Received September 25, 2019, accepted October 15, 2019, date of publication October 21, 2019, date of current version March 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2948800

Combining Graph Clustering and Quantitative Association Rules for Knowledge Discovery in Geochemical Data Problem

YASMINA MEDJADBA[1], DAN HU[1,2], WEI LIU[1], AND XIANCHUAN YU[1], (Senior Member, IEEE)

[1]College of Information Science and Technology, Beijing Normal University, Beijing 100875, China
[2]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Corresponding author: Yasmina Medjadba (yasmine.med1@yahoo.com)

This study was supported in part by the National Natural Science Foundation of China under Grant 41672323, in part by the Natural Science Foundation of Beijing under Grant L172029, and in part by the Public Welfare Industry Research of Natural Resources Ministry of China under Grant 201511079.

ABSTRACT Identifying geochemical patterns from backgrounds and generating associated mineralization remains challenging due to the complex structure of mineral deposits. To learn how to identify geochemical anomalies that are spatially associated with mineralization, we need in-depth knowledge of the dependence process. Quantitative association rules (QARs) are applied to discover remarkable relations and dependencies between attributes in a dataset, but it is difficult to generate relationships from geochemical data. In previous studies, no methodology to find association rules is proposed to deal with geochemical data problem, and the classical methods designed for Boolean and nominal attributes require previous discretization, which makes the whole process limited in processing complex data. In this paper, we proposed a hybrid method of graph clustering and quantitative association rules (GCQAR) as a new way of identifying significant geochemical patterns. Graph Clustering (GC) is used as partitioning paradigm because of its ability to handle large-scale datasets. The GC is based on modularity to effectively generate the groups of the graph, to avoid the over-partitioning, and to cover all the rules. In each partition, a set of geochemical quantitative association rules is produced. The results obtained in the experimental study performed on data collected in the field of Xiaoshan, Henan province, China. Our GCQAR has significant benefits in terms of recognition geochemical patterns compared to the traditional methods used in the field of geochemistry.

INDEX TERMS Recognition geochemical pattern, quantitative association rules, graph clustering, modularity.

I. INTRODUCTION

In recent decades, research on processing and recognition of geochemical anomalies that can be used in mineral exploration has made important progress. It is essential to look for the anomalies associated with mineral deposits [1], called significant anomalies. The anomalies are often interpreted as a basic sign of mineralization [1]. Besides, the distribution of geochemical elements is heterogeneous, and usually occurs at different temporal/spatial scales, and interconnects in various ways. Computational methods are necessary to extract knowledge from geochemical elements [2] that could help to identify hidden geochemical patterns related to mineralization [1]. Association rule is a machine learning method, and one of the most frequently used approaches to find relationships between different attributes in a database. It was first introduced in 1993 by Agrawal *et al.* [3], and the main target was to discover frequent patterns [4]. Thereafter, a large number of studies have been proposed to find quantitative association rules (QARs) [5]–[9].

Discovering frequent patterns plays a fundamental role to produce interesting relationships among quantitative data. Once the frequent patterns have been found, it is simple to generate association rules that satisfy both minimum support and minimum confidence [10].

The QARs are grouped into various categories [11] according to their computational techniques [12]–[16]. Commonly used methods are clustering-based approach [15]. Many of these clusters apply a domain partition technique and focus on logical interval generation using the notion of dense regions.

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

The difficulty of these methods lies in reaching optimal partitioning and might give rise to information loss. In addition, clustering methods are not all scalable for high dimensional cases and particularly considering that data can be highly skewed and very sparse.

A basic issue of the traditional association rules is to find frequent patterns in a database, this turns out to be even more problematic in geochemical data problem, due to the compositional nature [17], [18] of data, various dependencies exist, and the large-scale datasets that surpass the processing capability of the conventional system. In addition, geochemical exploration is based on the treatment of a huge number of variables from the relatively large area, and the elements in real-world are more or less associated in terms of certain relationships. Hence, traditional association rules have limitations in processing complex data. As far as we know, no previous research has investigated to identify geochemical pattern using association rules.

In order to properly address this issue, it is worthwhile to discover hidden structures from geochemical data to manage nonlinear and complex relationships before implementing quantitative association rules, because geochemical data usually coexist in heterogeneous geologic systems and connect with each other in difficult ways, so to identify appropriately significant anomalies. Furthermore, the geochemical anomalies generated are used as a direction of frequent patterns that lead to discovering significant patterns and the form of the rules. Geochemical patterns also have a sense of conditions for the rules, which would eliminate the discovery of certain redundant and uninteresting rules.

This work presents GCQAR, to discover significant patterns associated with mineral deposits from massive amounts of input data. The proposed method sequentially applies graph clustering and quantitative association rules. The geochemical anomalies identified are more meaningful in the context of mineralization, and had stronger spatial association with the known deposits in the study area.

This work is organized as follows: Section 2 introduces related works. The geochemical data and pre-processing are provided in Sections 3, 4. Section 5 provides details of our GCQAR method to generate quantitative association rules from geochemical data problem. In Section 6, experiment results and the comparison results with other approaches are provided. In Section 7, through experiments, we summarize the advantages and disadvantages of the GCQAR.

II. RELATED WORKS

Various approaches have been proposed to identify geochemical anomalies related to mineralization. Bölviken *et al.* [19] introduced the application of Fractal/multi-fractal models to quantify the spatial distribution of geochemical data. Later, a variety of fractal/multi-fractal models have been developed, such as the concentration-area (C-A) fractal model [20], [21], the spectrum-area (S-A) multifractal model [22], and the concentration-distance (C-D) fractal model [23], on the basis of scaling characteristics of geochemical data. Multivariate statistics such as principal component analysis (PCA) [24] and factor analysis (FA) [25], etc., are used to extract the multivariate geochemical data for mineral exploration. The previous methods are based on certain idealized assumptions, and their concern of only lower order, linear features makes them fail to support the complex nature of geochemical data.

A few works in literature are proposed to identify geochemical anomalies based on machine learning. In the research of supervised Learning, Abedi *et al.* [26] introduced support vector machine (SVM) to explore the Now Chun porphyry-Cu deposits, located in the Kerman province of Iran. Logistic regression (LR) [27], [28] is used to create a multivariate relationship between dependent (e.g., deposits or non-deposits) and independent variable (e.g., faults, geochemical anomaly) to estimate the probability of a specific event related to mineralization. Artificial Neural Networks [29]–[32] have shown advantages over many other methods in geochemical anomaly recognition. Chen *et al.* [30] employed a continuous restricted Boltzmann machine (CRAM) to recognize multivariate geochemical anomalies in the Baishan district in northeastern China. Hinton *et al.* [33] used a deep belief net (DBN) to identify multivariate geochemical anomalies. Carranza and Laborte [34] used random forest for data-driven modeling of mineral prospectivity with small number of prospects and data with missing values, in Abra (Philippines). A combination of m-branch smoothing, C4.5 decision tree and weights-of-evidence techniques was introduced by Chen *et al.* [35] for mineral prospectivity mapping. In the research of unsupervised learning, a deep autoencoder network was introduced by Xiong and Zuo [29] to encode and reconstruct a geochemical sample population with unknown complex multivariate probability distributions. Unsupervised clustering [36]–[40] mainly include k-means clustering [41], fuzzy c-means clustering [41], [42]. These clusters are implemented to describe the spatial distribution of data and define the locations of anomalies. Fouedjio [43] developed an agglomerative hierarchical clustering approach that considers the spatial dependency between observations. Self-organizing map (SOM) [44], [45] is used to identify relationships and patterns in multidimensional datasets. Although studies have been conducted by many authors, this problem is still insufficiently explored.

In other hand, the massive amount of data and applications have led to the development of numerous methods for generation of association relationships. In literature, most of the existing association rules are based on classical methods proposed by Agrawal, Imielinski and Swami such as Apriori [46]–[48], FP-Growth [49] and SETM [50]. These methods are designed to work perfectly with Boolean, nominal values and categorical. Apriori based on candidate creation, then investigation while other methods such as FP-Growth, tries to create a tree without candidate generation, and then finds the frequent items by scanning on the tree. Later, extensive studies were carried out to improve these methods and their

applications [46], [51]. However, these methods are based on the generation of a large number of rules suffering from a problem of choosing a threshold and take more database scan in order to calculate the frequency of itemset, which leads to an increase in execution time and memory overhead. Besides, the rules with numerical attributes cannot be discovered by these methods. Though the number of contributions that have been proposed to adapt these methods to deal with QARs, they all require previous discretization, where data are replaced by interval labels using data discretization or concept hierarchies. However, such simple discretization may lead to the generation of an enormous number of rules, most of which end up being unrelated or uninteresting. Even though minimum support thresholds help reduce the exploration of a good number of uninteresting rules, but several of them are still not interesting. Another large combination of strategies based on evolutionary algorithms (EA) [52], [53] that have been introduced to build a set of QARs. However, these methods require high implementation of knowledge exploration.

A new approach is therefore needed for the generation of frequent patterns from geochemical data problem. In this study, we have proposed a three-stage approach to this problem:

1) Implement graph clustering (GC) to generate clusters with significant frequent patterns (or geochemical patterns) from a complex background.
2) Obtain a set of QARs (RuleSet) from frequent patterns discovered in each cluster.
3) Evaluate the quality of the rules over the entire clusters with the aim of selecting the remarkable rules that present the best behavior between variables in the entire dataset.

III. STUDY AREA

The geological map of the study area is provided by the Institute of Geology and Mineral Resources and Development of Henan Bureau.

The investigation area is located in the southern margin of north China Platform and in the middle section of the Huaxiong Tailong Group (Fig. 1). It is an important metallogenic belt of the Yuxi Gold Mine. The stratigraphic zone of the investigation area belongs to the western Henan section of the north China stratigraphic zone, and is spanning the Xiong'er Mountain Community and the Dianchi-Cheng Mountain Community.

The study area has a typical double-layer structure. The first layer consists of the crystalline basement, which is the Taihua metamorphic complex group. The second layer is the caprock, which is distributed from the bottom to the top, namely, Lushan group, Xiong'er group, Guandaokou group, Fuyang group, Luojing group, Sinian, Cambrian, Cretaceous, Paleogene, Neogene, and Quaternary. The Taihua complex is exposed in the core of the Lushan fault, and is surrounded by the broad-angled Xiong'er group. The Quaternary system forms the loess area in the southeastern and northwestern fault basins, and the remaining strata are scattered. The metamorphic rocks in the inspection area are developed with an exposed area of $340km^2$, and constitute of the crystalline basement of the Lushan fault, which is an important gold-bearing geological body in the area. The metamorphic rocks are composed of six major types of rocks, including slightly metamorphic rock, amphibolites, quartzite, schist, felsic rock, and gneissic granite. Furthermore, the exposed area of the intrusive rocks is about $260km^2$, which more than 90% are Neoarchean metamorphic granitoid, and are concentrated in the crystalline basement of the Lushan faulted area. The gabbro, diabase, granite porphyry, indosinian syenite porphyry and late Yanshanian granite porphyry in the Mesoproterozoic bear period are scattered in the caprock zone. The middle Proterozoic Xiong'er volcanic rocks spread throughout the region, accounting for the total of the investigation area. The type of these rocks consist of volcanic lava and volcanic clastic rock. Lava rock can be divided into calc-alkaline series and alkali-calcium series according to alkalinity.

The fracture structures in the survey area are very developed and divided into four groups according to their distribution direction, namely: northeast (NE), northwest (NW), east-west (EW), and north-south (NS). There are more than 20 large-scale fractional zones, with the trend is about 60°, tend to northwest (NW), southeast (SE), and the inclination is generally between $60° \sim 80°$. The northwestward fracture is relatively developed and concentrated in belt production. A considerable part of this fracture is the ore-controlling structure of gold, silver, copper, lead, tungsten, barite and other minerals. The east-west fracture zones are generally of a huge scale, and there are many other normal fractures. Besides, two north-south fracture zones are developed in the area, Zhuyuangou-Yuwang fracture zone in the west and Dagugou-Taowangcun fracture zone in the east.

The Neoarchean granitic greenstone terrane and the middle Proterozoic Xiong'er group continental volcanic activity provide a source of gold for the group of gold deposits. The multi-stage tectonic-magmatic thermal events provide conditions for the group of gold deposits. In addition, copper lead, silver, tungsten polymetallic and non-metallic minerals based on barite have also formed a number of mineral deposits in the investigation area, so the metallogenic conditions are superior. The characteristics of geochemical elements in the study area are a comprehensive reflection of the geochemical fields in the Xiaoqinling gold ore field, the Xiong'er mountain gold and molybdenum polymetallic metallogenic belt. The distribution of the geochemical elements is uniform or uneven, while the majority of the elements are not highly differentiated. Only W is a strongly differentiated type.

The study area is divided into different regions, and two regions with different geological structures were selected for this study.

