KLSM}. It can be observed that 18 geo-objects are divided into two clusters which represents the stratigraphy and intrusions respectively. Also, there are some small clusters with the geo-objects that have a similar composition or a close spatial connection. For example, IN1-INE-IN2 are the intrusions located in the north; SOC2—SOC3 are geo-objects belong to Bayanhushu Formation; SOC1-SPCE2 are two adjacent geo-objects.

Fig. 4b shows the results of HCA based on D_{KLSC} with the eight-element datasets. In the dendrogram, geo-objects that have similar compositions or have a close spatial relationship tend to form clusters. For example: in clusters SPSE1-SPSE2, SPCW1-SPCW2 and IE1-IE2, the geo-objects have the same components that form clusters. The cluster SPCN-SPCE-SOC1-SPCE2-SOC2-SOC3 consists of geo-objects having close spatial relationships, and the cluster can be further divided into three small clusters SPCN-SPCE1, SOC1-SPCE2 and SOC2—SOC3, with each of them consisting of two adjacent geo-objects.

Fig. 4c shows a dendrogram based on D_{KLS}, which is a hybrid of results that uses D_{KLSC} and D_{KLSM}. In the figure, the intrusion geo-objects and strata geo-objects are generally well separated (except SPSE2). As well, there are many small clusters consisting of geo-objects that have similar compositions or spatial relationships, such as IN1-INE-IN2, IE1-IE2, SPCW2-SPCW1, SOC2—SOC3, SPCE1-SOC1-SPCE2.

4.2. Results using dataset G18E14

The results of HCA using dataset G18E14 are shown in Fig. 5. By comparing Figs. 5a and 4a, it is observed that for the dissimilarity D_{KLSM}, the dendrogram using fourteen elements is similar with the dendrogram using eight elements; the strata and intrusions are well separated into two clusters. Moreover, in Fig. 5a, there are more small clusters that present geo-objects with strong associations: IN1-INE-IN2, SPCW2-SPCW1, SOC1-SPCE2, SPCE1-SOC2-SOC3. For D_{KLSM}, the dendrogram using fourteen-element datasets (Fig. 5b) is similar to the dendrogram using eight-element datasets (Fig. 4b), but it loses the cluster of IE1-IE2. With respect to D_{KLS}, the fourteen-element datasets yield a better dendrogram (Fig. 5c); the strata and intrusion are correctly separated into two clusters, and there are several small clusters that correctly show the strong associations between the geo-objects, including SPSE1-SPSE2, SOC2—SOC3, SPCE1-SOC1-SPCE2, SPCW2-SPCW1.

By observing the results from the G18E8 and G18E14 datasets, it is apparent that some geo-objects are similar. However, there are some cases where the clusters to not form as expected (e.g., SOC1 combines with SPCE1 but not SOC2 and SOC3). More significantly, some intrusive rocks do not group together (e.g., IE1-IE2, IN1-IN2). This may be due to hydrothermal events that have changed the geochemical composition of some of the geo-objects.

4.3. Result using dataset G20E8 and G20E14

The results using D_{KLS} on dataset G20E8 are shown in Fig. 6a. It is observed that the strata and intrusions are correctly separated, and many adjacent geo-objects form small clusters such as QSW-SJSW, SPSE-QSE-SPS. Although the Mo mineralization associated elements are not in the dataset, the Mo deposit-associated geo-objects still forms clusters (e.g., INW—ICN, SPNW-SPCN). Moreover, the strata geo-objects cluster according to their geospatial locations and two clusters are formed; one is located in the south, and the other is for the geo-objects located in the north.

The results using D_{KLS} on dataset G20E14 are shown in Fig. 6b. There are many small clusters that represent adjacent geo-objects and a cluster that represents geo-objects associated with mineralization. The corresponding dendrogram based on 14 elements has less structure than the dendrogram that uses 8 elements (Fig. 6b). The geo-objects of the intrusion INW are misclassified and cluster with geo-objects associated

