Applied Geochemistry 51 (2014) 315–326

Contents lists available at ScienceDirect

Applied Geochemistry

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

A modified procedure for mixture-model clustering of regional geochemical data

Karl J. Ellefsen [a,*], David B. Smith [b,1], John D. Horton [b,2]

[a] U.S. Geological Survey, MS 964, Box 25046, Denver, CO, USA
[b] U.S. Geological Survey, MS 973, Box 25046, Denver, CO, USA

ARTICLE INFO

Article history:
Available online 18 October 2014
Editorial handling by M. Kersten

ABSTRACT

A modified procedure is proposed for mixture-model clustering of regional-scale geochemical data. The key modification is the robust principal component transformation of the isometric log-ratio transforms of the element concentrations. This principal component transformation and the associated dimension reduction are applied before the data are clustered. The principal advantage of this modification is that it significantly improves the stability of the clustering. The principal disadvantage is that it requires subjective selection of the number of clusters and the number of principal components. To evaluate the efficacy of this modified procedure, it is applied to soil geochemical data that comprise 959 samples from the state of Colorado (USA) for which the concentrations of 44 elements are measured. The distributions of element concentrations that are derived from the mixture model and from the field samples are similar, indicating that the mixture model is a suitable representation of the transformed geochemical data. Each cluster and the associated distributions of the element concentrations are related to specific geologic and anthropogenic features. In this way, mixture model clustering facilitates interpretation of the regional geochemical data.

Published by Elsevier Ltd.

1. Introduction

Regional-scale geochemical surveys typically involve the collection and chemical analysis of soil or stream-sediment samples at multiple sites across thousands to millions of square kilometers. The sample density varies enormously—from 1 site per 10 to 100 km^2 (e.g., Webb et al., 1978; Fauth et al., 1985; Thalmann et al., 1989; McGrath and Loveland, 1992) to 1 site per 1000 to 5000 km^2 (e.g., Reimann et al., 2003; Salminen et al., 2005; Caritat and Cooper, 2011; Smith et al., 2013). For each of the thousands of samples, the concentrations of multiple elements are usually measured. An important part of the geochemical interpretation is relating the spatial distribution of the element concentrations to features such as bedrock and surficial geology. The traditional method of establishing these relations involves comparing maps of the element concentrations to geologic maps. The traditional method is somewhat straightforward when the geochemical data comprise only a few elements, but is difficult when the data comprise many elements (e.g., 30 elements).

When there are many elements, a multivariate statistical method called "clustering" can help with the interpretation. The essential idea of clustering is that the regional geochemical data may be considered a mixture of data from different geochemical processes, and the clustering partitions the data into groups that are associated with the processes. The data from each geochemical process often are localized to a specific region and may be associated with geologic or anthropogenic features. When such associations occur, they greatly facilitate the interpretation of the geochemical data.

Clustering is a well-established method and is described in many multivariate statistics books (e.g., Johnson and Wichern, 2007, 671–706; Hastie et al., 2009, 501–528). Nonetheless, the application of clustering to geochemical data involves several significant difficulties: The data are compositional, so they cannot be directly analyzed with standard statistical methods (Pawlowsky-Glahn, 2003); some measurements are below the lower reporting limit (i.e., they are left-censored); those measurements just above the lower reporting limit tend to have large uncertainty; and modern data sets often include measured concentrations for about forty elements for each sample (i.e., the data sets have high dimension).

Templ et al. (2008) compared the efficacy of many different clustering procedures for processing regional geochemical data. Among their many findings, they reported that a particular method

* Corresponding author. Tel.: +1 303 236 7032.
E-mail addresses: ellefsen@usgs.gov (K.J. Ellefsen), dsmith@usgs.gov (D.B. Smith), jhorton@usgs.gov (J.D. Horton).
[1] Tel.: +1 303 236 1849.
[2] Tel.: +1 303 236 1921.

http://dx.doi.org/10.1016/j.apgeochem.2014.10.011
0883-2927/Published by Elsevier Ltd.

called "model-based clustering" works well (Fraley and Raftery, 2002). Model-based clustering is based on a particular probability model called a "finite mixture model" (McLachlan and Peel, 2000, 6), so we refer to this method as "mixture-model clustering" for the rest of this article. Reimann et al. (2008, 233–247) and Grunsky (2010) summarized how geochemical data can be analyzed with different clustering methods; Reimann et al. reported favorable results using mixture-model clustering. Morrison et al. (2011) present an application of mixture-model clustering to soil geochemical data from California (USA).

We modified a mixture-model clustering procedure that was originally presented by Templ et al. (2008), and we present the modified procedure in this article. We explain why the finite mixture model is appropriate for clustering of regional geochemical data. To demonstrate the modified procedure, we apply it to soil geochemical data collected in Colorado (USA). In addition, we compare clustering results from the modified procedure and the unmodified procedure.

2. Method

2.1. Mixture-model clustering

Consider a geochemical survey for which the survey area consists of two regions with different geochemical properties. Assume that the geochemical properties of each region may be characterized by a probability density function (pdf). The pdf for region 1 is $f(\mathbf{z}|\boldsymbol{\theta}_1)$ where $\boldsymbol{\theta}_1$ represents the parameters that characterize the pdf. Variable $\mathbf{z}$ represents the element concentrations that have undergone a transformation to make them suitable for standard statistical analysis; the transformation is described in Section 2.2. The pdf for region 2 is identical, except that its parameters are represented by $\boldsymbol{\theta}_2$. If a particular sample, which is indexed by i, is within region 1, then $f(\mathbf{z}_i|\boldsymbol{\theta}_1)$ usually will be large-valued, whereas $f(\mathbf{z}_i|\boldsymbol{\theta}_2)$ usually will be small-valued. In contrast, if sample i is within region 2, the relation is reversed.