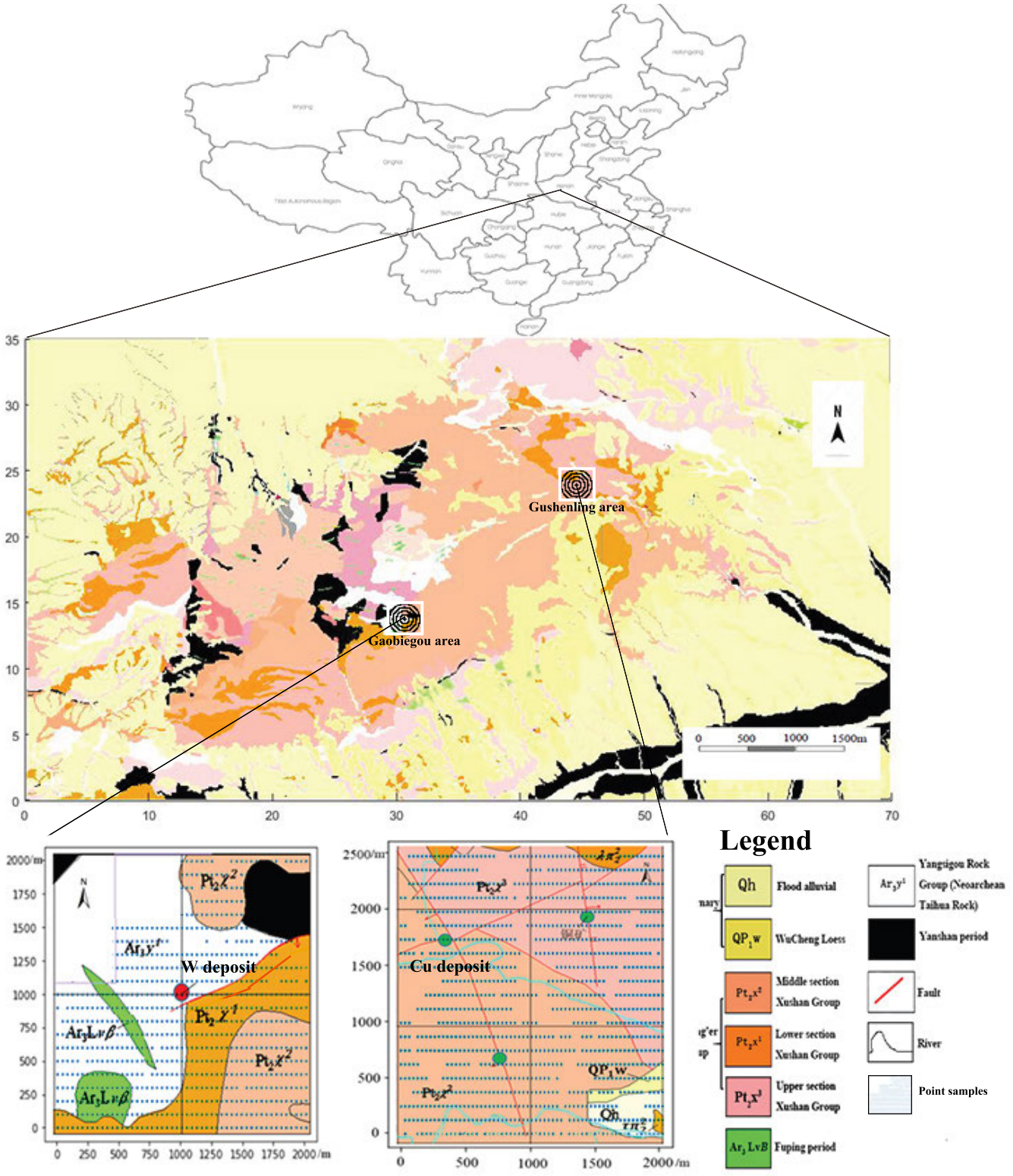


FIGURE 1. Geological and mineral ressources of Xiaoshan, Henan province, China (scale 1:10000).

Gaobeigou area is located in the north of Changshui Township, Luoning Province. The known deposits in this region are W-Ni-Zn-Mo-Au, with the presence of large and small W deposit in the middle of the area.

Gushanling area is located in Chenjiayuan Village and Dashitun Village in the south of Gongqian Township, Shaanxi Province, where the known deposit is Cu.

Within the study area, 12 elements were collected from soil sampling for further analysis, including Ag, As, Au, Bi,

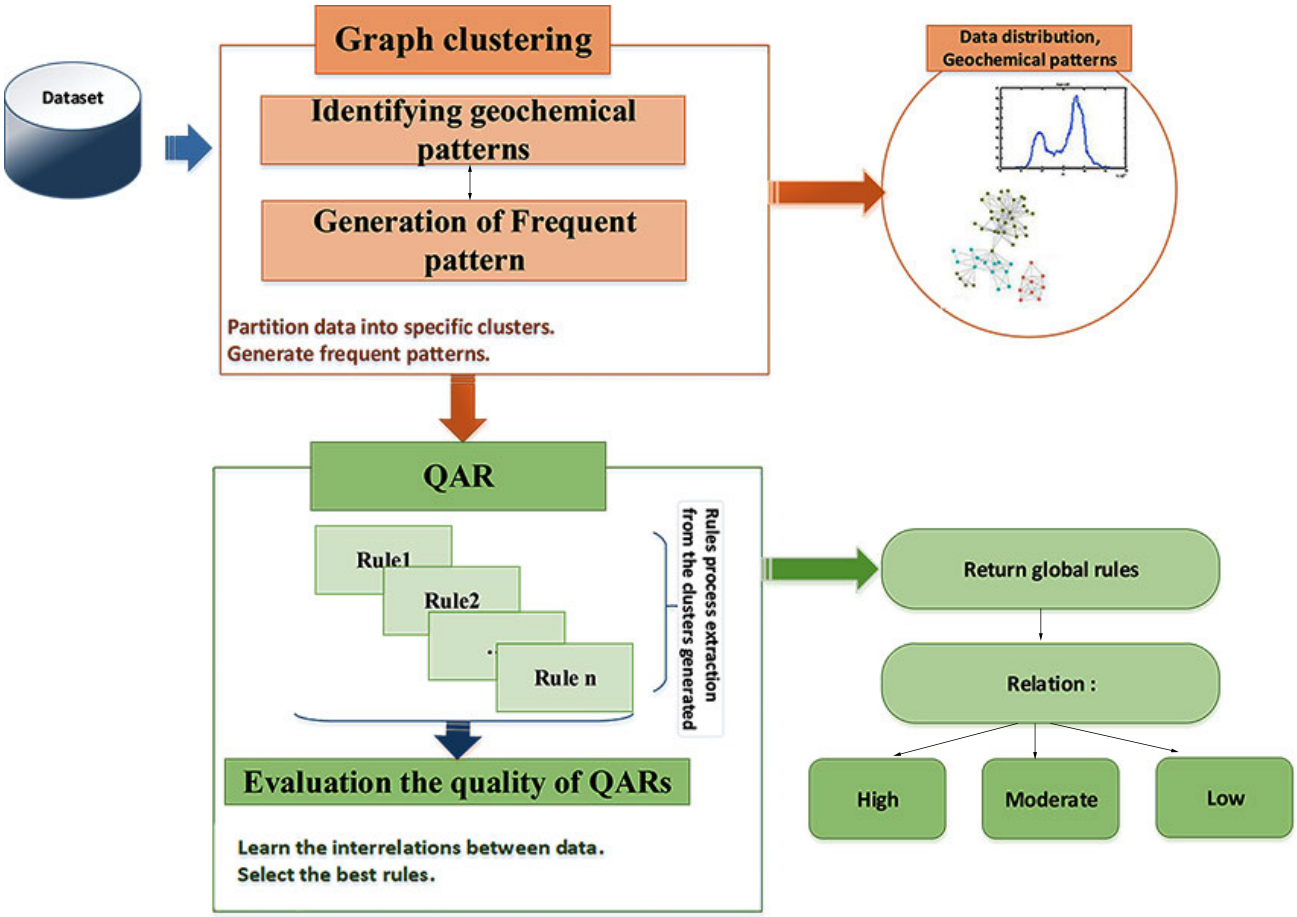


FIGURE 2. General scheme of GCQAR.

Co, Cu, Mo, Ni, Pb, Sb, W, Zn, with a total of 28270 sample points.

IV. DATA PREPROCESSING

Data preprocessing is often problem-dependent, and should be carefully employed since the input data significantly influence the results of many algorithms. It is suggested to prepare data in particular ways before implementing any methods. In addition, geochemical data listed as compositions and represented as vectors with a constant sum constraint, typically summing to 100%. This poses a difficulty when looking for statistical correlation in compositional data because values are relative, rather than absolute $n-1$ and can lead to spurious results [54], [55]. The log-ratio transformation [56] is a solution to the constraints of closed data. An isometric log-ratio transformation (ilr) [57]–[59] was employed to open the raw geochemical data prior to data analysis. The ilr transformation is presented as

$$ilr_i = \sqrt{\frac{rs}{r+s}} \log \frac{g(y_+)}{g(y_-)} \tag{1}$$

where $g(\cdot)$ is the geometric mean of the argument, y_+ is the group with r parts marked with $+1$ and y_- the group of s parts marked with -1.

After transformation, standardisation of feature values is required to provide relative measures of scale and a z-score [60] standardisation was selected for this purpose.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Here x is the transformed data, μ is the mean and σ is the standard deviation.

Later, to avoid high values of interesting measures [10] that lead to misleading results, all the inputs are transformed into [0,1] by using

$$X* = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

where $X*$ is the normalized value, X is the inputted value, X_{max} and X_{min} are the maximum and minimum values of X, respectively.

V. PROPOSED METHOD

A. GCQAR

The proposed method sequentially implements graph clustering and quantitative association rules to geochemical data problem; Fig. 2 describes the conceptual scheme. Graph clustering is first applied to identify geochemical data from complex background (Fig. 3). After GC method, detailed

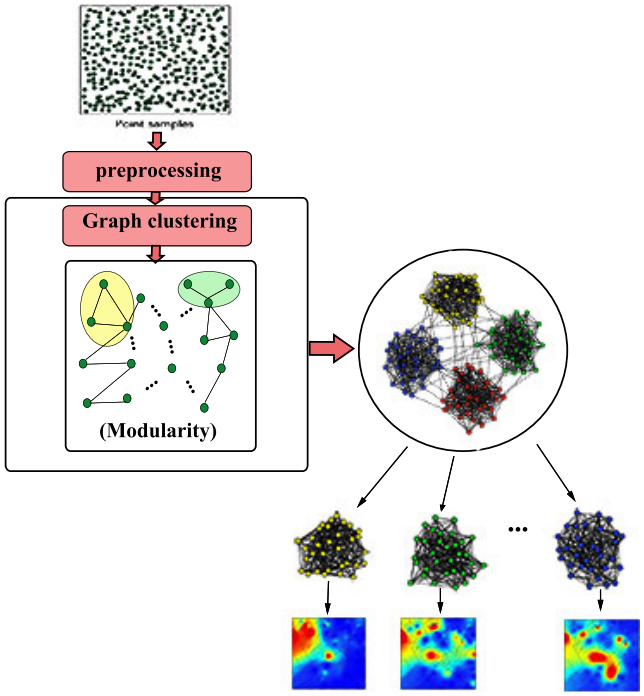


FIGURE 3. The framework of the Graph clustering, used to generate frequent pattern.

features in each cluster are examined based on the concept of quantitative association rules (Figs. 4-6), which allow the generation of unknown interrelations present in the clusters being studied. A result of this process can provide a useful coarse-grained representation of the data [61]. It can improve our understanding of the distribution of geochemical patterns and the interactions between the elements. Furthermore, it helps us to learn deeper structures of geochemical data and predict the future behavior of the elements. Details of the graph clustering and quantitative association rules used in this study are illustrated in the following subsections.

1) GRAPH CLUSTERING

Graph clustering [61], [62] is a field in cluster analysis that looks for groups of similar vertices (i.e., nodes) in a graph. Graph clustering represents data as vertices connected to one another by edges with a set of properties. It plays a basic role to model meaningful systems in different disciplines [68]. The ultimate goal of graph clustering is to partition vertices into several subgraphs, where the vertices are highly cohesive inside but sparsely to other subgraphs. There exist a number of approaches aim at discovering natural divisions of the graph, based on different measures of similarity. A more comprehensive description can be found in [63]–[67].

In the present study, we use modularity optimisation method [68], since it is suitable for handling large datasets. The groups can be quantified in terms of quality functions that give the best split.

Suppose geochemical dataset is a graph, contains n vertices. Each point sample is a node, and edges represent interactions among them. Given a sparse graph G(V,E) which consists of the node set V, the edge set E. The graph can be divided into two groups using a membership variable s. Let vertex v belongs to group 1 if $s_v = 1$ and $s_v = -1$ if it belongs to group 2, for a specific partition of the data into two groups. The number of edges between vertices v and u be M_{vu}, which will generally be 0 if there is no edge between vertices v and u or 1 if there is an edge between the two. The modularity Q

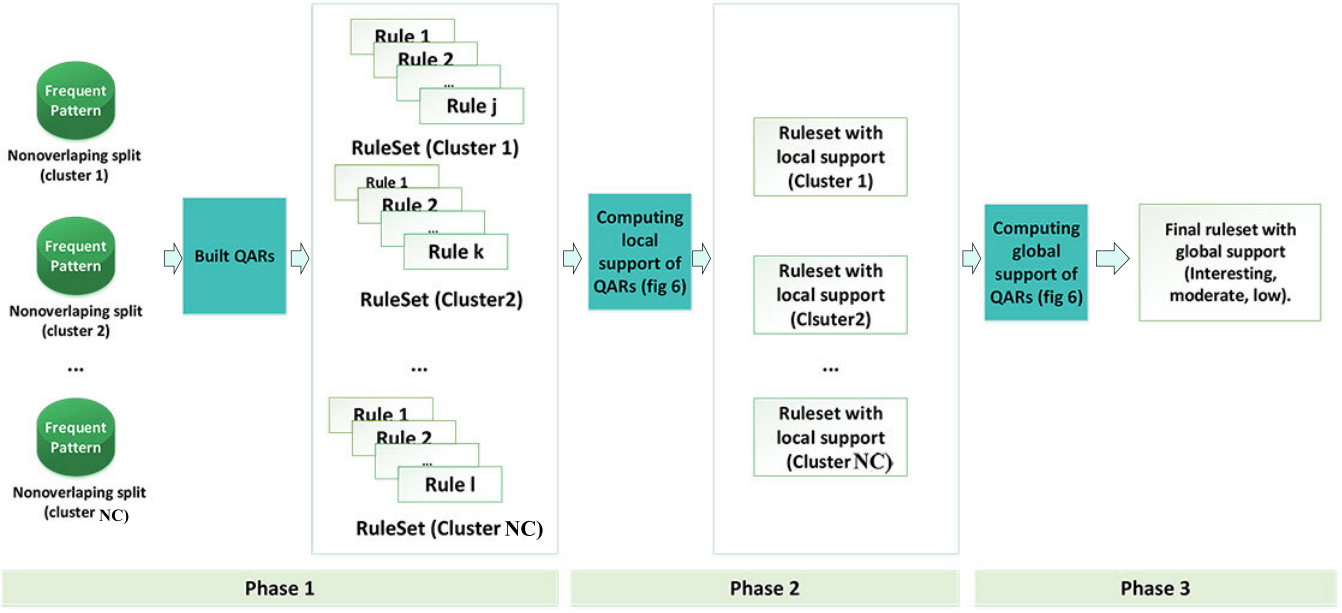


FIGURE 4. The process to mine QARs over the nonoverlapping splits.

