Available online at www.sciencedirect.com
ScienceDirect
Journal of Geochemical Exploration 93 (2007) 109 – 119
JOURNAL OF GEOCHEMICAL EXPLORATION
www.elsevier.com/locate/jgeoexp

Semi-hierarchical correspondence cluster analysis and regional geochemical pattern recognition

Hongjin Ji [a,*], Daoming Zeng [a], Yanxiang Shi [a], Yangang Wu [b], Xisheng Wu [a]

[a] *Department of Geochemistry, College of Geo-Exploration Science and Technology, Jilin University,*
6 Ximinzhu Street, Changchun, 130026, PR China
[b] *Department of Geophysics, College of Geo-Exploration Science and Technology, Jilin University,*
6 Ximinzhu Street, Changchun, 130026, PR China

Received 5 April 2006; accepted 29 October 2006
Available online 18 December 2006

Abstract

Semi-hierarchical correspondence cluster analysis (SHCCA), firstly developed in this paper, extracts the main advantages of correspondence analysis, hierarchical and non-hierarchical cluster analysis, and unifies the R- and Q-mode cluster analysis of large data set. A systemic program to recognize the regional geochemical patterns is built up based on this method. With this program, the complex tasks for data interpretation can be achieved by simple processes, and important geochemical information can be displayed by a single diagram, i.e. the multivariate regional geochemical image. As one of the applied examples of this program, the regional geochemical pattern recognition for a shallow covered area around Tahe in Heilongjiang Province is introduced. The results show that many hidden geochemical patterns related to the lithologies, structures, ore-forming conditions and prospecting targets etc are revealed by the geochemical image, and that the main geochemical patterns are related with certain geological and gravitational patterns. By finding contrasts between geochemical patterns and geological or gravitational patterns, the SHCCA results assist the geological mapping in this area. Geochemical data obtained in Chinese regional geochemical exploration provides useful information regarding geology and minerals, and the method described in this paper provides a new way to examine this type of resource.

Keywords: Multivariate statistical analysis; Correspondence cluster analysis; Large data set; Regional geochemistry; Data interpretation; Geological mapping

1. Introduction

Geochemical pattern recognition has been used in exploration geochemistry for a long time. Since the 1970s, pattern recognition techniques have been applied to recognize the geological and economic–mineralogical information hidden in geochemical data (Howarth, 1973; Castillo-Munoz and Howarth, 1976; Gustavsson and Bjorklund, 1976; Xie, 1979) and to establish multivariate geochemical background pattern (Lindqvist et al., 1987). Recently, it has also been applied to recognize the polluted sites and types (Hanesch et al., 2001), and to investigate the relations between regional geochemical patterns and large ore deposits (Xie et al., 2004).

Typical pattern recognition methods used in geochemical exploration mainly consist of discriminant analysis

* Corresponding author. Fax: +86 431 88524544.
E-mail address: jih_j@sina.com (H. Ji).

0375-6742/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gexplo.2006.10.002

and cluster analysis. The application of discriminant analysis is limited because training samples are often hard to obtain, so, only cluster analysis is considered in this paper. It is evident that cluster analysis, as a tool of geochemical pattern recognition, should provide, firstly, clustering of large data sets, because so many samples typically occur in regional geochemical exploration; Secondly, it is desirable that the relationship between R-mode and Q-mode cluster is shown, so that the types of sample can be interpreted in terms of the corresponding types of variable. But it is difficult to meet these conditions with existing clustering methods.

Classifying methods which adapt to large data sets are the various non-hierarchical cluster analysis methods (Mather, 1976; Treiger et al., 1995; Eddy et al., 1996; Velthuizen et al., 1997; Wei et al., 2003), and the *c*-means cluster analysis has been successfully applied to geochemical exploration (Rantitsch, 2000; Hanesch et al., 2001). With these methods, however, only Q-mode cluster analysis can be carried out, and the R-mode analysis needs to be executed by other methods (Shi and Carr, 2001; Reimann et al., 2002). In addition, the number of the classes needs to be subjectively chosen in *c*-means cluster analysis, and it is not so convenient in applications.

A clustering method which can show the corresponding relation between R- and Q-mode clusters is Hierarchical Correspondence Cluster Analysis (HCCA) developed by Ji and Zhong (1991), Ji et al. (1995a). This method has already been successfully applied in studies of the geochemistry of environment (Ge and Wu, 1995), element medicine (Ji et al., 2004), economics (Hong et al., 1998; Jia and Hong, 2000) and sociology (Ji et al., 1995b) etc. Unfortunately, it is not suited to large data sets, just as with other hierarchical cluster analysis.

In order to adapt geochemical pattern recognition to study large data sets, the main aim of this paper are: (1) to describe a new classification method semi-hierarchical correspondence cluster analysis (SHCCA), (2) to apply SHCCA to regional geochemical pattern recognition, (3) to introduce the method with a practical example of its application to a shallow covered area, and (4) to discuss some related problems.

2. Semi-hierarchical correspondence cluster analysis

The main theoretic basis of SHCCA is correspondence analysis (Benzécri, 1973; David et al., 1974). Let us suppose a data matrix with n samples and m variables to be $\boldsymbol{X}=(x_{ij})_{n\cdot m}$, where x_{ij} is the observed value of the jth variable in ith sample, $x_{ij}\geq 0$, and

$$r_i = x_{i1} + x_{i2} + \ldots + x_{im} > 0; i = 1, 2, \ldots, n \tag{1}$$

$$c_j = x_{1j} + x_{2j} + \ldots + x_{nj} > 0; j = 1, 2, \ldots, m. \tag{2}$$

If we use it to define diagonal matrices

$$\boldsymbol{R} = \mathrm{diag}(r_1, r_2, \ldots, r_n) \tag{3}$$

$$\boldsymbol{C} = \mathrm{diag}(c_1, c_2, \ldots, c_m) \tag{4}$$

the similarity matrix between variables in correspondence analysis can be expressed as