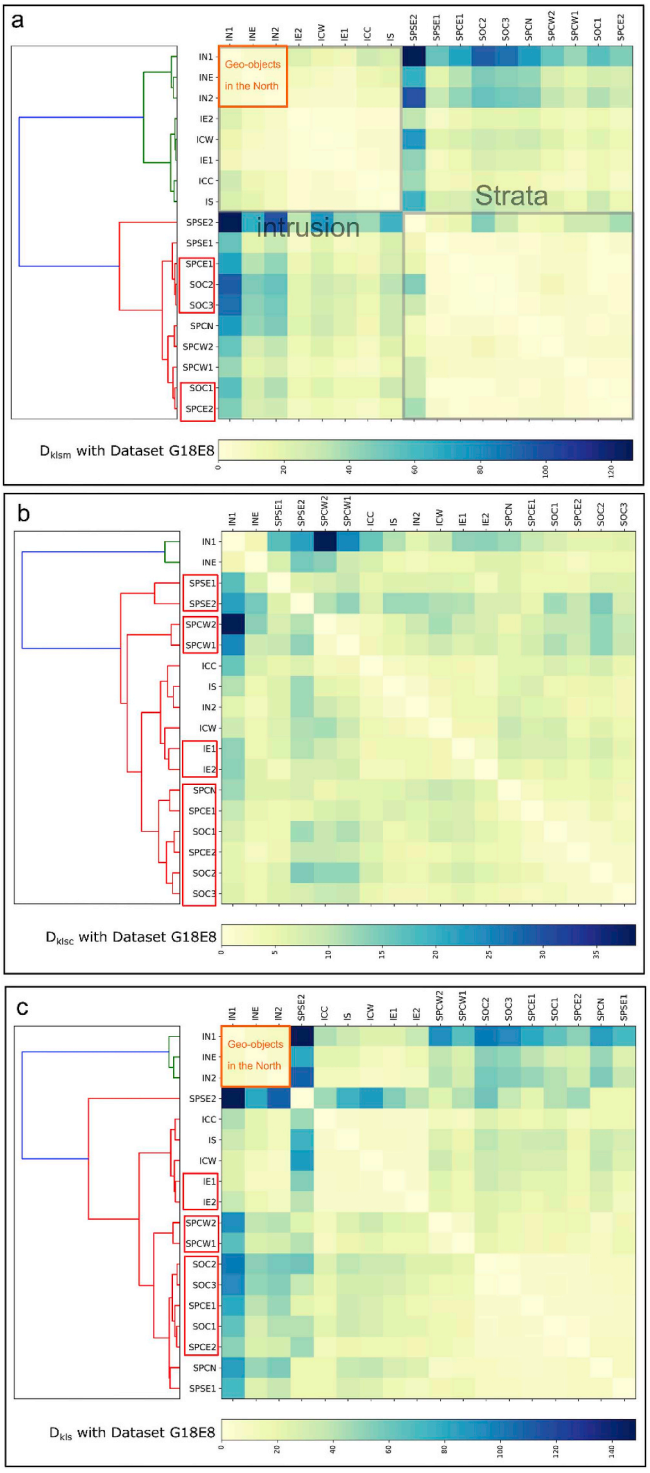

Fig. 4. Results of HCA using D_{KLSM}, D_{KLSC} and D_{KLS} as the dissimilarity measure with G18E8. (a) D_{KLSM} is applied, (b) D_{KLSC} is applied, (c) D_{KLS} is applied. In each subplot, the matrix on the right represents the dissimilarity between 18 geo-objects, a darker color means more differences, the dendrogram on the left was constructed from the matrix, and under the matrix is the color bar for the dissimilarity value. The adjacent geo-objects are highlighted with red box, cluster have specific meanings are highlighted in the matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