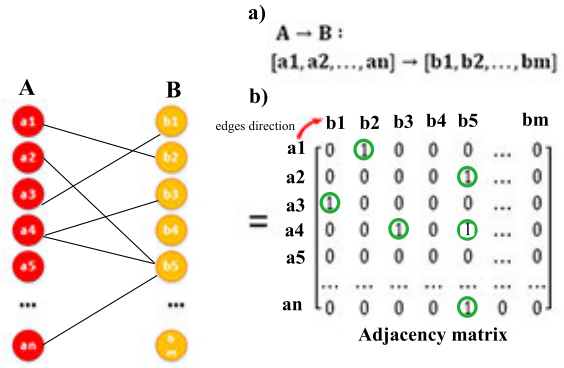


FIGURE 5. Adjacency matrix of a finite graph, the elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph.

is defined as

$$Q = \frac{1}{4m} \sum_{vu} (M_{vu} - \frac{k_v k_u}{2m}) s_v s_u \tag{4}$$

If edges are randomly placed between vertices v and u, then the expected number of edges is $\frac{k_v k_u}{2m}$, where k_v and k_u are the degrees of the vertices, $m = \frac{1}{2}\sum_v k_v$ is the total number of edges in the graph, where $2m = \sum_v k_v = \sum_{vu} M_{vu}$ and the modularity Q is given by the sum of $M_{vu} - \frac{k_v k_u}{2m}$ through all pairs of vertices v, u that fall in the same group. The whole procedure is repeated to subdivide the graph until every remaining subgraph is indivisible, and no further improvement in the modularity is possible. In this study, we focus on unweighted graphs.

The main process of Graph clustering algorithm used in this work is described as (Fig. 3):

Input: A graph G(V,E)

Require: unweighted graphs.

1) Each vertex belongs to a single group.
2) Consider each group pair, and assess the modularity score Q that could be achieved by joining them.
3) Join the two clusters that have positive, large values of the modularity (ΔQ) [68].
4) Repeat the steps 2 and 3 till only one group remains.
5) Return the splits that allowed obtaining the highest modularity score.

Output: The final partitions (disjoint modules).

In the initial work published in [68], it was described that the method was used to identify community compositions, and to reveal the structural features of networks. In the present study, we are specifically interested in the delineation of geochemical anomaly from complex background, and then the result obtained is used as frequent patterns.

GC method can find arbitrary shaped clusters, since geochemical data are not often spherical. Besides, we used modularity to identify disjoint groups that will generally lead to better results than the overlapping clusters. To keep particular features within the clusters for further analysis, and to avoid the generation of redundant rules [69].

2) QARs

Although the process of graph clustering creates groups in which geochemical patterns are brought into some degree of similarity in terms of the quality function known as modularity [68], the relation between the elements remains unclear. In addition, knowing the degree of association among the elements in the graph is also important to analyse their behaviors. In this section, our interest goes towards finding significant interrelation among nodes and explaining variations in

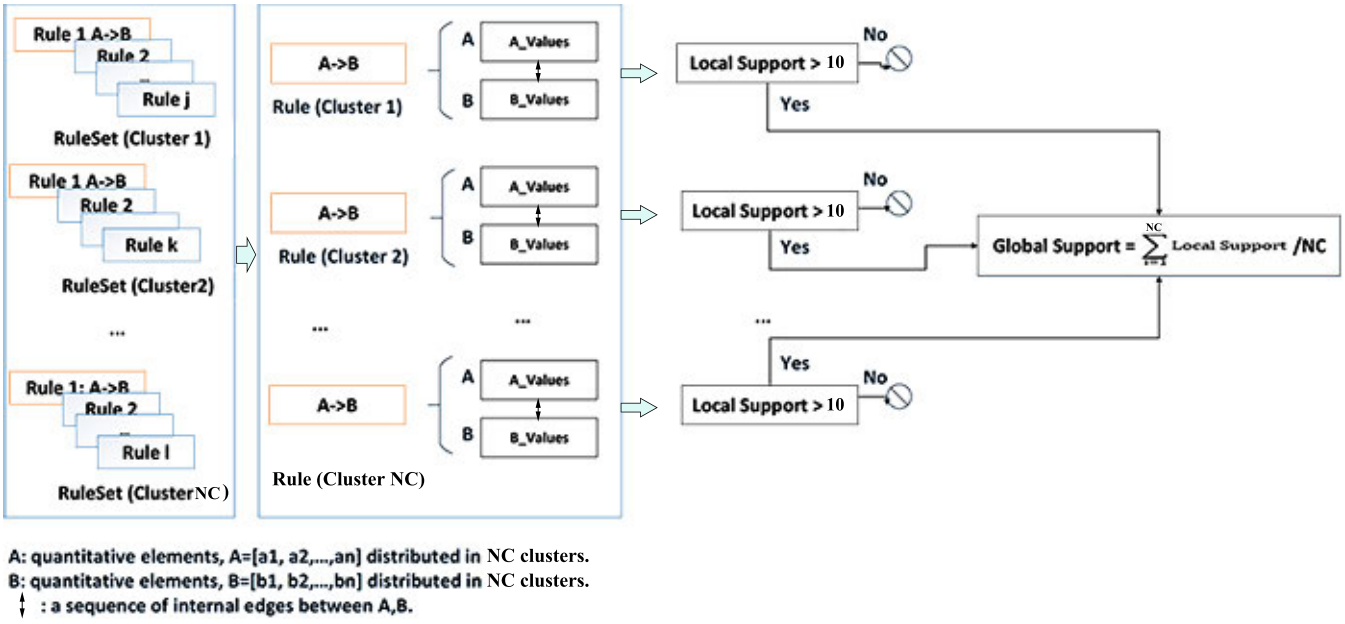


FIGURE 6. Flowchart of how the support of QARs is calculated.

TABLE 1. Summarized statistics of the chemical elements distributed in the soil sampling.

Lithology	Element/ppm	NS	Mean	Standard deviation	Skewness	Kurtosis	Minimum	Maximum
Gaobiegou area	Ag	896	0.08	0.07	16.07	286.19	0.04	1.50
	As		11.12	5.15	1.19	1.99	1.53	34.80
	Au/ppb		1.09	0.72	5.06	46.37	0.15	10.70
	Bi		0.28	0.07	-1.14	1.27	0.15	0.57
	Co		21.47	7.87	2.71	12.63	7.00	81.50
	Cu		23.46	13.06	3.98	24.02	3.57	155
	Mo		1.41	0.48	1.01	2.82	0.51	4.06
	Ni		29.49	8.38	0.22	0.77	5.42	65.40
	Pb		26.58	16.42	15.19	334.16	5.18	409.00
	Sb		0.90	0.57	2.79	23.15	0.15	7.05
	W		17.44	24.89	2.62	7.74	0.52	167.00
	Zn		65.98	22.65	10.73	207.11	33.60	535.00
Gushenling area	Ag	1136	0.07	0.01	1.08	6.56	0.04	0.17
	As		11.18	6.96	4.90	67.08	0.25	126.66
	Au/ppb		1.06	1.21	17.23	389.28	0.15	31.60
	Bi		0.30	0.17	3.92	25.70	0.15	1.88
	Co		23.75	6.92	1.14	3.74	5.60	67.50
	Cu		18.65	19.40	16.27	367.75	1.00	502.80
	Mo		1.10	0.38	1.30	4.78	0.50	3.98
	Ni		33.92	9.19	0.68	1.21	6.80	76.00
	Pb		21.22	9.85	3.69	29.48	2.50	120.60
	Sb		1.25	0.96	9.12	155.34	0.15	20.70
	W		2.66	1.28	2.54	11.03	0.50	13.11
	Zn		106.48	26.62	1.78	12.38	37.96	363.96

geochemical datasets, because understanding the interaction between elements through the obtained clusters, and exploring associated mineralization is worthwhile in geochemistry. The question now is how can we measure the interrelation between two given elements on a graph accordingly?

In order to address the question outlined here, we need to develop a new method to quantify the interrelation between the elements.

In this section, we introduce quantitative association rules to find useful information among the vertices. The QAR problem [5] is to identify all interesting rules of the form $A \rightarrow B$ where A is the antecedent and B is the consequent of the rule, $A, B \subseteq I$ and $A \cap B = \emptyset$. I represents itemset, A and B represent the set of items.

The learning phase of QARs used in this work consists of the following steps (Fig. 4):

- Obtain a set of QARs for each cluster, in which the input dataset is divided. The antecedent and consequent of the rules are arbitrarily selected. Besides, the length of the rules is always fixed to the number of nodes in each partition (Fig. 5(a)).
- Evaluate the quality of the rules over the entire splits, using the concept of support and confidence [70]. We focus on the following rule:
  If two elements are strongly related in the total splits, their relationship may lead to significant patterns (mineralization).
- Obtain the local support (L.Sup) of each rule in ruleset. The rule in each partition that does not satisfy a minimum threshold is removed (Fig. 6).
- Lately, the ruleset from each cluster is collected. Then, the local results generated (i.e., the local supports of ruleset) are merged to compute the global final result (global support of ruleset (G.Sup)) (Fig. 6).

However, the vertices lack additional attributes and there is nothing in the nodes themselves that allows the computation of a relationship. Besides, a path from one vertex to another one is a sequence of edges (Fig. 5(b)). Considering this information, we define local support (L.Sup) as the probability to find a sequence of internal edges *"e"* between each pair of vertices (i.e., elements) A and B in the same cluster.

$$L.Sup(A \rightarrow B) = \frac{\sum_{e \in E_l^e} e(A, B)}{1} \tag{5}$$

And the global support (G.Sup) represents the ratio of the number of internal edges between two elements to the number of clusters (NC).

$$G.Sup(A \rightarrow B) = \frac{\sum_{i=1}^{NC} L.Sup(A \rightarrow B)}{NC} \tag{6}$$

And confidence is defined as follows:

$$confidence(A \rightarrow B) = \frac{\sum_{e \in E_l^e} e(A,B)}{G.Sup(A)} \tag{7}$$

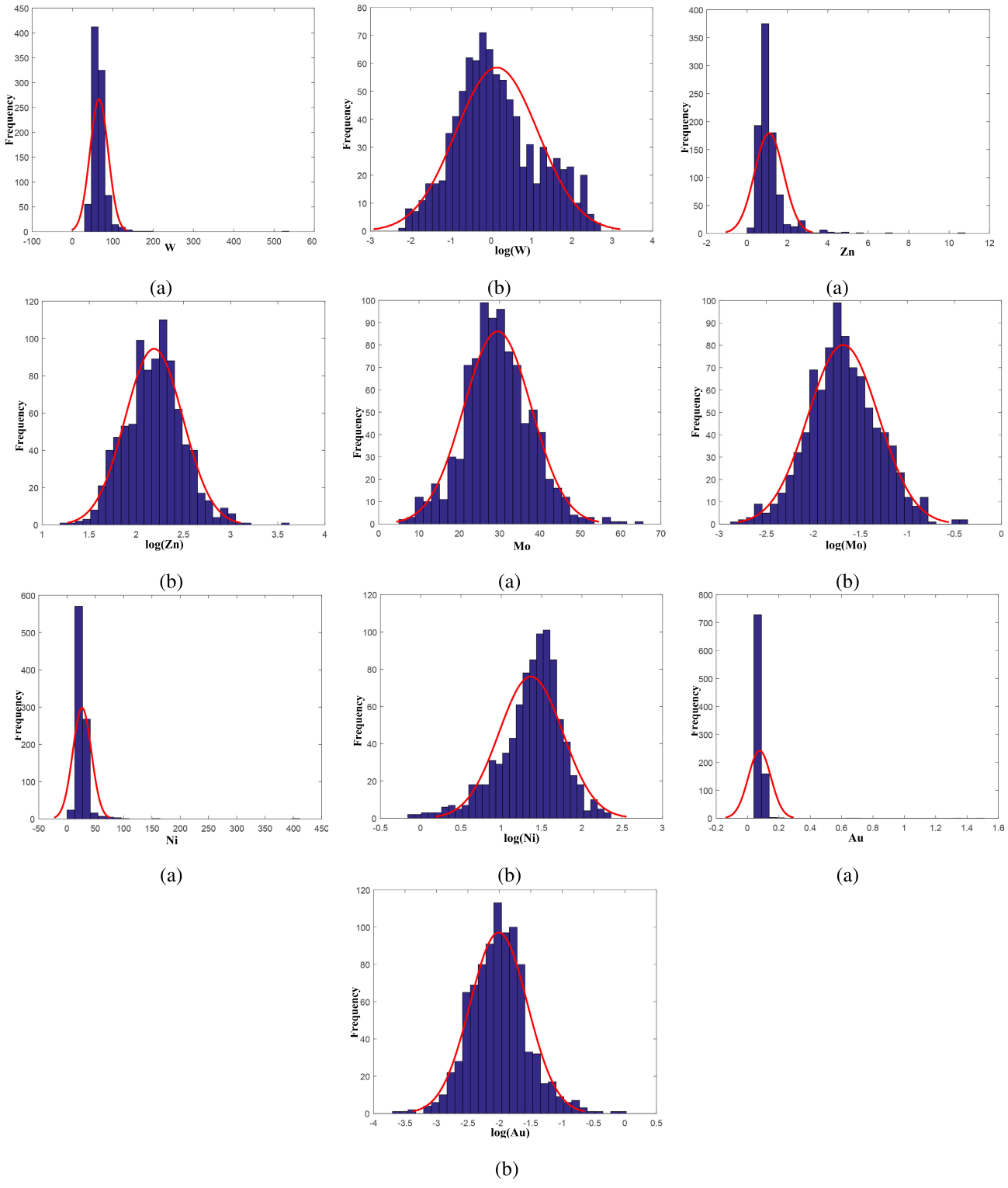


FIGURE 7. Histogram (a) for original data (b) transformed data, of Gaobiegou area.

where

$$G.Sup(A) = \frac{L.Sup(A)}{NC} \tag{8}$$

And

$$L.Sup(A) = \frac{|A|}{1} \tag{9}$$

where E_l^e contains the internal edges *"e"* belong to the lth cluster, and global support(A) is the ratio of the probability distribution of $|A|$ to the number of clusters (NC).

