Applied Geochemistry 75 (2016) 200–210

Contents lists available at ScienceDirect

Applied Geochemistry

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

Manual hierarchical clustering of regional geochemical data using a Bayesian finite mixture model

Karl J. Ellefsen*, David B. Smith [1]

U.S. Geological Survey, MS 964, Box 25046, Denver, CO, USA

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form
19 May 2016
Accepted 24 May 2016
Available online 26 May 2016

Keywords:
Clustering
Geochemical data
Compositional data
Finite mixture model
Bayesian modeling
Regional geochemistry

ABSTRACT

Interpretation of regional scale, multivariate geochemical data is aided by a statistical technique called “clustering.” We investigate a particular clustering procedure by applying it to geochemical data collected in the State of Colorado, United States of America. The clustering procedure partitions the field samples for the entire survey area into two clusters. The field samples in each cluster are partitioned again to create two subclusters, and so on. This manual procedure generates a hierarchy of clusters, and the different levels of the hierarchy show geochemical and geological processes occurring at different spatial scales. Although there are many different clustering methods, we use Bayesian finite mixture modeling with two probability distributions, which yields two clusters. The model parameters are estimated with Hamiltonian Monte Carlo sampling of the posterior probability density function, which usually has multiple modes. Each mode has its own set of model parameters; each set is checked to ensure that it is consistent both with the data and with independent geologic knowledge. The set of model parameters that is most consistent with the independent geologic knowledge is selected for detailed interpretation and partitioning of the field samples.

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 10–100 km^2 (e.g., Webb et al., 1978; Fauth et al., 1985; Thalmann et al., 1989; McGrath and Loveland, 1992) to 1 site per 1000–5000 km^2 (e.g., Reimann et al., 2003; Salminen et al., 2005; Caritat and de 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 difficult when the geochemical data comprise only a few elements, and the difficulty increases as the number of chemical elements increases.

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 at least two difficulties: (1) the data are compositional, so they cannot be directly analyzed with standard statistical methods (Pawlowsky-Glahn et al., 2015); and (2) modern data sets often include measured concentrations for about 40 elements for each sample (i.e., the data sets are large).

Several research groups have applied clustering to geochemical data. Templ et al. (2008) compared the efficacy of many different clustering procedures for processing regional geochemical data.

* Corresponding author.
E-mail addresses: ellefsen@usgs.gov (K.J. Ellefsen), dbsmith13@gmail.com (D.B. Smith).

[1] Retired from U.S. Geological Survey, MS 973, Box 25046, Denver, CO, USA.

http://dx.doi.org/10.1016/j.apgeochem.2016.05.016

Reimann et al. (2008, 233–247) and Grunsky (2010) summarized how geochemical data can be analyzed with different clustering methods. Both Templ et al. and Reimann et al. report favorable results using a particular algorithm called “model-based clustering” (Fraley and Raftery, 2002). Morrison et al. (2011) present an application of this model-based clustering to soil geochemical data from California (USA). Ellefsen et al. (2014) modified the clustering procedure that was originally presented by Templ et al. (2008); the modification makes the clustering more robust than it would be otherwise.

In this article, we investigate another clustering procedure, which is based on a hierarchy. At the highest level of the hierarchy, the field samples for the entire survey area are partitioned into two clusters; at the next level in the hierarchy, each of the two clusters is partitioned into two sub-clusters, and so on. Each level of the hierarchy shows geochemical processes occurring at different spatial scales. The clustering method is Bayesian finite mixture modeling; this method has been applied to many types of data (Gelman et al., 2014, p. 539–540) but not to regional geochemical data. The clustering procedure is applied to soil geochemical data collected in the State of Colorado, the United States of America; these data were clustered previously using a different procedure (Ellefsen et al., 2014).

2. Geochemical data

2.1. Survey area, sample collection, and chemical analysis

The geochemical survey area is the State of Colorado (Fig. 1), which has a land area of 269,837 km^2. The geology of Colorado is complex and heterogeneous but can be grouped into five major geologic regions. The regions (listed from largest to smallest) are the Great Plains, in the eastern half of the state; the Southern Rocky Mountains, a north-south swath in the middle of the state; the Colorado Plateau in the west and southwest; the Wyoming Basin in the northwest; and the Middle Rocky Mountains in the northwestern corner. Additional information about the geology of Colorado is reported in Tweto (1979) and numerous publications of the Colorado Geological Survey (http://geosurvey.state.co.us/Pages/CGSHome.aspx).