$$\boldsymbol{H} = \boldsymbol{C}^{-1/2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{-1/2} \tag{5}$$

where $\boldsymbol{X}^{\mathrm{T}}$ is the transposed matrix of $\boldsymbol{X}$. The largest eigenvalue of $\boldsymbol{H}$ is equal to 1, but this value corresponds to a meaningless eigenvector (Dong et al., 1979). Therefore, ignoring the eigenvalue of 1, we always select anther p ($<m$) greater eigenvalues to be $1 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$, and the corresponding unit eigenvectors

$$\boldsymbol{a}_j = (\boldsymbol{a}_{1j}, \boldsymbol{a}_{2j}, \ldots, \boldsymbol{a}_{mj})^{\mathrm{T}}; j = 1, 2, \ldots, p. \tag{6}$$

Let the matrices

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \tag{7}$$

$$\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p) \tag{8}$$

then two important results of the correspondence analysis can be expressed as

$$\boldsymbol{U} = \boldsymbol{C}^{-1/2}\boldsymbol{A}\boldsymbol{\Lambda}^{1/2} = (u_{ij})_{m\cdot p} \tag{9}$$

$$\boldsymbol{V} = \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{C}^{-1/2}\boldsymbol{A} = (v_{ij})_{n\cdot p} \tag{10}$$

where $\boldsymbol{U}$ is the $m\cdot p$ order factor loadings matrix for R-mode, its m rows can be regarded as m variable points in p-dimensional factor space, and $\boldsymbol{V}$ is the $n\cdot p$ order factor loadings matrix for Q-mode, its n rows can be regarded as n sample points in the same space. Based on the principle of correspondence analysis, the characteristics of the sample points can be interpreted by the neighboring variable points. So, the large data set can be easily classified by two steps.

The first step is to partition the types of the variable by hierarchical cluster analysis. The similarities between variable points i and j can be defined by the Euclidean distance

$$d_{ij} = \left[\sum_{k=1}^{p} (u_{ik} - u_{jk})^2\right]^{1/2}; i,j = 1, 2, \ldots, m \tag{11}$$

or the distance matrix $\boldsymbol{D}=(d_{ij})$; i, j=1, 2, …, m. By executing hierarchical cluster analysis with matrix $\boldsymbol{D}$, we can obtain a dendrogram of m variable points. Let us suppose that the m variables in the dendrogram can be classified into g groups: s_1, s_2, …, s_g, and all of the groups have practical meaning. Of course, unreasonable variable groups may sometimes appear in the result, it, as a first important question, will be discussed in Section 5 of the paper.

The second step is to recognize the types of sample by non-hierarchical cluster analysis. Similar to the c-means clustering, the m variable points can be regarded as the centers of clustering and the similarities between sample point i and variable point j can be defined by the Euclidean distance

$$d_{ij} = \left[\sum_{k=1}^{p} (v_{ik} - u_{jk})^2\right]^{1/2}; i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m \tag{12}$$

for sample point i, if

$$d_{ik} = \min(d_{i1}, d_{i2}, \ldots, d_{im}); i = 1, 2, \ldots, n \tag{13}$$

and the kth variable $x_k \in s_w$, then the sample point i will be classified into the group s_w.

It is clear that the main advantages of correspondence analysis, hierarchical and non-hierarchical cluster analysis are combined in this method. Since its basis is correspondence analysis, and the hierarchical and non-hierarchical cluster analysis is respectively used in half of the steps, it is called *semi-hierarchical correspondence cluster analysis*. It not only unifies the R- and Q-mode cluster analysis, i.e. shows the corresponding relations between the types of variable and sample, just as the HCCA, but also achieves the clustering of large data set, so it is a development of the HCCA.

Unlike c-means cluster analysis, the number of the classes in SHCCA can be selected according to the variable groups in the dendrogram without relying on subjective experience. In order to extract sufficient information, the number of factors in the correspondence analysis, p, can be chosen as the number of all positive eigenvalues less than unity.

3. Program for regional geochemical pattern recognition

A systemic program for regional geochemical pattern recognition can be built up based on the SHCCA method. Experience shows that the key tasks in geochemical data interpretation can be achieved with the following steps:

(1) Recognition of the types of element and sample: by executing SHCCA with the $n \cdot m$ data matrix from the study area, the m elements and n samples are divided into g groups or types, where each element group should possess practical geochemical significance.

(2) Recognition of geochemical units: the different types of sample are marked at the sampling locations with different symbols or colors, and then the geochemical units will be delineated. A geochemical unit is defined here as an area inhabited by the same type of samples which are nearby each other; but the presence of a few other types of sample in the unit can be ignored. The characteristics of the unit will be interpreted by the corresponding element group. For example, the acidic or alkaline units are represented by the element group: Si, K and Na etc, and the basic units are represented by the group: Fe, V, Ti and Mg etc.

(3) Recognition of geochemical fault traces: if a borderline of geochemical unit is approximately linear, it can be thought of as a geochemical fault trace, an inferential geological fault. If such a trace is followed by a unit corresponding to the volatile element group, As, Sb and Hg etc, it may be related to hydrothermal activities.

(4) Recognition of geochemical zoning: this may be recognized from the spatial relations among different types of geochemical unit.

(5) Recognition of the ore-forming conditions and prospecting targets: ore-forming geochemical conditions, related to the lithologies, structures, hydrothermal activities and zoning etc, may be recognized by comparing the various geochemical patterns with the presence of known mineralization. Sections where the geochemical conditions are similar to the known ore units can be regarded as prospecting targets.

(6) Expression of the results: the results can be shown using a single diagram, i.e. the multivariate *geochemical image*. It shows all of the geochemical patterns as above, other than usual information, such as the univariate anomalies or their composition etc. It is the main reason why the term "geochemical pattern recognition" is used in this paper. Such a diagram, makes it very convenient to compare the geochemical patterns with geological and geophysical patterns.