In other hand, in graph clustering the number of edges exceeds the number of nodes, thus to avoid high values of support and confidence, for each vertex, one edge is

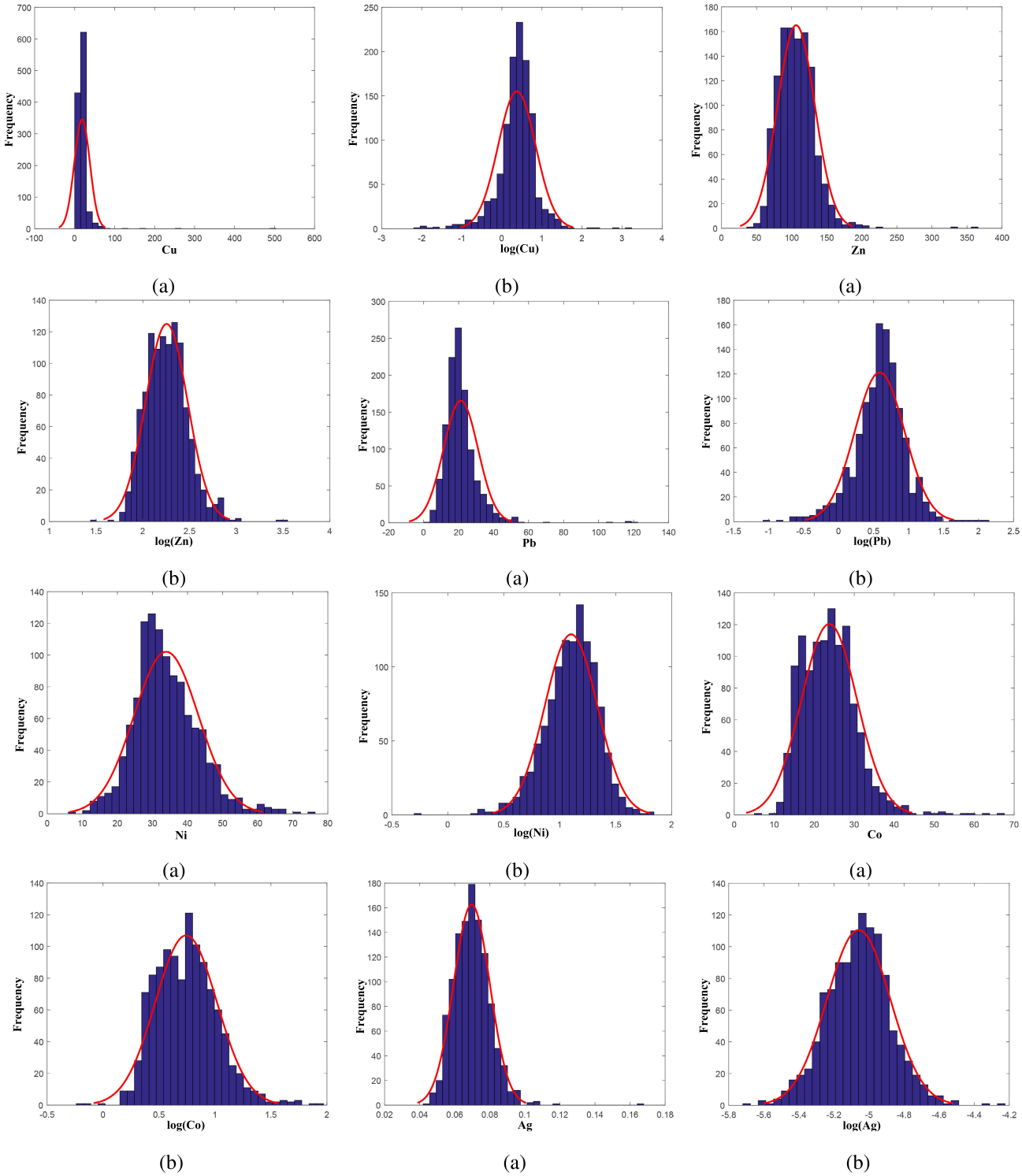


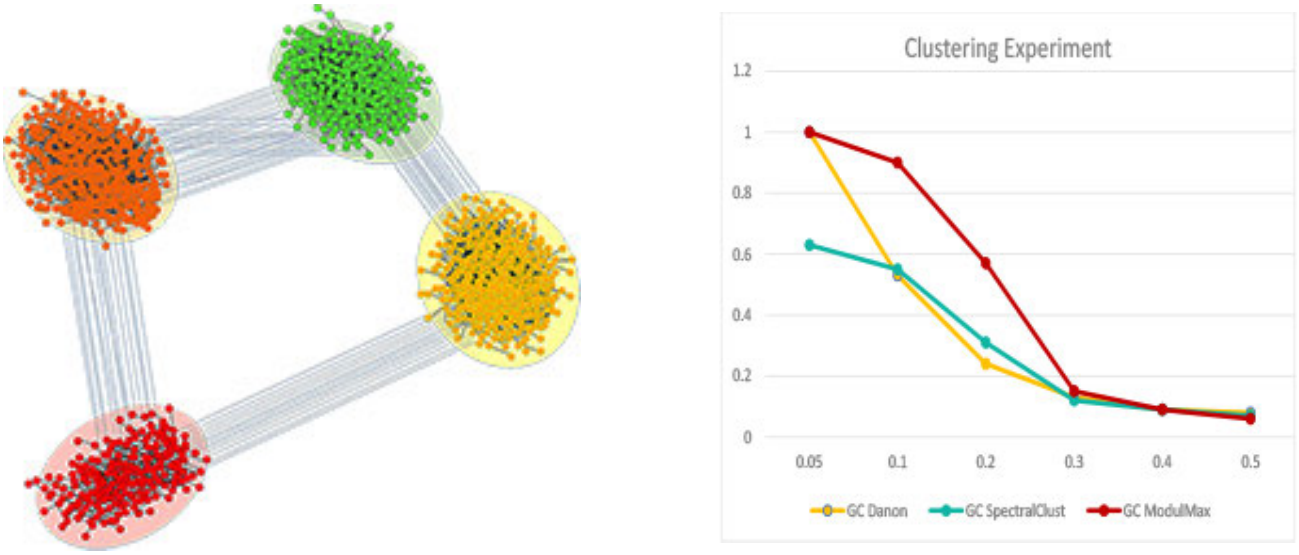
FIGURE 8. Histogram (a) for original data (b) transformed data, of Gushenling area.

calculated (the edge that starts from the antecedent of the rule (Fig. 5(b))), instead of considering all of them.

In addition, an edge between two given nodes can be defined with the adjacency matrix M, where its elements $M_{A,B} = 1$ when there is an edge from vertex A to vertex B, and $M_{A,B} = 0$ when there is no edge (Fig. 5(b)). In this paper, the edges from a vertex to itself (loops) are ignored.

The resulting QARs are presented as follows:

If the confidence is more than 50%, the relation is very significant and the edges between the two elements are effective.



(a) (b)

FIGURE 9. The visualization of (a) graph clustering generated by modularity maximization algorithm, and (b) partitions produced by danon algorithm, modularity maximization algorithm and spectral clustering for Gaobiegou area.

Algorithm 1 Support of the Ruleset

Input
Ruleset a set of rules discovered;
e(rule) $\in$ E a sequence of internal edges between the antecedent and the consequence of rule;
Require: Global support in NC $\simeq$ 1 $\simeq$ 100%;
Output: Local support of Ruleset;
Global support of Ruleset;
1. **For** each rule $\in$ Ruleset **do**
2. Compute local support of each rule (rule,e(rule)).
3. **end for**
4. **if** Local support of each rule $>$ 10 **then**
5. Compute global support of each rule.
6. **end if**

Algorithm 2 Confidence of the Ruleset

Input
Ruleset a set of rules discovered;
Local support of rule;
Local support of antecedent of rule;
Require: Confidence in NC $\simeq$ 1 $\simeq$ 100%;
1. **For** each rule $\in$ Ruleset **do**
2. Compute Confidence of rule.
3. **end for**

If the confidence is more than 39%, the relation is significant and the edges are mostly effective.

If the confidence is more than 10% the relation is low and the edges are ineffective.

VI. EXPERIMENTAL ANALYSIS

In our experiments, we implement GCQAR to regional geochemical pattern recognition for W-Zn-Mo-Ni-Au from 896 soil samples and Cu-Zn-Co-Pb-Ni-Ag from 1136 soil samples, of Gaobiegou and Gushenling area, respectively. In Xiaoshan, Henan province, China.

In this section, we will compare GC based on modularity optimization method [68] to the spectral partitioning that is used to generate overlapping groups (Luxburg) [71], Danon's greedy community detection agglomerative method (Martelot and Hankin) [72], and K-means (Serra and Tagliaferri) [73] that is widely used as partition method in geochemistry, so to demonstrate the features and operation of the proposed method for knowledge discovery in geochemistry.

The results were coded by lithology, using MAPGIS software package [74]. The experimental environments include an Intel Core i7-8550U 4.0-GHz CPU and 8 GB RAM.

A. STATISTICAL ANALYSIS

The statistical methods have performed in the description of the critical geochemical patterns [55]. The statistics have applied in as being descriptive such as mean, maximum, minimum, etc., for analyzing twelve elements (Tab. 1).

The elements concentrations are not normally distributed for Gaobiegou and Gushenling area (Figs.7, 8(a)).

Figs. 7, 8(b) show the histogram of the data after ilr-transformation. It can be seen that the distribution of the data has changed significantly.

B. GRAPH CLUSTERING RESULTS

The visualization of the graph clustering results is shown in Figs. 9-16 for Gaobiegou and Gushenling data, respectively.

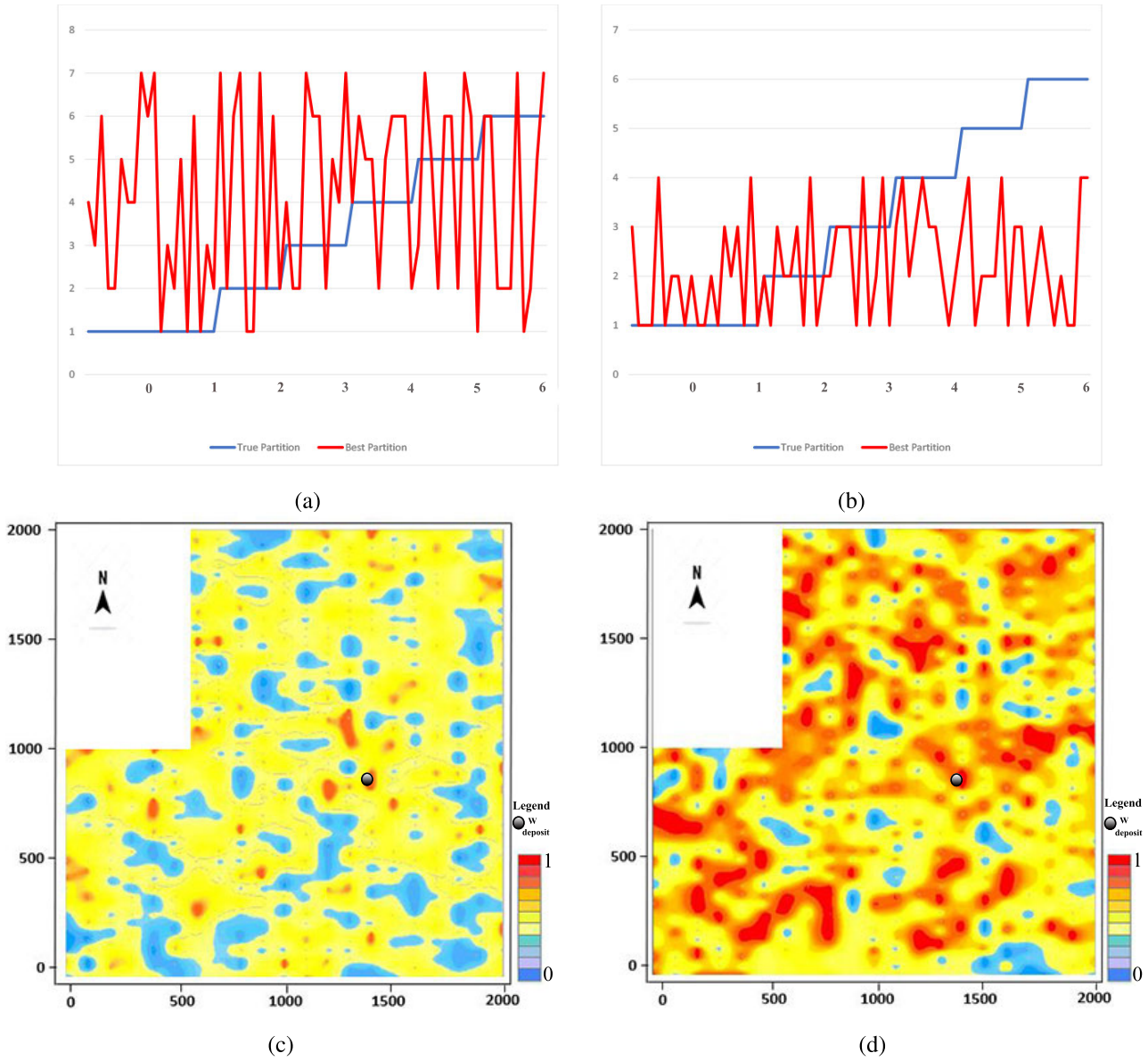


FIGURE 10. The partitions of the geochemical anomaly generated by (a) spectral clustering algorithm, (b) modularity maximization algorithm, and Geochemical anomaly maps obtained by (c) spectral clustering algorithm, (d) modularity maximization algorithm, for Gaobiegou area.