To select the sample locations, the State of Colorado was divided into 966 polygons for which the areas are all 280 km^2. Within each polygon, one point was selected at random to be the potential sample location. The actual sample location had to satisfy three criteria: (1) it had to be close to the potential sample location; (2) the landscape at the actual location had to be somewhat representative of the landscape in the polygon, as determined by the field geochemist; and (3) the soil at the actual location had no obvious contamination or other disturbance due to human activity, although the soil could be from an agricultural field or pasture. Six potential sample locations were difficult to access, so these were omitted from the survey. At each location, loose plant debris (if any) was removed from the ground surface, and the soil sample was collected from a depth interval of 0–15 cm.

Each soil sample was air dried at ambient temperature, disaggregated, and sieved through a 2-mm stainless steel screen. The sieved material was crushed to less than 150 μm in a ceramic mill and thoroughly mixed to ensure that it was homogeneous. The prepared samples were sent to a U.S. Geological Survey contract geochemical laboratory, where the concentrations of 44 elements were measured. Additional information, as well as the measured concentrations and sample locations, are reported in Smith et al. (2010). Summary statistics of the measured concentrations are listed in Table S1 that is within the Supplementary Material.

2.2. Data editing

We edited the soil geochemical data to make them suitable for clustering. First, field sample “06co437” was culled from the data set because it had an anomalously high copper (Cu) concentration that was likely caused by human activity. Second, silver (Ag), tellurium (Te), cesium (Cs), mercury (Hg), and selenium (Se) were removed from the data set because they had high percentages of their measured concentrations below their lower limits of determination (Table S1 in Supplementary Materials). Third, the left-censored concentrations for antimony (Sb), arsenic (As), bismuth (Bi), cadmium (Cd), indium (In), phosphorous (P), and sulfur (S) were assigned concentrations equal to 0.65 times their respective lower limits of determination (Palarea-Albaladejo et al., 2014). Because the percentages of left censored concentrations were small (Table S1 in Supplementary Materials), this assignment was assumed to have a negligible effect on the clustering. Finally, the element concentrations were scaled so that the units for all concentrations are “mg/kg.” After this editing, there were 959 field samples for which 39 element concentrations are reported.

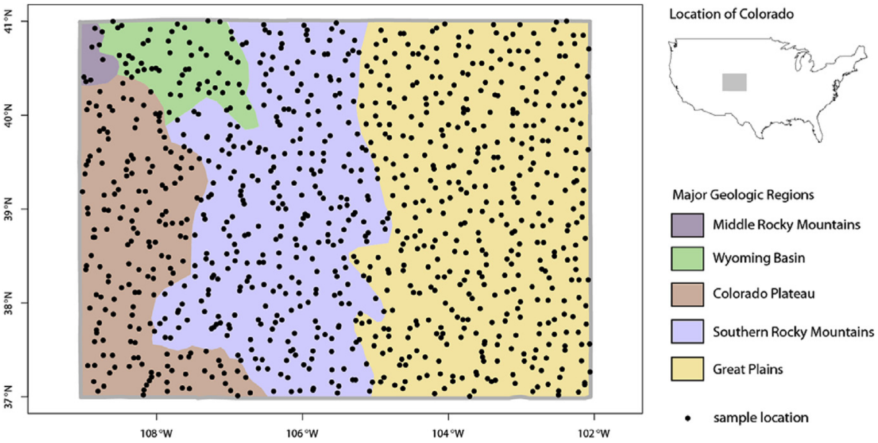


Fig. 1. Major geologic regions within the State of Colorado and sample locations.

3. Clustering procedure

3.1. Preprocessing and analysis

After editing, the data set includes concentrations for 39 elements. These missing concentrations have no effect on the clustering because of the property of subcompositional coherence (Pawlowsky-Glahn et al., 2015, p. 16). Nonetheless, sometimes it is helpful to interpret the clustering results when the scales of its concentrations match those of the measured concentrations. This scaling is easily accomplished if the missing-element concentrations are collectively represented by a single value. For example, if the concentrations of the 39 elements for one field sample sum to 97,634 mg/kg, then the sum of the missing-element concentrations must be 1,000,000−97,634 mg/kg, which equals 902366 mg/kg. Collectively, missing-element concentrations are calculated for all field samples and are appended to the element concentrations, making the effective number 40.