4. Applied example

High quality geochemical data for more than 30 elements have been obtained in Chinese regional geochemical exploration (Xie et al., 2004). This means that

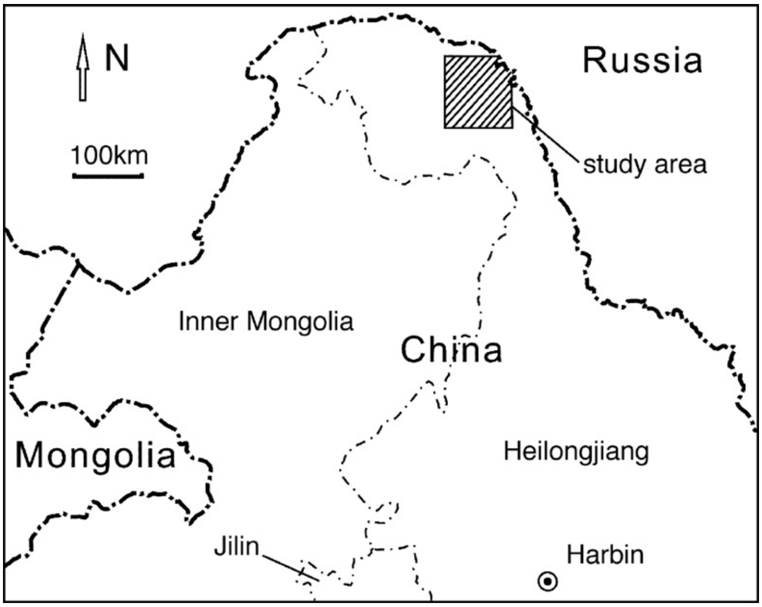

Fig. 1. Location map of the study area.

the regional geochemical pattern recognition program can be expediently carried out, and successful applications have been obtained from geology and mineral research in the eastern Kunlun and Heilongjiang areas. In this paper, only one example, related with the regional geological mapping in the shallow covered area around Tahe town in Heilongjiang province (a region on China's northeast border, abutting Russia and Inner Mongolia, Fig. 1), is introduced to illustrate the effects of the method.

4.1. Study area and data

The study area is located in northern Heilongjiang province and covers about 11,600 km^2. Because most parts of the area are covered by forest, swamp and residual soil the geological borderlines delineated in the geological map (Fig. 2) are tentative.

Based on the geological map, the study area can be divided into four main geological types. (1) Early Proterozoic basement rocks, mainly distributed in the south–east parts of the area. It includes magmatic rocks and metamorphic rocks; the former consists of migmatitic granites, gneissic granites and granodiorites etc, and enclose a lot of basic residual bodies; the latter consists of gneiss, schists and amphibolites etc. Biao et al. (1999) consider that this type should belong to the granite–greenstone mass and is closely related to the gold ore deposits in this area. (2) Paleozoic granites mainly distributed along the north bank of the Huma River. A smaller number are distributed beside the south bank as well as the south–west parts of the area. (3) Mesozoic rocks located in the northern parts of the area, where parts of the north–west area are volcanics, the other parts are geologically undivided volcanics and sedimentary rocks. (4) Mesozoic volcanics located in the south–west and south–middle parts of the area.

The geological fault structures and mineralization occurrences are marked in the geological map. The directions of the faults are shown to be both north–west and north–east, but they are also very tentative. The main known mineralization is gold and polymetallic. All of the gold ore occurrences, except for one that is located in a Mesozoic erathem near Baoxingou, are distributed within or near the Early Proterozoic rocks around Hanjiayuan-Donggou. The only polymetallic ore occurrence is located in Mesozoic erathem near Shiwuliqiao.

A study program to apply the regional geophysical and geochemical data to geological mapping is being carried out in this area for more detailed geological research. The example discussed in this paper is part of the results in the program. The geochemical data set, from the stream sediments of the <0.3 mm fraction obtained in Chinese regional geochemical exploration, is provided by the Heilongjiang Geology Survey and is utilized in this example, it contains 2727 samples and 35 components, i.e. the samples are all analysed: by X-ray fluorescence spectrometry (XRF) for SiO_2, Al_2O_3, Fe_2O_3, K_2O, Ti, P, Zr, Y, Nb, Rb, Pb and Cr; by