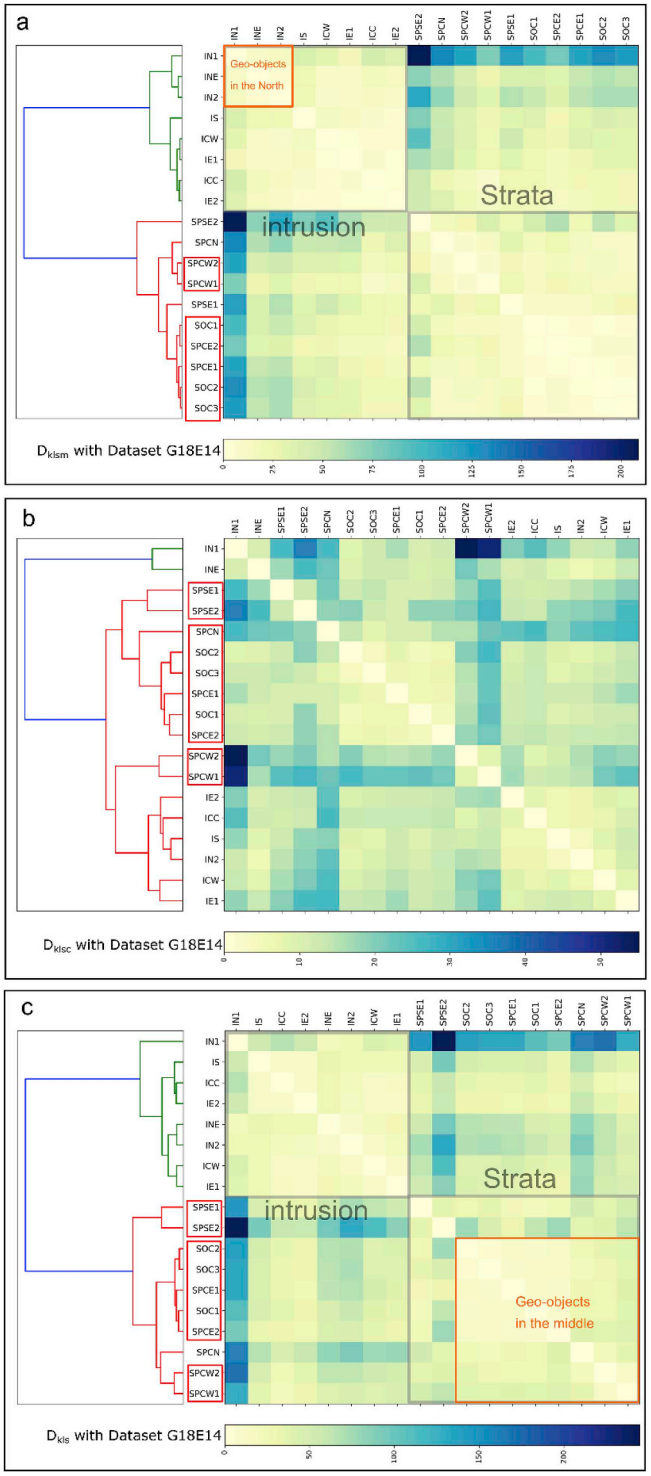

Fig. 5. Results of HCA using D_{KLSM}, D_{KLSC} and D_{KLS} as the dissimilarity measure with G18E14. (a) D_{KLSM} measure, (b) D_{KLSC} measure, (c) D_{KLS} measure.

with strata.

By comparing the results using different datasets, it is apparent that 1) when the geo-objects are less influenced by mineralization processes, adding elements that failed to pass the normality test, there is little impact on the results, although in some cases it yields better results (e.g., Fig. 5a vs. Fig. 4a); 2) When the variables close to being normally distributed, the extent of the area of the geo-objects might result in a more interpretable dendrogram and, 3) When the geo-objects are influenced by processes related to mineralization events, adding mineralization associated elements might lead to worse results. This implies that the "normal distribution" requirement on data is not a very strict condition.