Figs.9, 13 present the clusters generated by modularity maximization algorithm, and the final separation is achieved at *parameter* = 0.5, appears in the x-axis.

Danon algorithm and modularity maximization algorithm can automatically discover the optimal number of clusters, and their results are very close. The spectral clustering algorithm requires providing the maximum number of clusters.

Figs. 10, 14 (a,b) present the clusters of the geochemical anomaly generated by spectral clustering and modularity maximization algorithm for Gaobiegou and Gushenling area, respectively. The x-axis values describe the id of each node, and the y-axis values describe the number of clusters.

1) RESULTS USING DATA OF GAOBIEGOU AREA

In Gaobiegou area (Fig. 10d), the geochemical anomalies are typically detected at stratigraphy, which presents a set of metamorphic sedimentary clastic rocks, divided into two lithologic sections, and fit well into Tungsten deposit. In Fig. 10b the anomalies are characterised by large size, high intensity obvious concentration center and show a ring shape at W deposit, which must be given more focus. Both of spectral clustering (Fig. 10c) and k-means (Fig. 11a) method can identify the anomalies in different locations, with different shapes, but show low intensity.

Fig. 12 shows the distribution of four clusters separately of Gaobiegou area presented in Fig. 10d.

In cluster 1, the geochemical anomalies are mainly detected at the center of the investigation area, and appear with a ring shape, and cover well W deposit. The anomalies are also spread along the faults and related to faulting activities.

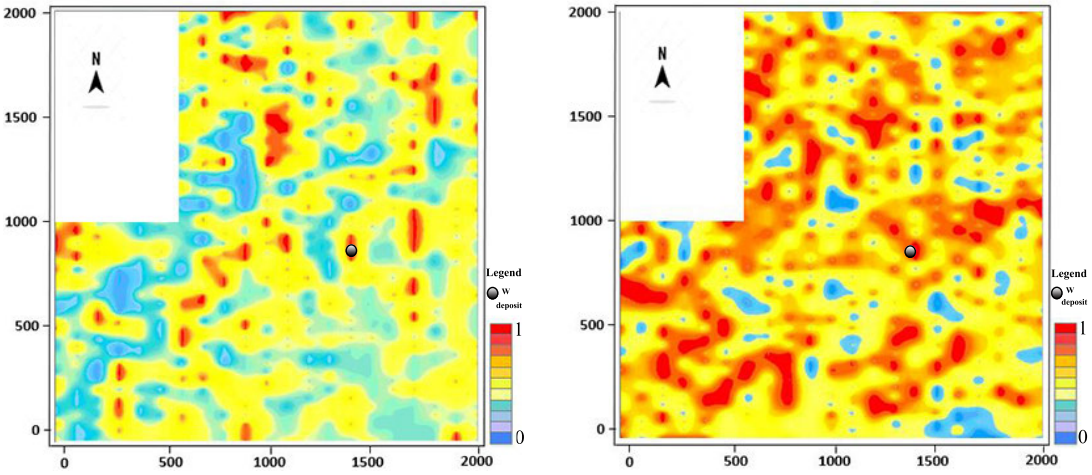


FIGURE 11. Geochemical anomaly maps obtained by (a) k-means clustering (k=6), and (b) modularity maximization algorithm.

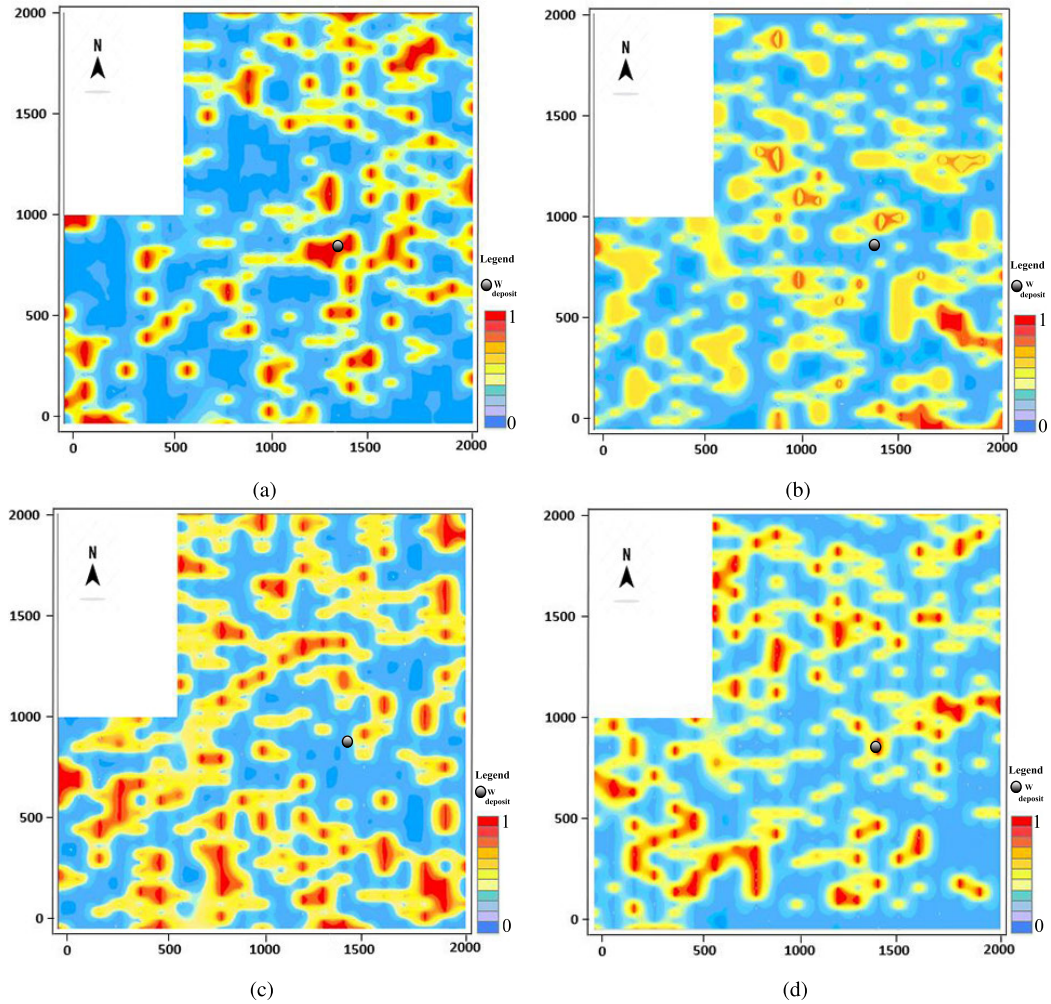


FIGURE 12. Geochemical anomaly maps, (a) cluster 1 (b) cluster 2 (c) cluster 3, and (d) cluster 4, obtained by modularity maximization algorithm, for Gaobiegou area.

In clusters 2 and 4, the geochemical anomalies are primarily identified at Yangsigou Rock Group. This is a set of metamorphic sedimentary clastic rocks, divided into two lithologic sections. The lithology of the lower rock section is black cloud and shallow granulite. It contains dolomitic shallow-grained rocks with dolomite schist and

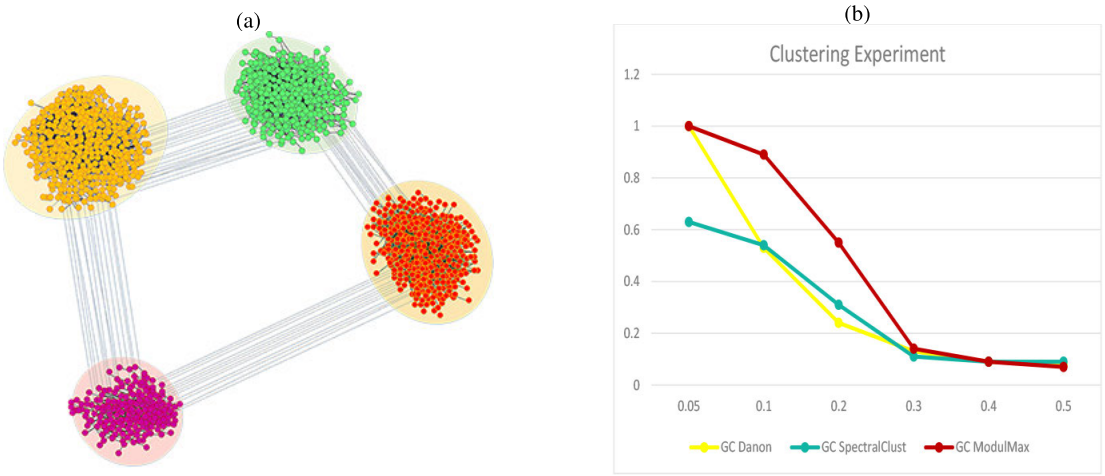


FIGURE 13. The visualization of (a) graph clustering generated by modularity maximization algorithm, and (b) partitions produced by danon algorithm, modularity maximization algorithm and spectral clustering, for Gushenling area.

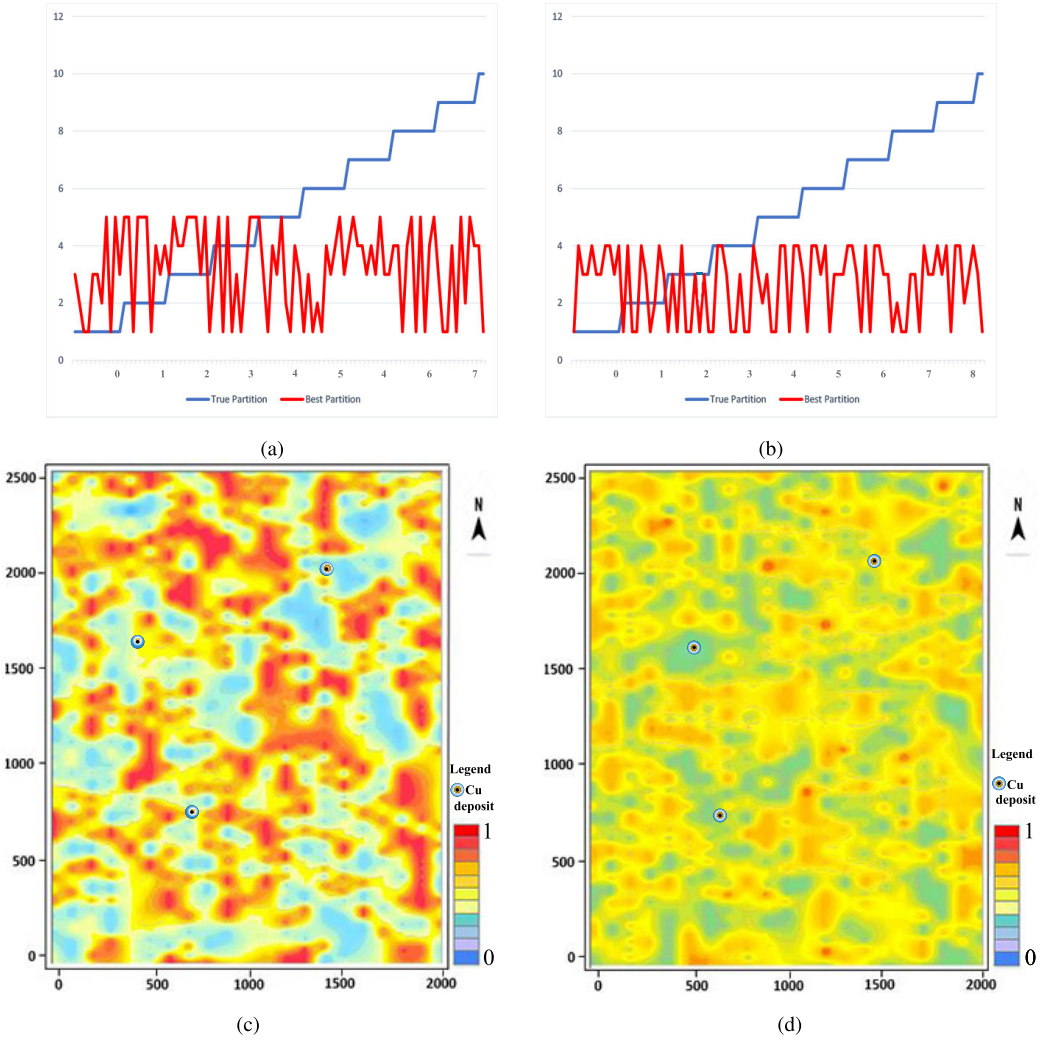


FIGURE 14. The partitions of the geochemical anomaly generated by (a) spectral clustering algorithm, (b) modularity maximization algorithm, and Geochemical anomaly maps obtained by (c) spectral clustering algorithm, (d) modularity maximization algorithm, for Gushanling area.

black cloud granulite. Rocks generally contain magnetite and garnet. The lithology of the upper rock section is black cloud granulite, black cloud sloping granulite, shallow granulite, dolomite quartz schist, magnetite, local garnet, graphite, etc. The sections are characterised by metamorphic minerals where Zn, Au mineralization occur. In addition,