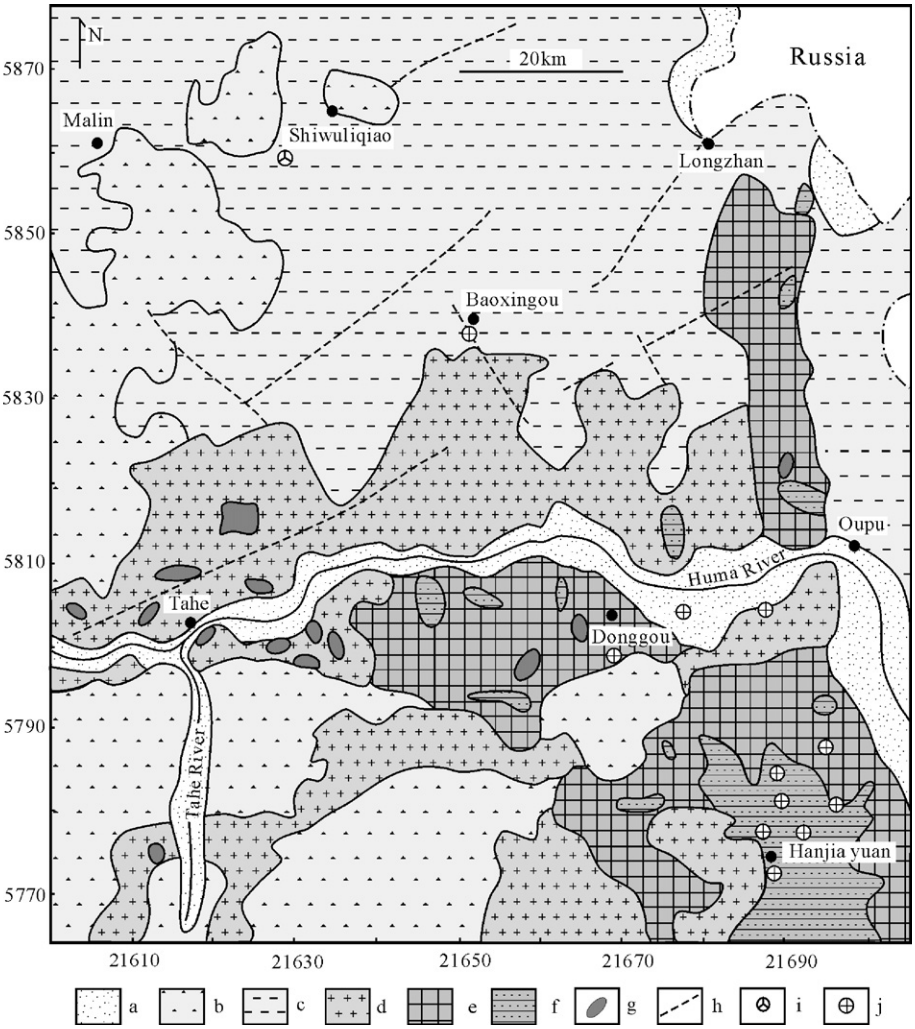

Fig. 2. Simplified geological map of Tahe area, Heilongjiang province. a: Quaternary alluvium; b: Mesozoic volcanics; c: Mesozoic volcanics and sedimentary rocks; d: Paleozoic granites; e: Early Proterozoic magmatic rocks; f: Early Proterozoic erathem metamorphic rocks; g: gabbro; h: fault structure; i: polymetallic ore occurrence; j: gold ore occurrence.

inductively coupled plasma atomic emission spectrometry (ICPAES) for CaO, MgO, Na_2O, Mn, Ba, Sr, Zn, Cu, Co, Ni, V, Li and Be; by hydride generation atomic fluorescence spectrometry (HGAFS) for As, Sb, Bi and Hg; by emission spectrometry (ES) for B, Sn and Ag; by polarography (POL) for W; by ion selective electrode (ISE) for F, and by graphite furnace atomic absorption spectrometry (GFAAS)for Au. Each sample represents a $2 \cdot 2$ km^2 grid square.

4.2. Recognition of the types of element and sample

The geochemical pattern recognition was performed by applying SHCCA to the data set of the $2727 \cdot 35$ matrix; the number of factors, p, chosen was 34. According to the dendrogram shown in Fig. 3, the 35 components are divisible into 6 element groups:

A. Au, Hg, As, Sb, B, Bi and W;
B. SiO_2, K_2O, Na_2O, and Zr;
C. Al_2O_3, Li, F, Cu, Pb, Zn and Ag;
D. Be, Nb, Sn, Y and Rb;
E. CaO, Mn, Sr and Ba;
F. Co, Fe_2O_3, P, Cr, Ni, V, Ti and MgO.

There is a distinct geochemical and geological significance to each element group. Most of the elements in group A have stronger migrating ability, and are usually related to the fault structures and hydrothermal activity (Liu et al., 1984). Group B and D

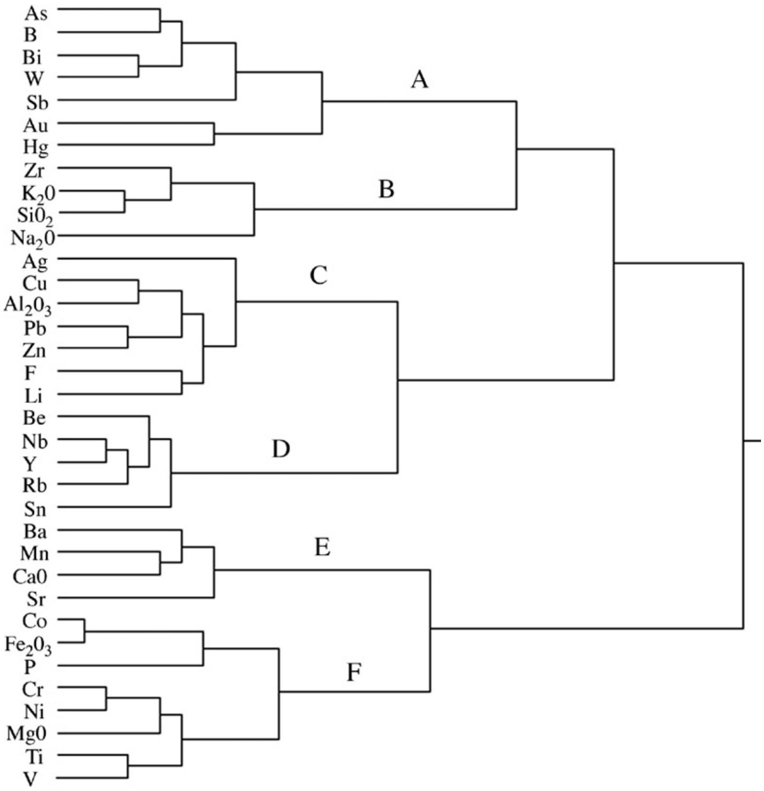

Fig. 3. Dendrogram of 35 variables from Tahe area in Heilongjiang province.

are related to the acidic or alkaline rocks, group E is characterized by the components of sedimentary rocks, and group F has the characteristics of basic components. The elements in group C can be greatly enriched as a result of volcanic activities (Liu et al., 1984).

The 2727 samples are classified into 6 types, corresponding to the 6 element groups, and all of them can also be reasonably interpreted by the practical geological case.