The pdf for the entire survey area $p(\mathbf{z})$ is a weighted sum of the two pdfs for the two regions: $p(\mathbf{z}) = \lambda_1 f(\mathbf{z}|\boldsymbol{\theta}_1) + \lambda_2 f(\mathbf{z}|\boldsymbol{\theta}_2)$. Weight λ_1 equals the area of region 1 divided by the area of the entire survey region. Weight λ_2 is similarly defined. Each weight is the relative contribution of its associated pdf to $p(\mathbf{z})$. When the sample locations for the geochemical survey are evenly distributed throughout the survey area, such as for a soil geochemical survey, the fraction of samples from region 1 is approximately λ_1, and the fraction of samples from region 2 is approximately λ_2. Pdf $p(\mathbf{z})$ is an example of a finite mixture model.

To use the finite mixture model in practice, it is generalized to J regions:

$$p(\mathbf{z}) = \sum_{j=1}^{J} \lambda_j f(\mathbf{z}|\boldsymbol{\theta}_j) \tag{1}$$

Pdf $p(\mathbf{z})$ is interpreted as a mixture of J pdfs, each of which represents the geochemical properties of J regions within the survey area. The weights are constrained: $0 \leqslant \lambda_j \leqslant 1$ and $\sum_{j=1}^{J} \lambda_j = 1$ (McLachlan and Peel, 2000, 6). We believe that a finite mixture model is a suitable mathematical representation for the geochemical properties of a survey area. Note that pdf $p(\mathbf{z})$ depends upon variables λ_j and $\boldsymbol{\theta}_j$ ($j = 1, \ldots, J$), but these variables are omitted from the notation to simplify it.

The extent to which sample i is associated with pdf j is expressed quantitatively with the conditional probability:

$$g_{ij} = \frac{\lambda_j f(\mathbf{z}_i|\boldsymbol{\theta}_j)}{p(\mathbf{z}_i)} \tag{2}$$

for $i = 1, \ldots, n$ where n is the number of samples (Fraley and Raftery, 2002). The greater the value of g_{ij}, the closer the association of sample i with pdf j. Those samples for which $g_{ij} \geqslant 0.5$ for pdf j constitute a "cluster." The number of clusters equals the number of pdfs in the finite mixture model. Thus, we call this procedure "mixture-model clustering."

2.2. Clustering procedure

Clustering begins by preparing the geochemical data, which involves eliminating those measurements with significant errors because they could affect adversely the clustering. The details of this step are most easily described with an example (see Section 3.2).

Clustering cannot be applied directly to element concentrations because they are a type of compositional data. Such data have two unique properties: they sum to a constant value (e.g., 1,000,000 mg/kg), and they are greater than or equal to zero (Aitchison, 1986, 25). The consequence of these two properties is that the algebraic operations for compositional data differ from those for non-compositional (conventional) data (Aitchison, 1986, 48–63). To overcome this problem, element concentrations are mathematically transformed with the isometric log-ratio (ilr) transform (Egozcue et al., 2003). The resulting ilr-transformed concentrations are a type of conventional data, which can be analyzed with standard statistical methods (Pawlowsky-Glahn, 2003; Mateu-Figueras et al., 2011).

The ilr-transformed concentrations are then transformed to principal components. This transformation is important for at least four reasons. First, the transformation reduces the dimension of the data, which, in turn, reduces the chances that the mixture-model clustering will encounter numerical problems (Fraley and Raftery, 2002). Second, the transformation removes correlations among the ilr-transformed concentrations, which could affect the mixture-model clustering. Third, the transformation arranges the data according to the amount of information it contains; this important issue is described in detail later in this section. Fourth, the transformation and the associated dimension reduction significantly improve the stability of the clustering. This important property is demonstrated in Section 4.3.

An important step in the transformation to principal components is calculating the mean vector and covariance matrix of the ilr-transformed concentrations. As suggested by Filzmoser et al. (2009), the calculation of these two quantities should be robust, meaning that the calculation should be somewhat insensitive to noise. To this end, the calculation is performed using minimum covariance determinant estimator (Rousseeuw and van Driessen, 1999), which is implemented in software function "covMcd" within the R statistical programming language (R Core Team, 2013). A key parameter in the calculation is α, which specifies the fraction of data that is excluded because its statistical distance from the mean vector is too large (Johnson and Wichern, 2007, 31–36). For example, if α is 0.35, then 35% of the data is excluded. Because those data with significant error have already been eliminated, we believe that parameter α should be less than 0.10; so, we pick several values between 0.0 and 0.10.

The principal components are still ilr-transformed concentrations — the origin has been translated so that the mean of the ilr-transformed concentrations is zero, and the coordinate system has been rotated so that the first coordinate accounts for most of the variance (i.e., information) in the data, the second coordinate for the next most variance, and so on. Thus, the principal components order the data according to the amount of information that the data contain. This ordering is apparent in a scree plot (Fig. 1), which shows the variance of each principal component. (This particular scree plot pertains to the actual principal components cal-