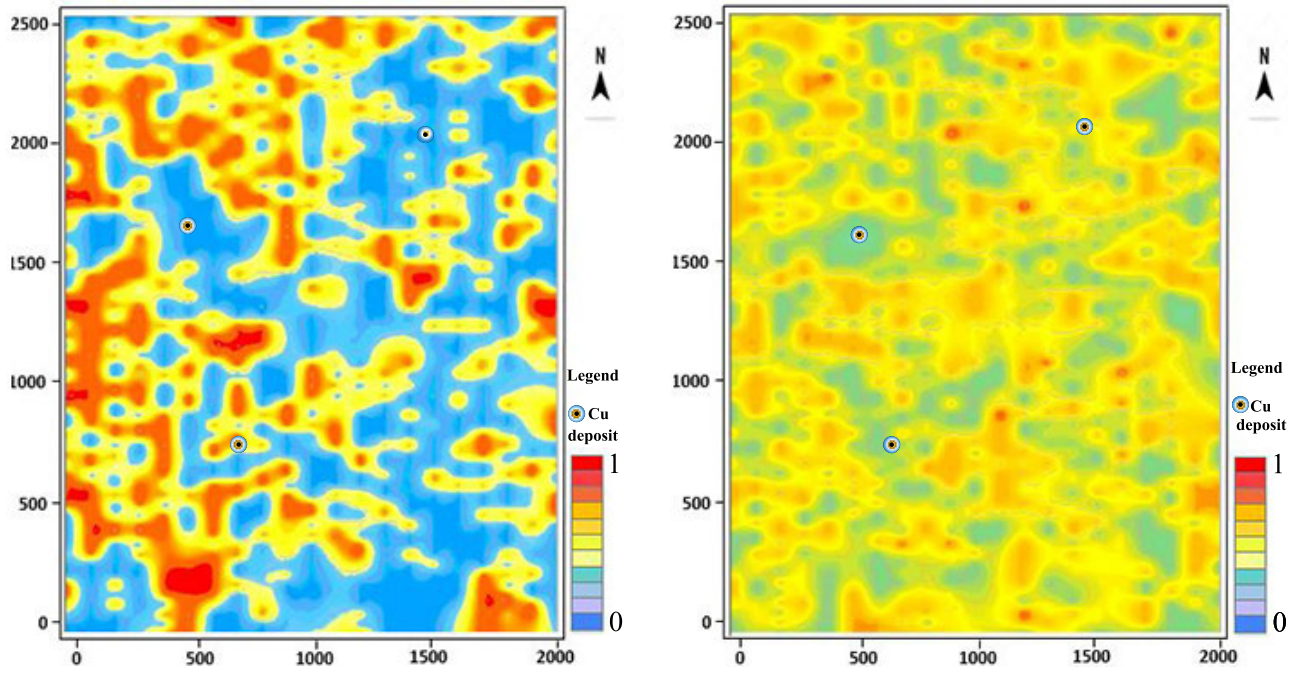


FIGURE 15. Geochemical anomaly maps obtained by (a) k-means clustering (k=6), and (b) modularity maximization algorithm.

the enrichment element W is relatively higher in these rocks.

In cluster 3, the geochemical anomalies are generally recognised at Xushan Group, and are slightly distributed at Yangsigou Rock Group. The Xushan Group is a set of medium-acid volcanic lava, which is mainly characterized by surface overflow, good layering, and forming a clear stacking layer. There are two lithological sections from bottom to middle. The lower section is gray-green andesite, andesite shale. The middle section is mainly the andesite porphyrite of the Great porphyry, where deposits of Au, Ni, Mo, and Zn are hosted. The elements W and Zn are higher in these sections. Besides, the anomalies are detected at the green amphibolite metamorphic domains of Fuping period, mainly the slanted amphibolite.

2) RESULTS USING DATA OF GUSHENLING AREA

In Gushenling area (Fig. 14d), the geochemical anomalies are typically detected at magmatic rock, where volcanic activity provides a source of deposits. In addition, the anomalies are spread along the rivers, and are characterised by high intensity and obvious concentration center. In Fig. 14c the anomalies detected by spectral clustering are obvious, but show low intensity at the same locations. However, the anomalies identified by k-means method (Fig. 15a) are generally detected at the west side of the investigation area.

Fig. 16 shows the distribution of four clusters separately of Gushenling area presented in Fig. 14d.

In cluster 1, the geochemical anomalies are typically detected at magmatic rock of Xushan group, where the lithology is divided into two sections. The first (middle) section is mainly the porphyrites of the Great porphyry. The second (Upper) section is andesites and almond-shaped andesites. The porphyry is the primary cause for the presence of mineral deposits such as Cu, Zn, and Pb. The anomalies fit will the Cu deposit. Besides, the anomalies are distributed along the rivers, which provide a source of mineral deposits.

In cluster 2, the geochemical anomalies are generally distributed along faults, and are related to faulting activities of the investigation area. Furthermore, the fault occasionally opens to allow pulses of high-pressure fluid to be released toward the top, which is particularly rich in the elements of interest, and is important in hosting mineralization. In addition, the geochemical anomalies fit well the Cu deposit.

In clusters 3 and 4, the geochemical anomalies are distributed at the magmatic rocks, which in general contain mineral deposit.

C. QUANTITATIVE ASSOCIATION RULES

In this section, we implement quantitative association rules to reveal important details within each cluster, as illustrated in the support and confidence (section. 5). The minimum support was fixed according to the proportion of each cluster. Regarding the reliability and the number of the rules generated, the minimum value for the support and confidence measures was set to be 0.1 and 0.4, respectively.

The results obtained by QARs are shown in Tables 2, 3.

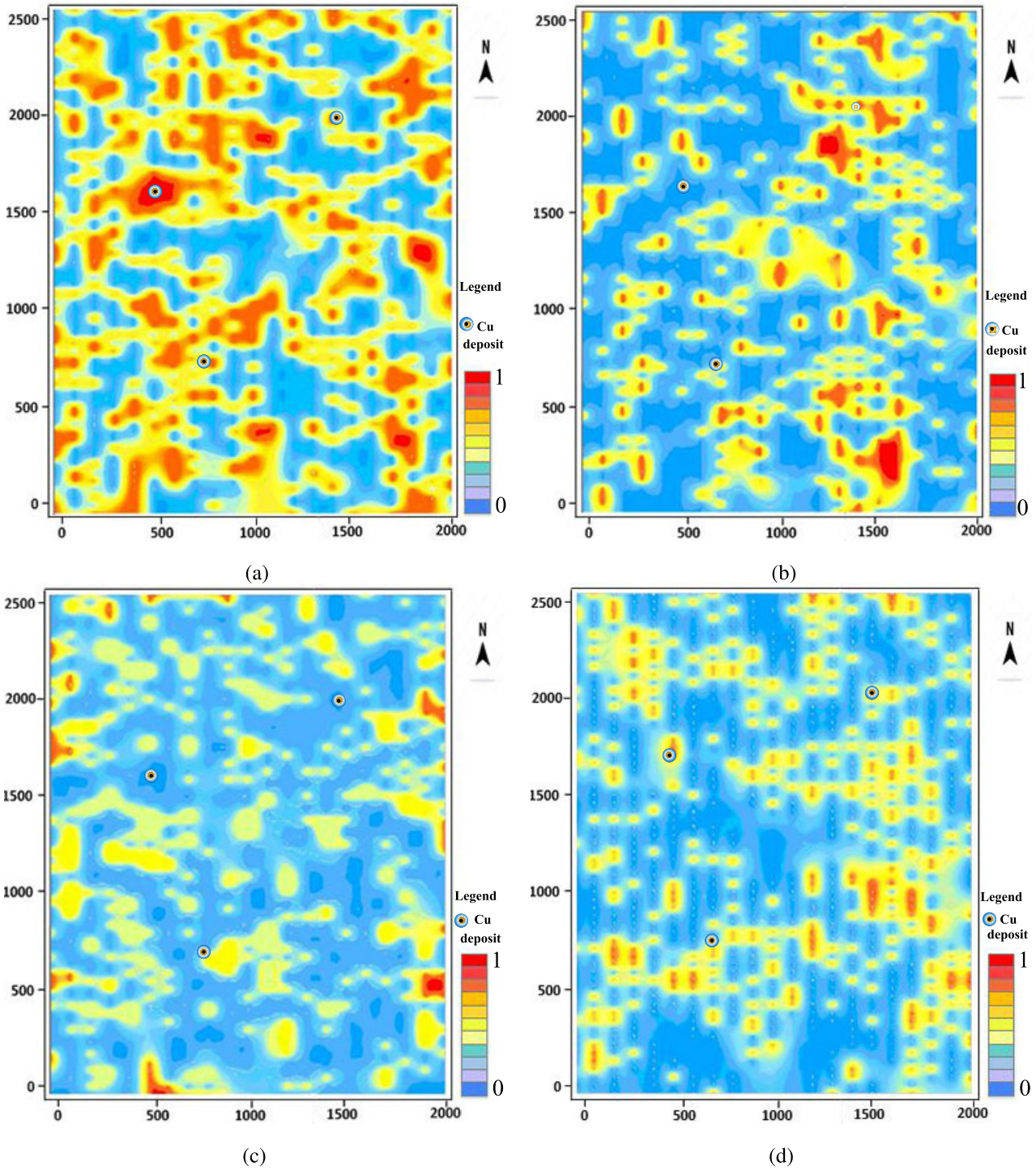


FIGURE 16. Geochemical anomaly maps, (a) cluster 1 (b) cluster 2 (c) cluster 3, and (d) cluster 4, obtained by modularity maximization algorithm, for Gushenling area.

Figs. 17, 18 show the local support of the elements in each cluster. The QARs proposed built a set of rules that cover different areas of the problem, which allow us to understand the anomalies generated.

In Gaobiegou area, the concentration of Tungsten (W) and Gold (Au) is high in four clusters. Meanwhile, Molybdenum (Mo), Nickel (Ni) and Zinc (Zn) are concentrated in three clusters.

As can be observed in Table 2 and Fig. 17, very significant association between W-Au, and Au-Ni with confidence of 0.67 and 0.55, respectively.

The strong association between W-Au can be explained by the fact that the tungsten occurs in vein deposits associated with granites along with gold, and can also be associated with various lithologies. There are, however, other possible explanations related to the geological process [75].

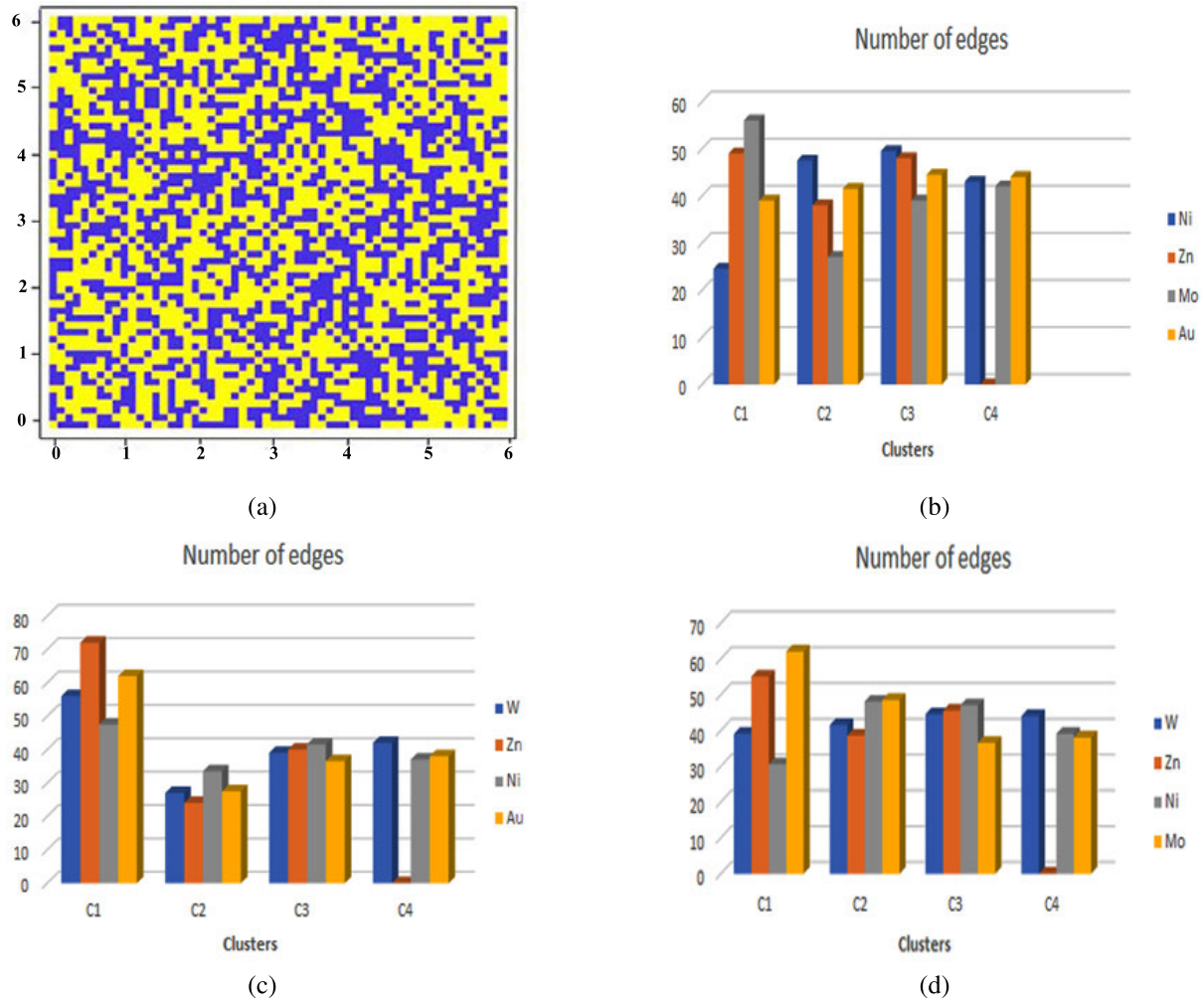


FIGURE 17. The visualization of (a) adjacency matrix of the graph clustering generated by modularity maximization algorithm, and (b) the internal edges related W with Ni, Zn, Mo and Au in four clusters, and (c) the internal edges related to Mo with W, Zn, Ni, and Au in four clusters, and (d) the internal edges related to Au with W, Zn, Ni, and Mo in four clusters, for Gaobiegou area.

TABLE 2. Support and confidence measures to evaluate geochemical rules.

Study area	Geochemical rule	Measure		Relation
		Support	Confidence	
Gaobiegou	$W \rightarrow Au$	42.3%	67%	Very significant
	$Au \rightarrow Ni$	41%	55%	Very significant
	$W \rightarrow Ni$	41.1%	50%	Significant
	$Mo \rightarrow Ni$	40%	50%	Significant
	$W \rightarrow Mo$	41%	50%	Significant
	$Au \rightarrow Mo$	39.1%	41%	Significant
	$Mo \rightarrow Zn$	34%	37%	Low

TABLE 3. Support and confidence measures to evaluate geochemical rules.