4.3. Recognition of the geochemical units

Samples belonging to the six classification groups are marked with different symbols (Fig. 4), the resulting geochemical units are approximately delineated in the regional geochemical image. In order to illustrate the practical significance of the image, the Bouguer gravitational anomaly map is given in Fig. 5. With these maps, the comparability and differences between the geochemical, gravitational and geological information can be seen. It is the comparability that proves the validity of the methods; it is the differences that show some problems and offer subjects for more detailed geological study.

The type C units indicate the main outline of the Mesozoic volcanics. In the southwest part of the area, the shape of these units is very similar to that of the volcanics. In the northwest part, the scope of these units includes and, in some places, exceeds that of the known volcanics, it may be attributable to the presence of unknown volcanics. Such a corresponding relation is indirectly testified by the gravitational information, because the type C units are consistent with the lower gravitational anomalies which represent the characteristics of the volcanics in neighboring area (Yang et al., 2002).

The type E units are related with Mesozoic sedimentary rocks. They are mainly distributed in the northeast part of the area, where more sedimentary rocks are exposed according to the geological mapping.

The type D units are mainly related to two kinds of acidic or alkaline rocks: One is the Paleozoic granites, beside the north bank of Huma Rive, the borderlines of the type D unit approximates that of the granites. Another kind is a certain type of volcanics, for example, in southeast part of Malin, where the shape of a type D unit, distributed along the north–west direction, is consistent with that of the volcanics.

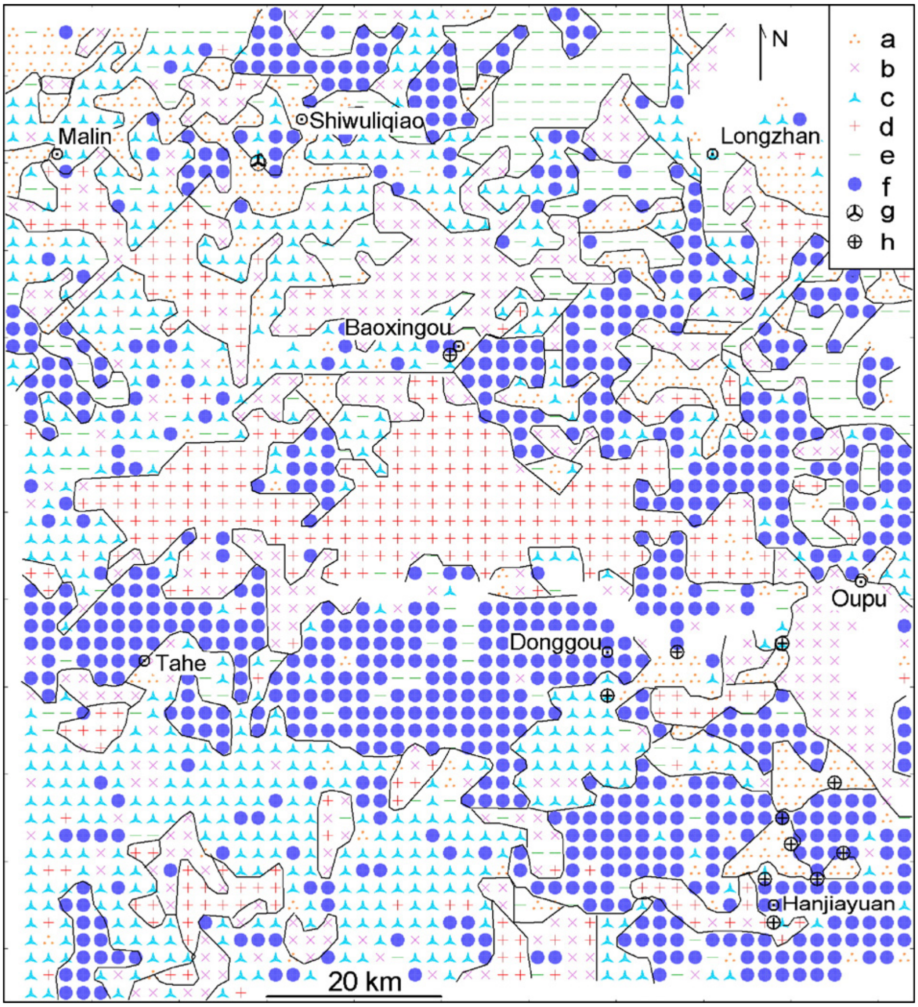

Fig. 4. Regional geochemical image of Tahe area in Heilongjiang province. a: Au, As, Sb, Bi, Hg, W and B group; b: SiO_2, K_2O, Na_2O and Zr group; c: Al_2O_3, Li, F, Cu, Pb, Zn and Ag group; d: Be, Nb, Sn, Rb and Y group; e: CaO, Sr, Mn and Ba group; f: Cr, Co, Ni, V, Ti, Fe_2O_3, MgO and P group; g: polymetallic ore occurrence; h: gold ore occurrences.

The type B units represent an evident geological or geochemical zoning in Paleozoic granites. For example, beside the north bank of Huma River they are all located at the edge of type D units or the granites. It may also represent certain type of volcanics or sedimentary rocks, for example, a type B unit located at north of Baoxingou.

Most of the type F units are related to Early Proterozoic rocks, including the magmatic rocks and metamorphic rocks. The distribution of this type of unit, especially in the southeast part of the area, is very similar to that of the Proterozoic rocks. It is interesting that both the type F units and Proterozoic rocks in this district are all consistent with the higher gravitational anomalies, and the corresponding relation between the Proterozoic rocks and higher gravitational anomalies is also supported by studies in neighboring area (Yang et al., 2002). Otherwise, part of type F units distributed in the Mesozoic erathem may represent basic volcanics or sedimentary rocks.