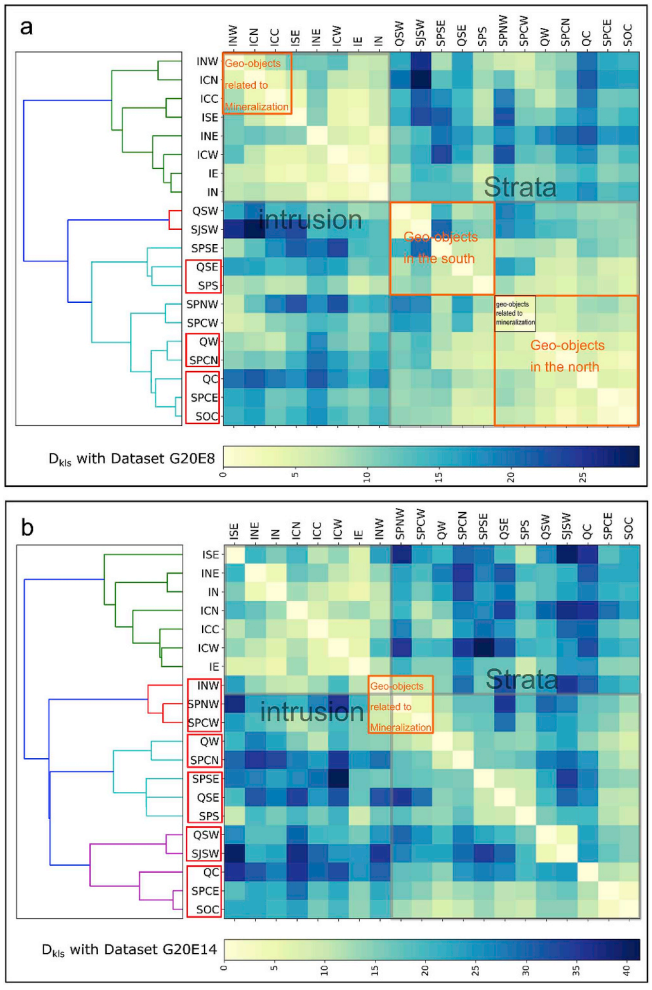

Fig. 6. Results of HCA using D_{KLS} on different datasets (a) G20E8; (b) G20E14.

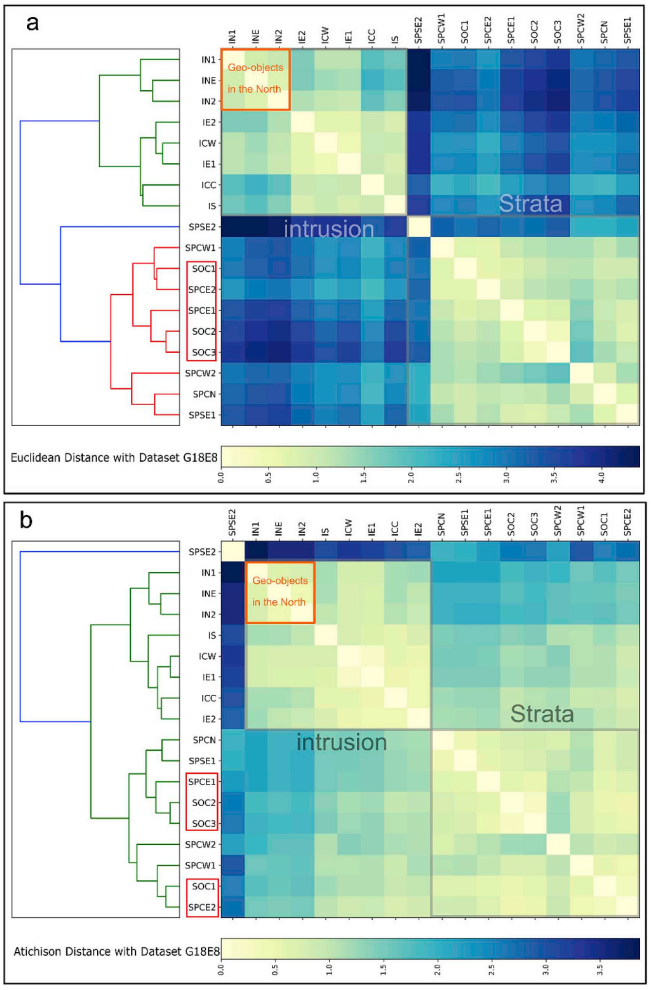

Fig. 7. Results of HCA using Euclidean distance and Aitchison distance as the dissimilarity measures with dataset G18E8. (a) Euclidean distance; (b) Aitchison distance.

4.4. Comparison with euclidean distance and aitchison distance

The Euclidean distance and Aitchison distance between mean vectors of datasets can also be used as a measure of dissimilarity in HCA. Fig. 7a shows a dendrogram using the Euclidean distance based on the log-transformed and standardized dataset G18E8. The dendrogram is very similar with the dendrogram using D_{KLSM} in Fig. 4a. The strata and intrusions are correctly separated into two clusters, and there are a few clusters that represent geo-objects, which have close relationships such as IN1-INE-IN2, SPCW1−SOC1-SPCE2-SPCE1-SOC2-SOC3.