Study area	Geochemical rule	Measure		Relation
		Support	Confidence	
Gushenling	$Ag \rightarrow Pb$	50%	70%	Very significant
	$Cu \rightarrow Pb$	40%	62%	Very significant
	$Ag \rightarrow Ni$	30%	41%	Significant
	$Ag \rightarrow Cu$	38.2%	42%	Significant
	$Cu \rightarrow Co$	39.1%	47%	Significant
	$Cu \rightarrow Ni$	30%	41%	Significant
	$Ag \rightarrow Zn$	27%	27%	Low
	$Pb \rightarrow Ni$	20.3%	23%	Low

Significant associations for Mo-Ni, W-Mo, and Au-Mo with confidence of 0.50, 0.50 and 0.41, while the association of Mo with Zn is a little lower with confidence of 0.37.

Hence, the anomaly values divided into three categories : the high anomaly ($>$ 0.50), moderate anomaly (0.50-0.39), and low anomaly ($\geq$ 0.10) (Fig. 19). The high anomaly area occupies 3.4% of the total area, the moderate anomaly occupies 39.1% of the total area.

In Gushenling area, the concentration of Silver (Ag) and Lead (Pb) is high in four clusters, Copper (Cu) and Cobalt (Co) are concentrated in three clusters. Meanwhile, Nickel (Ni) and Zinc (Zn) are only concentrated in two clusters.

As can be observed in Table 3 and Fig. 18, very significant association between Ag-Pb with confidence of 0.7. These

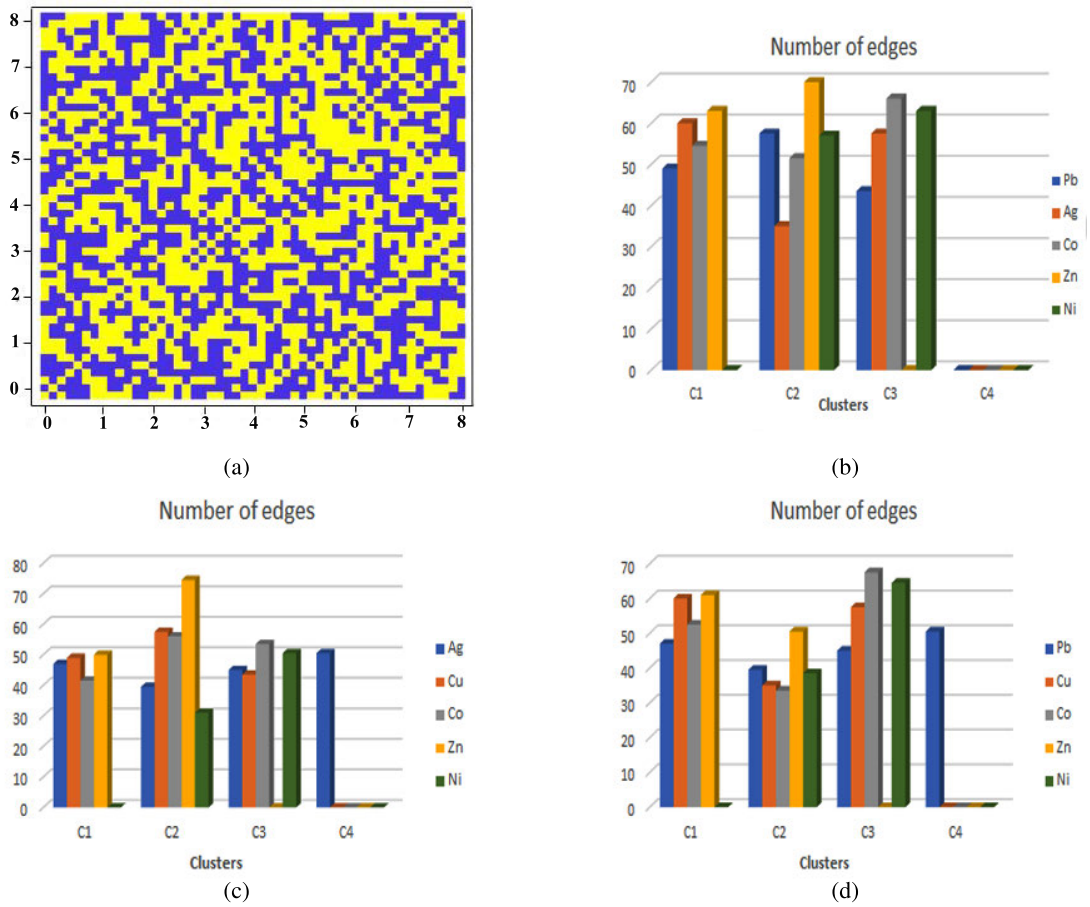


FIGURE 18. The visualization of (a) adjacency matrix of the graph clustering generated by modularity maximization algorithm, and (b) the internal edges related Cu with Pb, Ag, Co, Zn and Ni in four clusters, and (c) the internal edges related to Pb with Ag, Cu, Co, Zn and Ni in four clusters, and (d) the internal edges related to Ag with Pb, Cu, Co, Zn and Ni in four clusters, for Gushenling area.

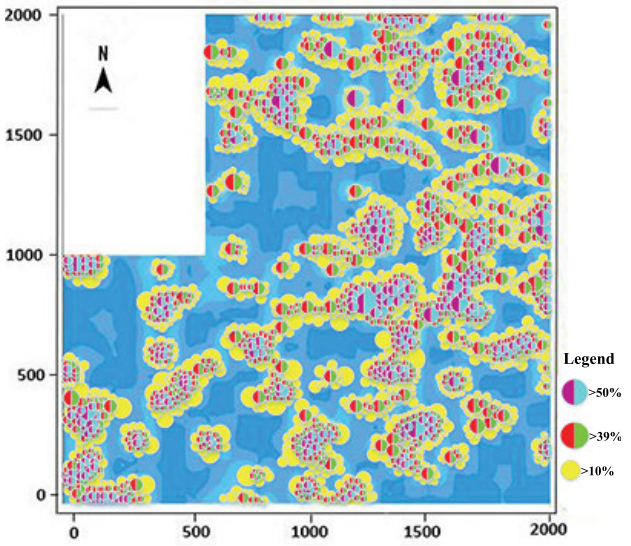


FIGURE 19. Geochemical anomaly map generated by QARs, for Gaobiegou area.

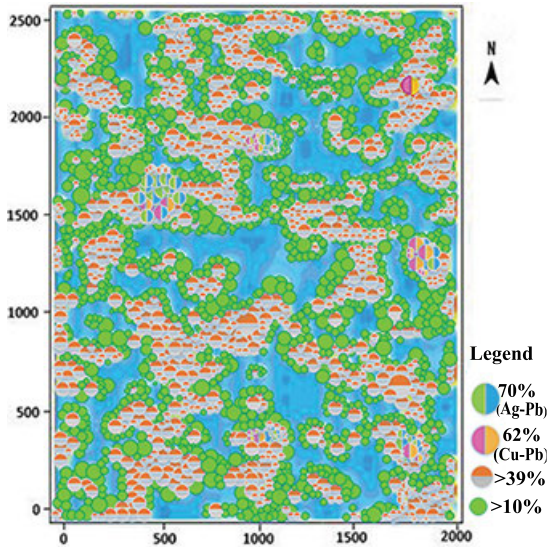


FIGURE 20. Geochemical anomaly map generated by QARs, for Gushenling area.

elements are typically geochemically coherent, and their strong association probably indicates similar characteristics in the hydrothermal mineralization process and probably come from the same geological process.

A significant association exists between Cu-Pb with confidence of 0.62. This strong association suggests their presence genetically related to the volcanic and/or subvolcanic quartz porphyry.

Very significant to significant associations for Ag with Ni (0.41), Ag with Cu (0.42), Cu with Co (0.42), and Cu with Ni (0.41). Meanwhile, the association of Ag with Zn is a little lower with confidence of 0.25. This can be explanted by low concentration of the Zn samples in the study area.

Thus, the anomaly values divided into three categories : the high anomaly ($>$ 0.50), moderate anomaly (0.50-0.39), and low anomaly ($\geq$ 0.10) (Fig. 20). In Gushanling area, the high anomaly area occupies 2.4% of the total area, the moderate anomaly occupies 40.1% of the total area.

From the results, superior results are achieved with GCQAR than a result that is generated by k-means and spectral clustering. Therefore, our results cast a new light on learning the normal element behavior and highlighting anomalies related to it in geochemical data problem.

VII. CONCLUSION AND FUTURE DIRECTIONS

In this study, GCQAR method was implemented to recognize geochemical anomalies. The proposed method sequentially applies graph clustering and quantitative association rules. The results of this work lead to the following conclusions:

The hybrid methodology combining graph clustering and QAR is a useful method for recognizing geochemical anomalies. Graph clustering is used to segment data into meaningful groups, and QAR is performed to learn the normal behavior of the elements and to highlight anomalies related to them.

The GCQAR has significant benefits in terms of recognition of significant geochemical patterns compared to the traditional methods used in the field of geochemistry.

The GCQAR can be used to not only delineate geochemical anomaly zones, but also to improve our understanding of mineralization. It can be a very suitable method for examining nonlinear and complex relationships caused by a variety of geological processes. Thus, the GCQAR is a potential method to be considered for use in geochemistry problem.

It can find high-dimensional clustering and provide the most suitable intervals of values belonging to the rules without implementing a discretization process. Moreover, it helps find reduced sets of significant rules from large dataset. This study will bridge a knowledge gap in terms of recognizing geochemical patterns formed over various lithology. Despite the success demonstrated, a significant limitation of GC (modularity maximization algorithm) is time consuming.

More broadly, the research is also needed to determine negative quantitative association rules. In future work, we plan to use association rules to isolate the overlapping groups by analysing the relation between the external edges, and considering negative quantitative association rules.

ACKNOWLEDGMENT

The authors are grateful to Dr. Byung-Gyu Kim and anonymous reviewer for their constructive comments which greatly improved our manuscript. The authors are grateful to Dr. Yunzhen Chang for his help in collecting geological, and geochemical data of the study area.

REFERENCES

[1] A. Gartman and J. R. Hein, "Mineralization at oceanic transform faults and fracture zones," in *Transform Plate Boundaries and Fracture Zones* Santa Cruz, CA, USA: Candice Janco, 2019, ch. 5, pp. 105–118.

[2] S. A. Meshkani, B. Mehrabi, A. Yaghubpur, and Y. F. Alghalandis, "The application of geochemical pattern recognition to regional prospecting: A case study of the Sanandaj–Sirjan metallogenic zone, Iran," *J. Geochem. Explor.*, vol. 108, pp. 183–195, Mar. 2011.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*. New York, NY, USA: ACM, 1993, pp. 207–216.

[4] S. Nasreen, M. A. Azam, K. Shehzad, U. Naeem, and M. A. Ghazanfar, "Frequent pattern mining algorithms for finding associated frequent patterns for data streams: A survey," *Proc. Comput. Sci.*, vol. 37, pp. 109–116, Sep. 2014.

[5] B. Alata and E. Akin, "An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules," *Soft Comput.*, vol. 10, pp. 230–237, Feb. 2006.

[6] J. Alcala-Fdez, N. Flugy-Pape, A. Bonarini, and F. Herrera, "Analysis of the effectiveness of the genetic algorithms based on extraction of association rules," *Fundamenta Informaticae*, vol. 98, pp. 1–14, Jan. 2010.

[7] J. M. Luna, J. R. Romero, and S. Ventura, "Grammar-based multi-objective algorithms for mining association rules," *Data Knowl. Eng.*, vol. 86, pp. 19–37, Jul. 2013.

[8] D. Martin, A. Rosete, J. Alcalá-Fdez, and F. Herrera, "A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules," *IEEE Trans. Evol. Comput.*, vol. 18, no. 1, pp. 54–69, Feb. 2014.

[9] X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," *Expert Syst. Appl.*, vol. 36, pp. 3066–3076, Mar. 2009.

[10] M. Martínez-Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme, "Obtaining optimal quality measures for quantitative association rules," *Neurocomputing*, vol. 176, pp. 36–47, Feb. 2016.

[11] D. Adhikary and S. Roy, "Trends in quantitative association rule mining techniques," in *Proc. IEEE 2nd Int. Conf. Recent Trends Inf. Syst. (ReTIS)*, Jul. 2015, pp. 126–131. [Online]. Available: https://ieeexplore.ieee.org/document/7232865

[12] K. C. C. Chan and W. H. Au, "An effective algorithm for mining interesting quantitative association rules," in *Proc. ACM Symp. Appl. Comput.* San Jose, CA, USA: ACM, 1997, pp. 88–90. doi: 10.1145/331697.331714.

[13] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, "Mining optimized association rules for numeric attributes," *J. Comput. Syst. Sci.*, vol. 58, pp. 1–12, Feb. 1999.

[14] D. Li, M. Zhang, S. Zhou, and C. Zheng, "A new approach of self-adaptive discretization to enhance the apriori quantitative association rule mining," in *Proc. 2nd Int. Conf. Intell. Syst. Design Eng. Appl.* Washington, DC, USA: IEEE Computer Society, 2012, pp. 44–47.

[15] Y. Guo, J. Yang, and Y. Huang, "An effective algorithm for mining quantitative association rules based on high dimension cluster," in *Proc. 4th Int. Conf. Wireless Commun., Netw. Mobile Comput.* Dalian, China: IEEE, 2008, pp. 1–4.

[16] A. Salleb-Aouissi, C. Vrain, C. Nortet, X. Kong, V. Rathod, and D. Cassard, "QuantMiner for mining quantitative association rules," *J. Mach. Learn. Res.*, vol. 14, pp. 3153–3157, Oct. 1999.

[17] A. Buccianti and E. Grunsky, "Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes?" *J. Geochem. Explor.*, vol. 141, pp. 1–5, Jun. 2014.