Some problems in the geological mapping are revealed and some significant inferences are given by comparing these maps. For example, between Longzhan and Oupu, a continuous Proterozoic geological body along south–north direction is delineated in geological map which follows the shape of a higher gravitational anomaly, but the type F units over there are not so continuous. Thus, the Proterozoic rocks implied by gravitational anomaly may be partly covered, but are not so continuous as they are shown in the geological map. This inference is validated by primary field examination.

Near Tahe the granites are geologically delineated in this district, but the larger type F units over there imply that the rocks exposed in this district should possess abundant

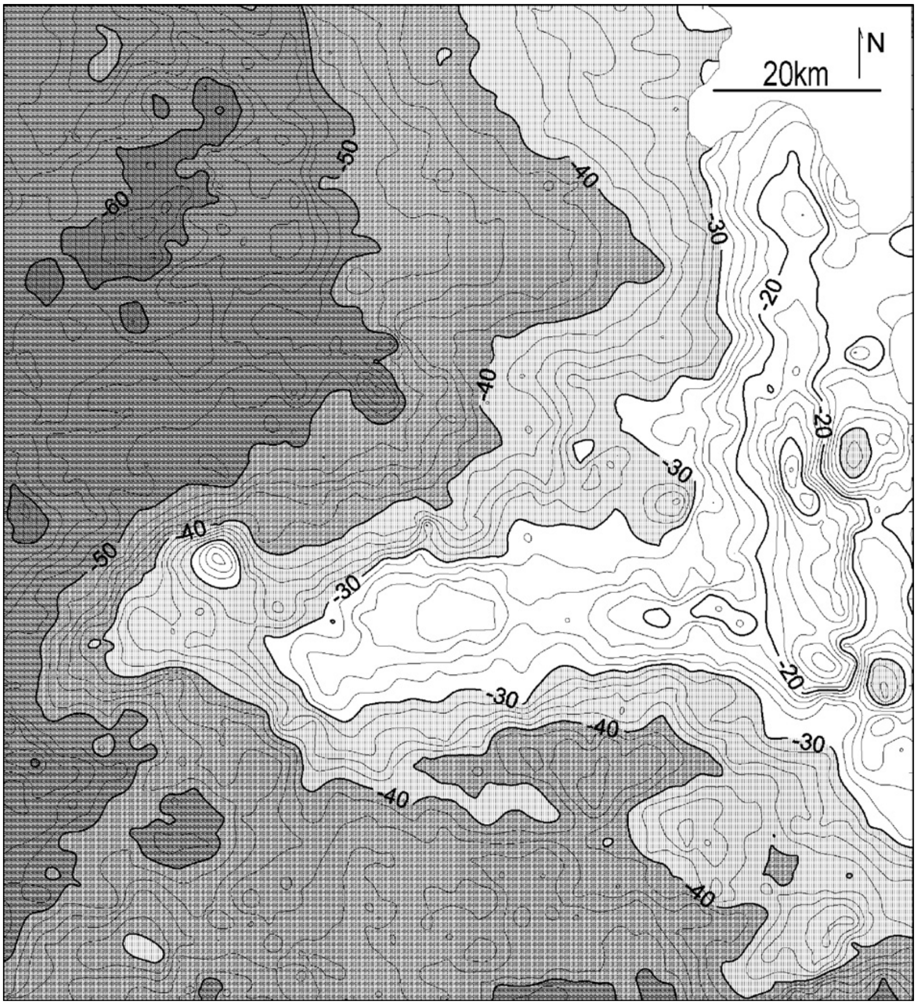

Fig. 5. Bouger gravitational anomaly map of Tahe area in Heilongjiang province.

basic components. Hence, the “granites” over there may not be true granites. Field examination showed that the main rock exposed in this district is actually granodiorite. Many similar examples occur in other districts, and they offer scope for revision of the geological mapping.

4.4. Recognition of the geochemical fault traces

The geochemical fault traces, which can be regarded as a kind of inferential geological fault, are clearly shown by the regional geochemical image. Especially those followed by (close to or bordering) type A units are closely related with the mineralization.

Between Malin and Shiwuliqiao, two groups of the fault traces followed by type A units are distributed along the north–east and north–west directions respectively. It is over there that a polymetallic ore occurrence appears. It indicates that the polymetallic mineralization may be controlled by the fault structures and related to hydrothermal activities.

Near the Baoxingou gold ore occurrence, following a type A unit (only one sample), three fault traces along different directions converge, where the fault trace along north–west direction is consistent with a fault marked in geological map. Around Donggou and Hanjiayuan, the fault traces followed by type A units cut the type F units, and 10 gold ore occurrences closely appear over there.

A large geochemical structural trace related to gold mineralization is an approximate triangle inhabited by the dense type F units, which covered about 5000 km^2 and is intruded by other types of unit. Its three vertices are located near Tahe, Longzhan and Hanjiayuan respectively. Its borderlines (expect for the east borderline, which is limited by the study area) show the characteristics of the fault traces. It is interesting that the higher gravitational anomalies also appear to be an equal triangle, and that all of the gold ore occurrences are located inside or beside the triangle. So, this triangular area, characterized by abundant basic components, higher gravitational value and numerous gold ore occurrences, is very important for geology and mineral exploration, and it may belong to a certain geochemical block (Xie et al., 2004).

4.5. Recognition of the ore-forming conditions and prospecting targets

The geochemical ore-forming conditions of the known ore occurrences can be inferred from the regional geochemical image. All of the known gold ore occurrences meet three necessary conditions, i.e. they are all synchronously located inside or beside: (1) the triangular geochemical block; (2) the type F units; and (3) the geochemical fault traces followed by type A units. The condition (2) may reflect the relation between gold ore occurrences and Eearly Proterozoic basement rocks that has been recognized geologically (Biao et al., 1999). Conditions (1) and (3) express relations between gold ore occurrence and the geochemical information that are hard to find with a macroscopical geological approach.