Clustering cannot be applied directly to chemical concentrations because they are a type of compositional data. Such data have two unique properties: they are positive, real-valued numbers and they contain only relative information (Pawlowsky-Glahn et al., 2015, p. 8). The consequence of these two properties is that the algebraic operations for compositional data differ from those for non-compositional (conventional) data (Pawlowsky-Glahn et al., 2015, p. 23−31). To overcome this problem, element concentrations are mathematically transformed with the isometric log-ratio (ilr) transform (Pawlowsky-Glahn et al., 2015, p. 40). The resulting ilr-transformed concentrations are a type of conventional data, which can be analyzed with standard statistical methods (Mateu-Figueras et al., 2011). One consequence of this transformation is a change in the dimension of the data: Before transformation, the data comprise 959 field samples and 40 effective element concentrations; after transformation, the data comprise 959 field samples and 39 ilr coordinates.

If the ilr coordinates are transformed with the robust principal component transformation (Filzmoser et al., 2009), then model-based clustering is more stable than it would be otherwise (Ellefsen et al., 2014). It is assumed that this transformation is similarly beneficial here. Another benefit is that the dimension of the data is significantly reduced. The robust principal component transformation is identical to a conventional principal component transformation (Johnson and Wichern, 2007, 430−439), except that the mean vector and the covariance matrix are calculated in a manner that is relatively insensitive to noise (Rousseeuw and van Driessen, 1999).

The principal components are still ilr coordinates—the only changes are that the origin of the coordinate system has been translated and that the coordinate axes have been rotated. These changes are apparent in the distributions of the principal components (Fig. 2). The distributions are centered at zero because of the coordinate translation, and the spread of the distributions decreases as the component number increases because of the coordinate rotation. The distribution of each component appears unimodal. This lack of multiple modes indicates that the grouping of the principal components into clusters is subtle.

The variances from the diagonal of the covariance matrix are plotted as a function of the principal component number (Fig. 3); in principal component analysis, this plot is called a "scree plot" (Johnson and Wichern, 2007, 444−445). Above each bar in the screen plot is the "cumulative percentage of the total variance." To understand this quantity, consider the variances for just the first three principal components, 1.795, 0.797, and 0.538. The cumulative variances are 1.795, 2.592, and 3.130. These cumulative variances are expressed as percentages of the total variance, which is 5.097. Thus, the cumulative percentages of the total variance are 35.22%, 50.86%, and 61.41%. These cumulative percentages mean that the first component accounts for 35.22% of the total variance, the first and second components for 50.86%, and the first, second, and third components for 61.41%.

A suitable subset of principal components must be selected for the clustering. The selection criterion is that the chosen components must account for most of the variance in the principal components, which is equivalent to most of the information in the geochemical concentrations. Thus, the subset always includes the lower-order components (i.e., components 1, 2, and so on). The relevant issue is determining the last component in the subset. The key issue in selecting this subset involves the signal (i.e., the geochemical information) and the noise. We define noise as errors in the measurements of the element concentrations. After the ilr transformation and the principal component transformation, this noise is spread among all principal components. The ratio of the signal to the noise should be high for the first principal component and should diminish gradually as the component number increases. At some component number, the ratio is small enough that the associated principal component is not contributing useful information to the clustering. A way to estimate this component number is perform clustering for a wide range in the component numbers. In a previous analysis of these data (Ellefsen et al., 2014), we found that the clustering yielded similar results when the largest principal component corresponded to 90 percent or more of the cumulative percentage of total variance. Consequently, for this analysis, we chose of a threshold to 96 percent, which corresponds to 22 principal components. We believe that this threshold is conservative.

In addition to reducing the noise in the data, using only a subset of principal components significantly reduces the amount of data but minimally reduces the amount of information. In this application, using 22 principal components instead of 39 corresponds to a 44% reduction. The consequence is that the amount of computation for the modeling is significantly reduced.