[18] V. Pawlowsky-Glahn and J. J. Egozcue, "Spatial analysis of compositional data: A historical review," *J. Geochem. Explor.*, vol. 164, pp. 28–32, May 2016.

[19] B. Bölviken, P. R. Stokke, J. Feder, and T. Jössang, "The fractal nature of geochemical landscapes," *J. Geochem. Explor.*, vol. 43, pp. 91–109, Apr. 1992.

[20] Q. Cheng, F. P. Agterberg, and S. B. Ballantyne, "The separation of geochemical anomalies from background by fractal methods," *J. Geochem. Explor.*, vol. 51, pp. 109–130, Jul. 1994.

[21] R. Zuo, E. John, M. Carranza, and Q. Cheng, "Fractal/multifractal modelling of geochemical exploration data," *J. Geochem. Explor.*, vol. 122, pp. 1–3, Nov. 2012.

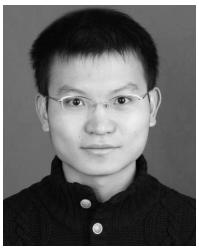
[22] Q. Cheng, Y. Xu, and E. Grunsky, "Integrated spatial and spectrum method for geochemical anomaly separation," *Natural Resour. Res.*, vol. 9, pp. 43–52, Mar. 2000.

[23] C. Li, T. Ma, and J. Shi, "Application of a fractal method relating concentrations and distances for separation of geochemical anomalies from background," *J. Geochem. Explor.*, vol. 77, pp. 167–175, Mar. 2003.
[24] E. C. Grunsky and B. W. Smee, "The differentiation of soil types and mineralization from multi-element geochemistry using multivariate methods and digital topography," *J. Geochem. Explor.*, vol. 67, pp. 287–299, Dec. 1999.
[25] E. C. Grunsky, P. de Caritat, and U. A. Mueller, "Using surface regolith geochemistry to map the major crustal blocks of the Australian continent," *Gondwana Res.*, vol. 46, pp. 227–239, Jun. 2017.
[26] M. Abedi, G.-H. Norouzi, and A. Bahroudi, "Support vector machine for multi-classification of mineral prospectivity areas," *Comput. Geosci.*, vol. 46, pp. 272–283, Sep. 2012.
[27] V. Nykänen and J. J. Ojala, "Spatial analysis techniques as successful mineral-potential mapping tools for orogenic gold deposits in the northern Fennoscandian Shield, Finland," *Natural Resour. Res.*, vol. 16, pp. 85–92, Jun. 2007.
[28] P. Mejía-Herrera, J.-J. Royer, G. Caumon, and A. Cheilletz, "Curvature attribute from surface-restoration as predictor variable in Kupferschiefer copper potentials," *Natural Resour. Res.*, vol. 24, pp. 1–16, Jun. 2014.
[29] Y. Xiong and R. Zuo, "Recognition of geochemical anomalies using a deep autoencoder network," *Comput. Geosci.*, vol. 86, pp. 75–82, Jan. 2016.
[30] Y. Chen, L. Lu, and X. Li, "Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly," *J. Geochem. Explor.*, vol. 140, pp. 56–63, May 2014.
[31] J. A. Anderson, "A simple neural network generating an interactive memory," *Math. Biosci.*, vol. 14, pp. 197–220, Aug. 1972.
[32] R. Zuo, Y. Xiong, J. Wang, and E. J. M. Carranza, "Deep learning and its application in geochemical mapping," *Earth-Sci. Rev.*, vol. 192, pp. 1–14, May 2019.
[33] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[34] E. J. M. Carranza and A. G. Laborte, "Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines)," *Comput. Geosci.*, vol. 74, pp. 60–70, Jan. 2015.
[35] C. Chen, B. He, and Z. Zeng, "A method for mineral prospectivity mapping integrating C4.5 decision tree, weights-of-evidence and m-branch smoothing techniques: A case study in the eastern Kunlun Mountains, China," *Earth Sci. Informat.*, vol. 7, pp. 13–24, Mar. 2014.
[36] S. Zhang, K. Xiao, E. J. M. Carranza, F. Yang, and Z. Zhao, "Integration of auto-encoder network with density-based spatial clustering for geochemical anomaly detection for mineral exploration," *Comput. Geosci.*, vol. 130, pp. 43–56, Sep. 2019.
[37] X. Yu, F. Xiao, Y. Zhou, Y. Wang, and K. Wang, "Application of hierarchical clustering, singularity mapping, and Kohonen neural network to identify Ag-Au-Pb-Zn polymetallic mineralization associated geochemical anomaly in Pangxidong district," *J. Geochem. Explor.*, vol. 203, pp. 87–95, Aug. 2019.
[38] K. J. Ellefsen, D. B. Smith, and J. D. Horton, "A modified procedure for mixture-model clustering of regional geochemical data," *Appl. Geochem.*, vol. 51, pp. 315–326, Dec. 2014.
[39] G. Tepanosyan, L. Sahakyan, C. Zhang, and A. Saghatelyan, "The application of Local Moran's I to identify spatial clusters and hot spots of Pb, Mo and Ti in urban soils of Yerevan," *Appl. Geochem.*, vol. 104, pp. 116–123, May 2019.
[40] S. Zaremotlagh, A. Hezarkhani, and M. Sadeghi, "Detecting homogenous clusters using whole-rock chemical compositions and REE patterns: A graph-based geochemical approach," *J. Geochem. Explor.*, vol. 170, pp. 94–106, Nov. 2016.
[41] M. G. Di Giuseppe, A. Troiano, D. Patella, M. Piochi, and S. Carlino, "A geophysical k-means cluster analysis of the Solfatara-Pisciarelli volcano-geothermal system, Campi Flegrei (Naples, Italy)," *J. Appl. Geophys.*, vol. 156, pp. 44–54, Sep. 2018.
[42] K. J. Ellefsen, D. B. Smith, and J. D. Horton, "A modified procedure for mixture-model clustering of regional geochemical data," *Appl. Geochem.*, vol. 51, pp. 315–326, Dec. 2014.
[43] F. Fouedjio, "A hierarchical clustering method for multivariate geostatistical data," *Spatial Statist.*, vol. 18, pp. 333–351, Nov. 2016.
[44] M. Brehme, K. Bauer, M. Nukman, and S. Regenspurg, "Self-organizing maps in geothermal exploration—A new approach for understanding geochemical processes and fluid evolution," *J. Volcanol. Geothermal Res.*, vol. 336, pp. 19–32, Apr. 2017.
[45] B. S. Penn, "Using self-organizing maps to visualize high-dimensional data," *Comput. Geosci.*, vol. 31, pp. 531–544, Jun. 2005.
[46] S. Singh, R. Garg, and P. K. Mishra, "Performance optimization of MapReduce-based Apriori algorithm on Hadoop cluster," *Comput. Electr. Eng.*, vol. 67, pp. 348–364, Apr. 2018.
[47] X. Wang, C. Song, W. Xiong, and X. Lv, "Evaluation of flotation working condition recognition based on an improved Apriori algorithm," *IFAC-PapersOnLine*, vol. 51, pp. 129–134, Oct. 2018.
[48] C. Aori and M. Craus, "Grid implementation of the Apriori algorithm," *Adv. Eng. Softw.*, vol. 38, pp. 295–300, May 2007.
[49] J. Wang and Z. Cheng, "FP-Growth based regular behaviors auditing in electric management information system," *Proc. Comput. Sci.*, vol. 139, pp. 275–279, Oct. 2018.
[50] M. Houtsma and A. Swami, "Set-oriented mining for association rules in relational databases," in *Proc. 11th Int. Conf. Data Eng.* Taipei, Taiwan: IEEE, 1995, pp. 25–33.
[51] Y. Zeng, S. Yin, J. Liu, and M. Zhang, "Research of improved FP-Growth algorithm in association rules mining," *Sci. Program.*, vol. 3, p. 6, Jan. 2015.
[52] X. Yao, "Evolutionary Computation," in *Evolutionary Optimization*. Boston, MA, USA: Springer, 2002, pp. 27–53.
[53] D. Martn, A. Rosete, J. Alcalá-Fdez, and F. Herrera, "QAR-CIP-NSGA-II: A new multi-objective evolutionary algorithm to mine quantitative association rules," *Inf. Sci.*, vol. 258, pp. 1–28, Feb. 2014.
[54] C. Reimann and P. Filzmoser, "Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data," *Environ. Geol.*, vol. 39, pp. 1001–1014, Jul. 2000.
[55] J. Aitchison, "The statistical analysis of compositional data," *Math. Geol.*, vol. 16, pp. 531–564, Aug. 1984.
[56] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, "Isometric logratio transformations for compositional data analysis," *Math. Geol.*, vol. 35, no. 3, pp. 279–300, Apr. 2003.
[57] J. L. Shelton, M. A. Engle, A. Buccianti, and M. S. Blondes, "The isometric log-ratio (ilr)-ion plot: A proposed alternative to the Piper diagram," *J. Geochem. Explor.*, vol. 190, pp. 130–141, Jul. 2018.
[58] A. Buccianti and R. Zuo, "Weathering reactions and isometric log-ratio coordinates: Do they speak to each other?" *Appl. Geochem.*, vol. 75, pp. 189–199, Dec. 2016.
[59] Z. Wang and W. Shi, "Robust variogram estimation combined with isometric log-ratio transformation for improved accuracy of soil particle-size fraction mapping," *Geoderma*, vol. 324, pp. 56–66, Aug. 2018.
[60] A. Tlacuilo-Parra, R. Morales-Zambrano, N. Tostado-Rabago, M. A. Esparza-Flores, B. Lopez-Guido, and J. Orozco-Alcala, "Inactivity is a risk factor for low bone mineral density among haemophilic children," *Brit. J. Haematol.*, vol. 140, pp. 562–567, Mar. 2010.
[61] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, Jun. 2010.
[62] P. G. Sun and X. Sun, "Complete graph model for community detection," *Phys. A, Stat. Mech. Appl.*, vol. 471, pp. 88–97, Apr. 2017.
[63] H. Zhou, J. Li, J. Li, F. Zhang, and Y. Cui, "A graph clustering method for community detection in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 551–562, Mar. 2017.
[64] C. Shi, Y. Cai, D. Fu, Y. Dong, and B. Wu, "A link clustering based overlapping community detection algorithm," *Data Knowl. Eng.*, vol. 87, no. 9, pp. 394–404, Sep. 2013.
[65] C. W. Loe and H. J. Jensen, "Comparison of communities detection algorithms for multiplex," *Phys. A, Stat. Mech. Appl.*, vol. 431, pp. 29–45, Aug. 2015.
[66] R. Liu, S. Feng, R. Shi, and W. Guo, "Weighted graph clustering for community detection of large social networks," *Proc. Comput. Sci.*, vol. 31, pp. 85–94, Nov. 2014.
[67] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *J. Netw. Comput. Appl.*, vol. 108, pp. 87–111, Apr. 2018.
[68] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, pp. 8577–8582, Jun. 2006.
[69] D. Martín, M. Martínez-Ballesteros, D. García-Gil, J. Alcalá-Fdez, F. Herrera, and J. C. Riquelme-Santos, "MRQAR: A generic MapReduce framework to discover quantitative association rules in big data problems," *Knowl.-Based Syst.*, vol. 153, pp. 176–192, Aug. 2018.
[70] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, "Selecting the best measures to discover quantitative association rules," *Neurocomputing*, vol. 126, pp. 3–14, Feb. 2014.

[71] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Aug. 2007.
[72] E. L. Martelot and C. Hankin, "Multi-scale community detection using stability as optimisation criterion in a greedy algorithm," in *Proc. KDIR*, Paris, France, 2011, pp. 216–225.
[73] A. Serra and R. Tagliaferri, "Unsupervised Learning: Clustering," in *Encyclopedia of Bioinformatics and Computational Biology*. Oxford, U.K.: Academic, 2019, pp. 350–357.
[74] X.-H. Ming and C.-S. Yu, "The programming of drilling log drawing system based on MAPGIS," in *Proc. Comput. Techn. Geophys. Geochem. Explor.*, vol. 26, pp. 85–90, Feb. 2004.
[75] A. Beaudoin, G. Perrault, and M. Bouchard, "Distribution of gold, arsenic, antimony and tungsten around the Dest-Or Orebody, Noranda district, Abitibi, Quebec," *J. Geochem. Explor.*, vol. 28, pp. 41–70, Jun. 1987.

YASMINA MEDJADBA was born in Batna, Algeria. She received the master's degree in computer network and information security from the University of Batna, Algeria, in 2013. She is currently pursuing the Ph.D. degree in computer science with the College of Information Science and Technology Beijing Normal University, China. Her research interests include pattern recognition and anomaly detection, machine learning, and signal processing.

DAN HU was born in 1977. She received the B.S. and M.S. degrees in mathematics from Sichuan Normal University, Chengdu, China, in 1999 and 2002, respectively, and the Ph.D. degree in applied mathematics from Beijing Normal University, Beijing, China. She is currently a Research Scholar with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, working in the field of machine learning.



WEI LIU was born in 1987. He is currently pursuing the Ph.D. degree in computer science with the College of Information Science and Technology, Beijing Normal University, China. His research interests include machine learning, big data, and astronomical data classification.

XIANCHUAN YU (SM'09) was born in 1967. He received the Ph.D. degree in mathematical geology from Jilin University. He is the Director of the School Key Laboratory of Spatial Multisource Information Fusion and Analysis. He is currently a Professor with the Computer Science Department, College of Information Science and Technology, Beijing Normal University. He is the Academic Leader of Intelligent Information Processing, Beijing Normal University. He has been a Vice Director of China Mathematical Geology Information Processing Professional committee, since 2012, a Vice Chairman of the Branch China National Committee of the International Mathematical Geoscience Society (IAMG), since 2009, an Executive Member of Chinese Computer Graphics and GIS Society, a Committee Standing Member of the Chinese Aerospace Optical Society of Technical, and a Professional Committee Member of CCF Collaborative Computing. His current research interests include blind source separation, modeling uncertainty in geoscience, remote image processing, fuzzy sets, and mineral resources appraisement.