The particular significance of the geochemical conditions is shown by the Baoxingou ore occurrence, which is geologically located in Mesozoic rocks and hence different from the other gold ore occurrences which are inside or beside the Early Proterozoic rocks. But, just as the others, it is controlled by the same triangular geochemical block, i.e. all of the gold ore occurrences have the same geochemical conditions. Wang (1999) has suggested that these geochemical conditions may be the “gene” controlling the deposits, just as organisms are controlled by DNA.

Similarly, the ore-forming conditions of the polymetallic ore occurrence near Shiwuliqiao are also shown by the geochemical image, it mainly related with types C, A units and the geochemical fault traces.

According to these ore-forming conditions of known ore occurrences, it is not difficult to find previously unknown prospecting targets from the geochemical image. The main gold targets occur in the northeast part of the triangular geochemical block, and the main polymetallic targets occur between Malin and Shiwuliqiao, in the northwest part of the map.

5. Discussion

The first question that should be discussed is the success of the results obtained with SHCCA. Our experience shows that the usefulness of the results depends on obtaining a reasonable clustering of the variables, just as Fig. 3. But sometimes unreasonable element groups can appear in the results. In such a case, the results can be improved by: selecting a more appropriate method to link up a dendrogram (Mather, 1976); resisting the influences of outliers, or eliminating a few variables that are hard to interpret (Howarth and Sinding-Larsen, 1983; Ji and Chen, 1993; Tao and Xia, 2003).

The second question is what are the limitations of application of the SHCCA to geochemical pattern recognition. The sufficient recognition of different types of geochemical pattern depends on having sufficient indicator elements. The number of recognizable types would be limited if the useful elements are limited. Otherwise, just as the other multivariate methods, the results of SHCCA can only be applied to the study area itself, but cannot be compared with other areas. The method, however, should be very useful in such a case when the internal geochemical patterns of a certain district are emphasized.

The third question is related to the application of geochemical pattern recognition in mining districts. Actually, SHCCA has been successfully applied to many such districts since the 1990s, but only one example, from Jishan gold mine in Shandong province, has been published (Ji et al., 2005) and it had no introduction to the SHCCA method. In that example, 4 types of geochemical unit were recognized using 11 elements, a typical geochemical zoning of these types, from center to periphery on the ground of the deposit, was found to be: type of Co, Ni and Mo – type of Au, Ag, Cu and Pb – type of Hg, As and Sb – type of Ba, and both the known industrial ore bodies and hidden targets were effectively indicated by this pattern (Ji et al., 2005).

The fourth question is related to the peculiarity of the methods. The distinct characteristics of the method developed in this paper is to accomplish the main tasks of data interpretation with a simple multivariate classifying processes, and to show all of the interpreting results with a single diagram, i.e. the regional geochemical image. Actually, the essential task performed by most of the usual methods, including univariate methods to recognize the geochemical anomalies from background, is also classification (Ji et al., 2005), but complex steps and too many maps are needed in usual methods. Otherwise, the limitation of the univariate methods, which has been pointed out by many authors (Howarth and Sinding-Larsen, 1983; Lindqvist et al., 1987), has been improved in SHCCA. Therefore, the SHCCA can be regarded as a development of the traditional approach.

The last question is related to the exploitation of the data resources obtained from regional geochemical exploration. The example in the Tahe area shows that abundant small-scale information, which is hard to find with usual methods, is contained in the regional geochemical data. Geochemical exploration will play a more important part if all of the information, other than just detecting geochemical anomalies, can be fully exploited by advisable methods.

6. Conclusions

The semi-hierarchical correspondence cluster analysis, developed in this paper, is a new contribution for multivariate statistical analysis. It regroups the main advantages of correspondence analysis, hierarchical and non-hierarchical cluster analysis, and unifies the R- and Q-mode cluster analysis of the large data set.

A pattern recognition method, based on semi-hierarchical correspondence cluster analysis, can be selected as a data interpretation program in regional geochemical exploration. It carries out the main tasks of geochemical data interpretation with simple steps, and shows more important geological and geochemical information in only one diagram.

The abundant information resources related to geology and mineralization are hidden in regional geochemical data, and a new way to exhume this type of resource is offered by semi-hierarchical correspondence cluster analysis.

Acknowledgements

We would like to thank Professor Xia Lixian for his kindly attention to this research. Professors Sun Fengyue and Zhu Yongzheng are thanked for their cordial cooperation to study this method in many programs during the last decades. Professor Yan Guangsheng is thanked for his warmhearted support to apply this method to the Tahe area. Professors Wang Xueqiu, Liu Dawen, Hao Libo, Lu Jilong and Engineer Chen Man are thanked for their important help to provide the related information. Professors Richard Howarth and Rudy Swennen are thanked for their very patient corrections and valuable suggestions to this paper. We also thank the Heilongjiang Geology Survey which has provided the original geological, geochemical and gravitational data needed in this research.

References

Benzécri, J.P., 1973. L'Analyse des Données. Tome 1: La Taxonomie; 2: L'Analyse des Correspondence. Dunod, Paris. 619 pp.

Biao, S., Li, Y., He, X., Zhou, X., Ma, L., 1999. The geochemical characteristics of the Xinghuadukou Group in Lulin Forestry Center, Tahe, Heilongjiang Province. Reg. Geol. China 18 (1), 28–33 (in Chinese, with English abstract).

Castillo-Munoz, R., Howarth, R.J., 1976. Application of the empirical discriminant function to regional geochemical data from the United Kingdom. Geol. Soc. Amer. Bull. 87, 1567–1581.

David, M., Campiglio, C., Darling, R., 1974. Progress in R- and Q-mode analysis: correspondence analysis and its application to the study of geological processes. Can. J. Earth Sci. 11, 131–146.