3.2. Finite mixture model

The clustering method is based on a Bayesian finite mixture model (Gelman et al., 2014, p. 519−521). In this model, the data are represented by vector $\boldsymbol{y}_i$. The probability distribution for vector $\boldsymbol{y}_i$ is assumed to be the sum of two multivariate normal probability density functions (pdfs):

$$\boldsymbol{y}_i \sim \lambda N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\lambda) N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \tag{1}$$

Variables λ and $(1-\lambda)$ are the proportions that specify the contribution of each pdf; constraints on their values are presented later. For the first pdf, the mean is vector $\boldsymbol{\mu}_1$, and the covariance matrix is $\boldsymbol{\Sigma}_1$. If the dimension of $\boldsymbol{y}_i$ is represented by D, then the dimension of vector $\boldsymbol{\mu}_1$ is D, and the dimension of matrix is $\boldsymbol{\Sigma}_1$ is $D \times D$. The parameters for the second pdf are defined similarly.

It is helpful to relate the parameters in the finite mixture model to the principal components. Dimension D equals 22 because the first 22 principal components are used (section 3.1). Vector $\boldsymbol{y}_i$ represents the first 22 principal components for field sample i. The distribution of each element in vector $\boldsymbol{y}_i$ is the distribution of the corresponding principal components (Fig. 2). Collectively, these 22 distributions are fit by two multivariate normal distributions that have dimension 22.

The parameters in the finite mixture model are specified using pdfs, which are called "prior pdfs." These prior pdfs should provide some information about the parameters to constrain their possible values (Gelman and King, 1990; Wasserman, 2000); the