The dendrogram using the Aitchison distance and dataset G18E8 is shown in Fig. 7b. This dendrogram has similar features, but the features show less contrast than the dendrogram using the Euclidean distance (Fig. 7a). It is worth noting that the geo-object SPSE2 is an isolated leaf, does not belong to any main cluster represents intrusions or strata.

A comparison of the different proximity measures including the Euclidean distance, Aitchison distance and D_{KLSM} generate similar dendrograms since they are based on the mean. Although the Aitchison distance is scale invariant, it does not perform any better than the Euclidean distance, which is not scale invariant. Although the Aitchison and Euclidean distances are metrics (Aitchison et al., 2000), the results are not better than the measure of D_{KLSM}. Moreover, measures based on the KL-divergence are scale invariant and can provide a new perspective (D_{KLSC}) on the observed patterns of the data.

5. Conclusion

The application of the HCA method shows that measures of KL-divergence can describe the dissimilarity of pairwise geochemical datasets based on geo-objects, and the HCA method can give a comprehensive view of geo-object associations. Additionally, the decomposition components of D_{KLS}, D_{KLSC} and D_{KLSM}, can further characterize dissimilarity in two aspects: the rock types that are reflected via D_{KLSM}, and the spatial relationships and component similarities of geo-objects that are revealed via D_{KLSC}. This indicates that the information about rock type is mainly explained by the mean (center of the dataset), and the information about the spatial and component characteristics is likely explained by the covariance (shape of dataset).

Although the method presented here requires a hypothesis of a normally distributed dataset the results show that strict normality is not essential. Similar phenomena also can be found in other applications with geochemical data. For example, Templ et al. (2008) show that MCLUST, a clustering algorithm based on a hypothesis that the clusters that are formed from normally distributed data (Fraley and Raftery, 1999), provide the most reliable and interpretable results compared with other clustering techniques. The method presented here shows that is essential to prepare the data to closely approximate normality.

The results of the current research not only provide a powerful data mining tool that reveals the relationships between geo-objects through geochemical data, but also reveal that measures derived from KL-divergence can characterize geochemical characteristics for a more meaningful interpretation.

CRediT authorship contribution statement

Jie Yang: Writing-original draft, Formal analysis, Writing-original draft. **Eric Grunsky:** Writing-review & editing. **Qiuming Cheng:** Supervision.

Acknowledgments

We thank three anonymous reviewers for their helpful comments. This research benefited from financial support from the National Key Research and Development Program of China (2016YFC0600501), the National Natural Science Foundation of China (No. 41430320, 41602337) and a Chinese Geological Survey project (Minerals and Geological Prospecting on Shallow Covered Areas of Jinning, Inner Mongolia, No. DD20160045).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2018.11.003.

Computer code availability

A program has been prepared in Python. The source code can be found on GitHub: https://github.com/sinixyang/KLD-Based-HCA. The library includes Numpy, Pandas, and Matplotlib, which are required to run the program. More information about the program can be found in the file “README.md.”