Dong, W., Zhou, G., Xia, L., 1979. Theory of Quantification and its Application. Jilin People's Publishing House, Changchun. 197 pp. (in Chinese).

Eddy, W.F., Mockus, A., Ouc, S., 1996. Approximate single linkage cluster analysis of large data sets in high-dimensional spaces. Comput. Stat. Data Anal. 23, 29–43.

Ge, X., Wu, X., 1995. A primary study on relation of regional geochemistry to endemic disease in Heilongjiang province. J. Changchun Univ. Earth Sci. 25 (1), 63–69 (in Chinese, with English abstract).

Gustavsson, N., Bjorklund, A., 1976. Lithological classification of tills by discriminant analysis. J. Geochem. Explor. 5, 393–395.

Hanesch, M., Scholger, R., Dekkers, M.J., 2001. The application of fuzzy *c*-means cluster analysis and non-linear mapping to a soil data set for the detection of polluted sites. Phys. Chem. Earth, Part A Solid Earth Geod. 26 (11–12), 885–891.

Hong, F., Jia, Z., Long, Z., 1998. Correspondence cluster analysis and the decision of state investment fields. J. Southwest Inst. Ethn. Groups (Philos. Soc. Sci.) 19 (2), 132–137 (in Chinese).

Howarth, R.J., 1973. The pattern recognition problem in applied geochemistry. In: Jones, M.J. (Ed.), Geochemical Exploration 1972. Institution of Mining and Metallurgy, London, pp. 259–273.

Howarth, R.J., Sinding-Larsen, R., 1983. Multivariate analysis. In: Howarth, R.J. (Ed.), Statistics and Data Analysis in Geochemical Prospecting. Handbook of Exploration Geochemistry, vol. 2. Elsevier, Amsterdam, pp. 207–291.

Ji, H., Zhong, C., 1991. New graphic procedures in regional geochemical exploration. J. Comput. Tech. Geophys. Geochem. Explor. 13 (2), 98–104 (in Chinese, with English abstract).

Ji, H., Chen, Y., 1993. Correspondence cluster analysis for qualitative data and its application. J. Comput. Tech. Geophys. Geochem. Explor. 15 (4), 300–306 (in Chinese, with English abstract).

Ji, H., Zhu, Y., Wu, X., 1995a. Correspondence cluster analysis and its application in exploration geochemistry. J. Geochem. Explor. 55, 137–144.

Ji, H., Li, C., Wang, L., Wang, A., 1995b. The application of correspondence cluster analysis to the teaching research. Appl. Stat. Manag. 5, 10–13 (in Chinese).

Ji, H., Yu, Y., Shi, Y., Qi, Y., Chao, L., 2004. An application of correspondence cluster analysis in study of element medicine. Chin. J. Health Stat. 21 (4), 248–249 (in Chinese).

Ji, H., Sun, F., Chen, M., Hu, D., Shi, Y., Pan, X., 2005. Geochemical evaluation for uncovered gold-bearing structures in Jiaodong area. J. Jilin Univ. (Earth Sci. Ed.) 35 (3), 308–312 (in Chinese, with English abstract).

Jia, Z., Hong, F., 2000. Application of correspondence cluster analysis to decide the investment for railway construction. J. Southwest Univ. Natl. (Philos. Soc. Sci.) 21 (1), 11–14 (in Chinese).

Lindqvist, L., Lundholm, I., Nisca, D., Esbensen, K., Wold, S., 1987. Multivariate geochemical modelling and integration with petrophysical data. J. Geochem. Explor. 29, 279–294.

Liu, Y., Cao, L., Li, Z., Wang, H., Chu, T., Zhang, J., 1984. Elements Geochemistry. Science Press, Beijing. 480 pp. (in Chinese).

Mather, P.M., 1976. Computational Methods of Multivariate Analysis in Physical Geography. John Wiley & Sons, London. 532 pp.

Rantitsch, G., 2000. Application of fuzzy clusters to quantify lithological background concentrations in stream-sediment geochemistry. J. Geochem. Explor. 71, 73–82.

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

Shi, M., Carr, J.R., 2001. A modified code for R-mode correspondence analysis of large-scale problems. Comput. Geosci. 27, 139–146.

Tao, F., Xia, L., 2003. Stepwise correspondence analysis and study of physical characteristics of Chinese Ethnic Groups. J. Biomathematics 18 (2), 139–145.

Treiger, B., Bondarenko, I., Malderen, H.V., Grieken, R.V., 1995. Elucidating the composition of atmospheric aerosols through the combined hierarchical, non-hierarchical and fuzzy clustering of large electron probe microanalysis data sets. Anal. Chim. Acta 317, 33–51.

Velthuizen, R.P., Hall, L.O., Clarke, L.P., Silbiger, M.L., 1997. An investigation of mountain method clustering for large data sets. Pattern Recogn. 30 (7), 1121–1135.

Wang, X., 1999. Exploration geochemistry of giant ore deposit. In: Xie, X., Shao, Y., Wang, X. (Eds.), Geochemistry Tend Towards 21 Century. Geological Press, Beijing, pp. 35–47 (in Chinese).

Wei, Chih-Ping, Lee, Yen-Hsien, Hsu, Che-Ming, 2003. Empirical comparison of fast partitioning-based clustering algorithms for large data sets. Exper Systems with Applications 24, 351–363.

Xie, X., 1979. Regional Geochemical Prospecting. Geological Press, Beijing. 192 pp. (in Chinese).

Xie, X., Liu, D., Xiang, Y., Yan, G., Lian, C., 2004. Geochemical blocks for predicting large ore deposits-concept and methodology. J. Geochem. Explor. 84, 77–91.

Yang, Y., Li, Y., Fan, J., 2002. Metallogenic system in gold-silver-copper deposits, Tayuan district, Heilongjiang province. J. Jilin Univ. (Earth Science Edition) 32 (3), 229–232 (in Chinese, with English abstract).