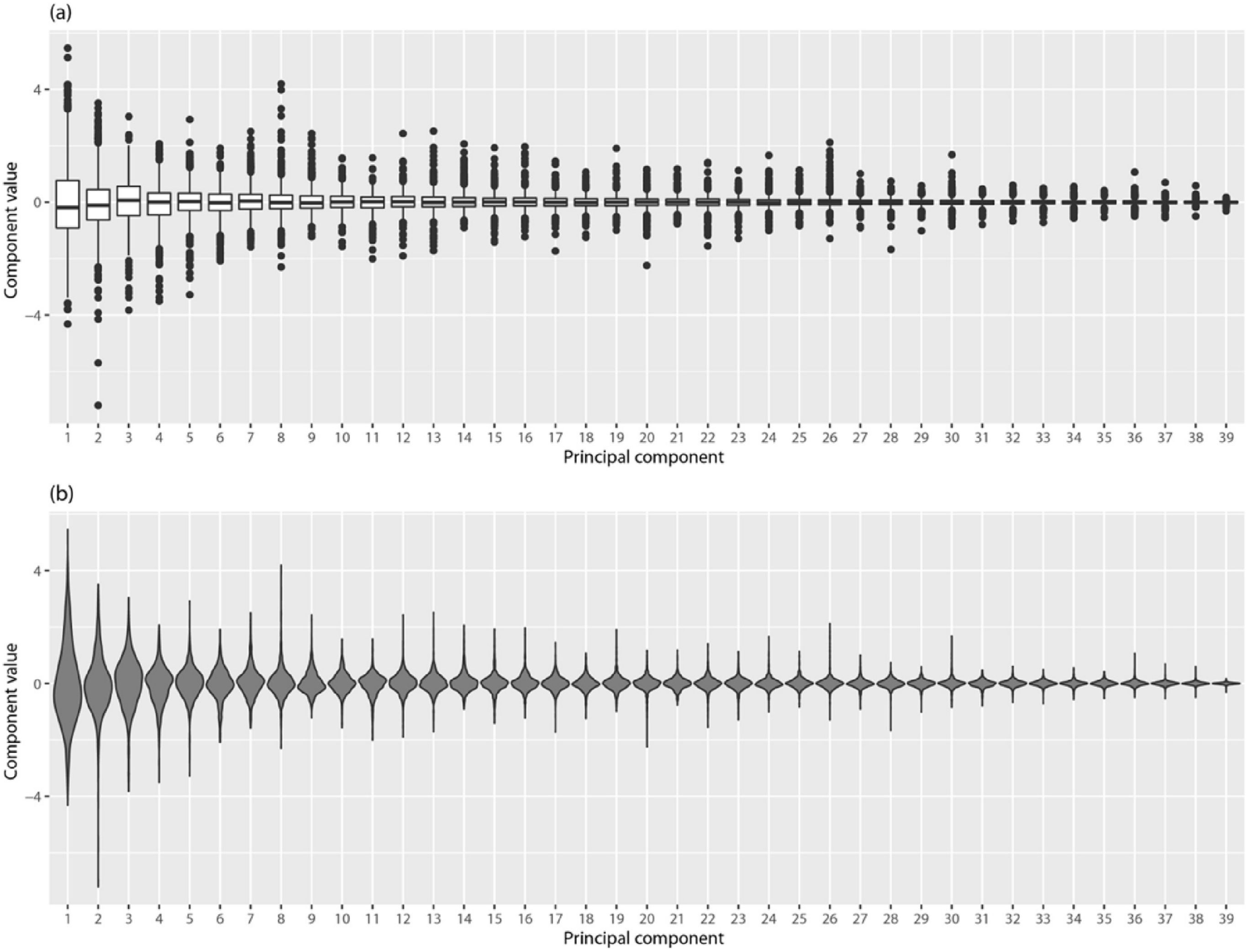


Fig. 2. (a) Boxplots and (b) violinplots of the principal components.

information may be derived from substantive knowledge of the parameters, the data, or both.

To develop a suitable prior pdf for the components of the mean vectors, consider just the distribution of principal component 1 (Fig. 2). This distribution will be represented as the sum of two univariate normal distributions within the finite mixture model (Eq. (1)). The mean for the first univariate distribution $\mu_{1,1}$ must be within the range of principal component 1 (namely, approximately from 4 to 5); the mean probably will be within the largest part of the distribution—that is, approximately between the lower and upper hinges of the boxplot. So, its prior pdf is chosen to be

$$\mu_{1,1} \sim N(0,9)$$

(namely, a normal distribution with a mean of 0 and a variance of 9). Because of the moderately large variance, $\mu_{1,1}$ is only weakly constrained. For similar reasons, the same prior pdf is chosen for the mean for the second univariate distribution $\mu_{2,1}$, as well as for all other elements of the mean vectors. Because the spreads of the principal component distributions decrease as the component number increases (Fig. 2), the variances of the prior pdfs could decrease similarly. However, we have not yet encountered the need for such an elaborate specification of the prior pdfs.

A common prior pdf for a covariance matrix is the inverse-Wishart distribution, because this distribution is conjugate to the multivariate normal distribution, which facilitates Gibbs sampling of the posterior pdf. However, this type of sampling is not used here, so conjugacy is not required. Instead, a simpler prior pdf is used. The covariance matrix for the first pdf in the finite mixture model $\boldsymbol{\Sigma}_1$ is decomposed into a vector of standard deviations $\boldsymbol{\tau}_1$ and a correlation matrix $\boldsymbol{\Omega}_1$:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\tau}_1' \boldsymbol{\Omega}_1 \boldsymbol{\tau}_1 \qquad (2)$$

where the symbol ′ indicates transposition. Consequently, prior pdfs must be specified for $\boldsymbol{\tau}_1$ and $\boldsymbol{\Omega}_1$. Covariance matrix $\boldsymbol{\Sigma}_2$ is similarly decomposed, and prior pdfs must be specified for $\boldsymbol{\tau}_2$ and $\boldsymbol{\Omega}_2$.

To assign a prior pdf for the components of the standard deviation vectors, again consider just the distribution of principal component 1 (Fig. 2). The standard deviation for the first univariate distribution $\tau_{1,1}$ depends upon the associated mean $\mu_{1,1}$, which is unknown. Nonetheless, assume that the mean $\mu_{1,1}$ is between the lower and upper hinges of the boxplot. The univariate distribution must be wide enough to represent the distribution of principal component 1 (Fig. 2). So, the prior pdf for $\tau_{1,1}$ is chosen to be

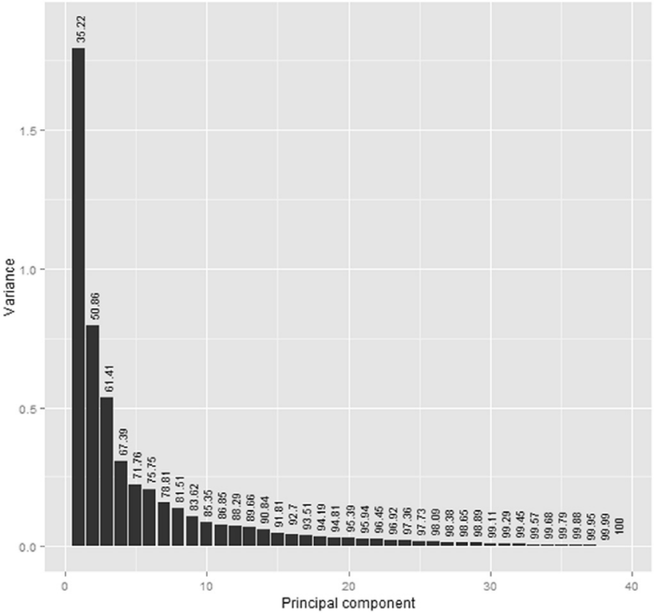


Fig. 3. Scree plot. The number above each bar is "cumulative percentage of the total variance," which is defined in the text.