References

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 271–275.
Allen, A.P., Li, B.L., Charnov, E.L., 2001. Population fluctuations, power laws and mixtures of lognormal distributions. Ecol. Lett. 4, 1–3.
Amari, S.-i., Cichocki, A., 2010. Information geometry of divergence functions. Bull. Pol. Acad. Sci. Tech. Sci. 58, 183–195.
Basseville, M., 2013. Divergence measures for statistical data processing—an annotated bibliography. Signal Process. 93, 621–633.
Cheng, Q., Agterberg, F., Ballantyne, S., 1994. The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51, 109–130.
Cheng, Q.M., 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.M., 2012. Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. J. Geochem. Explor. 122, 55–70.
Cheng, Q.M., Agterberg, F.P., 2009. Singularity analysis of ore-mineral and toxic trace elements in stream sediments. Comput. Geosci. 35, 234–244.
Cichocki, A., Amari, S.-i., 2010. Families of alpha-beta-and gamma-divergences: flexible and robust measures of similarities. Entropy 12, 1532–1568.
Clement, P., Desch, W., 2008. An elementary proof of the triangle inequality for the Wasserstein metric. Proc. Am. Math. Soc. 136, 333–339.
Contreras-Reyes, J.E., Arellano-Valle, R.B., 2012. Kullback–Leibler divergence measure for multivariate skew-normal distributions. Entropy 14, 1606–1626.
Everitt, B., Landau, S., Leese, M., 2011. Cluster Analysis, fifth ed. John Wiley & Sons Ltd., United Kingdom.
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.
Fraley, C., Raftery, A.E., 1999. MCLUST: software for model-based cluster analysis. J. Classif. 16, 297–306.
Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochem. Explor. Environ. Anal. 10, 27–74.
Hernandez, P.A., Perez, N.M., Salazar, J.M.L., Ferrell, R., Alvarez, C.E., 2004. Soil volatile mercury, boron and ammonium distribution at Canadas caldera, Tenerife, Canary Islands, Spain. Appl. Geochem. 19, 819–834.
Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95.
Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 264–323.
Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. Lond. A, Math. Phys. Sci. 453–461.
Jones, E., Oliphant, T., Peterson, P., 2001. SciPy: open source scientific tools for Python. http//www.scipy.org.
Kailath, T., 1967. The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. Circ. Theor. 15, 52–60.
Kanamori, T., Fujisawa, H., 2014. Affine invariant divergences associated with proper composite scoring rules and their applications. Bernoulli 20, 2278–2304.
Kremer, B., Owocki, K., Krolikowska, A., Wrzosek, B., Kazmierczak, J., 2012. Mineral microbial structures in a bone of the Late Cretaceous dinosaur Saurolophus angustirostris from the Gobi Desert, Mongolia - a Raman spectroscopy study. Palaeogeogr. Palaeoclimatol. Palaeoecol. 358, 51–61.
Kullback, S., 1978. Information Theory and Statistics. Dover Publications.
Li, Y., Zuo, R.G., Bai, Y., Yang, M.G., 2014. The relationships between magnetic susceptibility and elemental variations for mineralized rocks. J. Geochem. Explor. 146, 17–26.
Liu, Y., Cheng, Q., Xia, Q., Wang, X., 2014. Identification of REE mineralization-related geochemical anomalies using fractal/multifractal methods in the Nanling belt, South China. Environ. Earth Sci. 72, 5159–5169.
Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions. Internet Math. 1, 226–251.
Mokhtari, A.R., Rodsari, P.R., Fatehi, M., Shahrestani, S., Pournik, P., 2014. Geochemical prospecting for Cu mineralization in an arid terrain-central Iran. J. Afr. Earth Sci. 100, 278–288.
Mumm, A.S., Dart, R.C., Say, P., 2012. Hematite/Maghemite trace element geochemistry in base metal exploration. J. Geochem. Explor. 118, 1–13.
Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. Comput. J. 26, 354–359.
Murtagh, F., Contreras, P., 2012. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 2, 86–97.
Nezhad, M.T.K., Tabatabaii, S.M., Gholami, A., 2015. Geochemical assessment of steel smelter-impacted urban soils, Ahvaz, Iran. J. Geochem. Explor. 152, 91–109.
O'Shea, B., Jankowski, J., 2006. Detecting subtle hydrochemical anomalies with multivariate statistics: an example from 'homogeneous' groundwaters in the Great Artesian Basin, Australia. Hydrol. Process. 20, 4317–4333.
Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional Data Analysis: Theory and Applications. Wiley.
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. Environ. Geol. 39, 1001–1014.
Reimann, C., Filzmoser, P., Garrett, R.G., 2002. Factor analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 17, 185–206.
Steerneman, T., 1983. On the total variation and Hellinger distance between signed measures; an application to product measures. Proc. Am. Math. Soc. 88, 684–688.
Tao, J.Z., R, Wang, T., 2009. Strategically Mineral Prospecting in Dalaimiao District. Geological Survey institute of Inner Mongolia.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 23, 2198–2213.
Walt, S.v.d., Colbert, S.C., Varoquaux, G., 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13, 22–30.
Zhang, C.S., Fay, D., McGrath, D., Grennan, E., Carton, O.T., 2008. Statistical analyses of geochemical variables in soils of Ireland. Geoderma 146, 378–390.
Zhang, C.S., Manheim, F.T., Hinde, J., Grossman, J.N., 2005. Statistical characterization of a large geochemical database and effect of sample size. Appl. Geochem. 20, 1857–1874.