$$\tau_{1,1} \sim \text{Truncated}Cauchy(0,3)$$

(namely, a Cauchy distribution with a center of 0, a scale parameter of 3, and truncation at 0 so that $\tau_{1,1}$ is always positive). Because of the moderately large scale parameter and the long tails of the distribution, $\tau_{1,1}$ is only weakly constrained. For similar reasons, the same prior pdf is chosen for the standard deviation for the second univariate distribution $\tau_{2,1}$, as well as for all other elements of the standard deviation vectors. Because the spreads of the principal component distributions decrease as the component number increases (Fig. 2), the scale parameter of the prior pdfs could decrease similarly. Again, we have not yet encountered the need for such an elaborate specification of the prior pdfs.

When assigning a prior pdf for the correlation matrices, it is tempting to assume that lack of correlation among the principal components should be reflected in the correlation matrices. This assumption is incorrect: It will be shown later that one cluster of field samples (and hence one subset of the principal components) is primarily associated with the first pdf in the model; another cluster of field samples (and hence another subset of the principal components) is primarily associated with the second pdf. The principal components within a subset are slightly to moderately correlated with one another, and this correlation must be taken into account by using the correlation matrix. So, the prior pdf for $\boldsymbol{\Omega}_1$ is chosen to be

$$\boldsymbol{\Omega}_1 \sim LkjCorr(2)$$

(namely, a LKJ distribution (Lewandowski et al., 2009) with a shape parameter of 2). When the shape parameter is greater than 1, then the LKJ distribution has a mode corresponding to the identity matrix; as the shape parameter increases, the LKJ distribution becomes increasingly concentrated about this mode (Gelman et al., 2014, p. 582). The prior pdf for $\boldsymbol{\Omega}_2$ is identical.

Proportion λ must satisfy several criteria. It must be positive, real-valued, and between 0 and 1. (The last criterion ensures that the sum of λ and $1-\lambda$ equals 1.) The distribution for λ must be symmetric with respect to 1/2, so that neither λ nor $1-\lambda$ is favored. These criteria are satisfied by the beta pdf, when its two shape parameters have equal values. When the two, equal-valued shape parameters are greater than 1, the beta pdf has a symmetric mode at 1/2, which is exactly the desired shape of the prior pdf. Consequently, the prior pdf for λ is chosen to be

$$\lambda \sim Beta(4,\ 4)$$

(namely, a beta distribution for which both shape parameters are 4).

The procedure to estimate the model parameters is described in section "Sampling the posterior pdf" that is within the Supplementary Materials. Additional, important information is in sections "Checking the fit of the model to the data" and "Sensitivity analysis," within the Supplementary Materials.

3.3. *Classifying the field samples*

The interpretation of the geochemical data requires knowing which field samples are associated with each pdf in the finite mixture model (Eq. (1)). This association is specified with conditional probability: The conditional probability that field sample i is associated with the first pdf in the model, given the data $\boldsymbol{y}_i$, is designated p_{i1} and is calculated with

$$p_{i1} = \frac{\lambda N(\boldsymbol{y}_i|\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1)}{\lambda N(\boldsymbol{y}_i|\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1) + (1-\lambda)N(\boldsymbol{y}_i|\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2)} \tag{4}$$

Gelman et al. (2014, p. 539–540). A similar formula may be presented for p_{i2}, the conditional probability that field sample i is associated with the second pdf in the model, given the data $\boldsymbol{y}_i$. However, probability p_{i2} may be calculated with the simple formula $p_{i2} = 1-p_{i1}$. This procedure for specifying association is a type of statistical classification (Hastie et al., 2009, p. 9–22).

Conditional probabilities p_{i1} and p_{i2} are calculated from the samples of the posterior pdf (section "Sampling the posterior pdf" in the Supplementary Materials): Samples of $\boldsymbol{\tau}_1$ and $\boldsymbol{\Omega}_1$ are used to calculate samples of $\boldsymbol{\Sigma}_1$ with equation (2). Similarly, samples of $\boldsymbol{\tau}_2$ and $\boldsymbol{\Omega}_2$ are used to calculate samples of $\boldsymbol{\Sigma}_2$. The samples of λ, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ are used to calculate samples of p_{i1} with equation (4), which, in turn, are used to calculate samples of p_{i2}. The samples of p_{i1} and p_{i2} are summarized by their medians.

3.4. *Interpretation to check model*

An important aspect of checking the model involves interpreting the model results to ensure that they make sense, which means that they are consistent with independent knowledge of geology and geochemistry. The model parameters, which relate to the principal components, cannot be interpreted directly in terms of the geology and geochemistry. Consequently, the model parameters (except the proportion λ) are transformed back to concentrations. To this end, the covariance matrices for each pdf are calculated from the standard deviation vector and correlation matrix using equation (2). Next, the coordinate system, in which the mean vectors and covariance matrices are defined, is rotated and then translated—these operations account for the corresponding operations that were applied during the principal component transformation (section 3.1). Finally, the mean vectors and covariance matrices are transformed from ilr coordinates to concentrations. As a result of this transformation, the mean vector and covariance matrix for each pdf become, respectively, a "compositional center" and a "variation matrix" (Hron and Filzmoser, 2015; Pawlowsky-Glahn et al., 2015, p. 37, 66, and 109).

A compositional center is a vector that comprises the centers for the chemical elements. The centers for the different elements span an enormous range, making it difficult to interpret them. Consequently, the centers are translated—a mathematical operation called "perturbation difference" (Pawlowsky-Glahn et al., 2015, p. 25). The composition that is used for the translation is the sample center for all field samples from the geochemical survey (Pawlowsky-Glahn et al., 2015, p. 66).

Recall that the Monte Carlo sampling generates many samples of the mean vector for each pdf. Consequently, there are many samples of the compositional center for each pdf. The translation is applied to all Monte Carlo samples of the compositional centers. The resulting distributions for the chemical elements are shown in Fig. 4a and b for the first and the second modes, respectively. To compare the distributions, it is helpful to have a reference, which is chosen to be the barycenter (Pawlowsky-Glahn et al., 2015, p. 25). The barycenter is the translated sample center and has the same value for all chemical elements.

Consider, for example, the two distributions for yttrium (Y) for the first mode (Fig. 4a). The credible interval is entirely above the barycenter for pdf 2 but is entirely below the barycenter for pdf 1. The interpretation of this relation is that, for those field samples associated with pdf 2, yttrium is relatively enriched compared to the entire survey area. Conversely, for those field samples associated with pdf 1, yttrium is relatively depleted compared to the entire survey area. For many chemical elements, the relative enrichments and depletions differ for the two modes (Fig. 4a and b).

The Monte Carlo sampling generates many samples of the covariance matrix for each pdf, so there are many samples of the variation matrix for each pdf. It is difficult to show the distribution for each element in the variation matrix, so the distribution is summarized by its median. The variation matrix with these medians is symmetric with respect to its diagonal. To minimize redundant information, the upper triangle of the variation matrix for pdf 1 and the lower triangle of the variation matrix for pdf 2 are combined into a single matrix. The range of the variances in the combined matrix is often large, so the variances are scaled by the square root. Consequently, the matrix elements represent standard deviations. The combined, scaled variation matrices are shown in Fig. 4c and d for the first and second modes, respectively.

In Fig. 4c, practically all pixels in the upper triangle are greater than the corresponding pixels in the lower triangle; that is, the scaled variances for pdf 1 are greater than the corresponding scaled variances for pdf 2. The interpretation of this result is that the geochemistry for those field samples associated with pdf 1 are much more variable than the geochemistry for those field samples associated with pdf 2. The combined, scaled variation matrices for the first and second modes are similar (Fig. 4c and d).

The only model parameter that has not been discussed yet is the proportion λ. Instead, we discuss the conditional probabilities, which are calculated from the proportion (section 3.3) and which are more helpful to the interpretation than the proportion is. Recall that the conditional probabilities indicate the association between a field sample and either pdf 1 or pdf 2. It is helpful to see the associations for all field samples as a map. To this end, the conditional probabilities p_{i1} and p_{i2} for field sample i, for which there are many Monte Carlo samples, are summarized by their medians $\tilde{p}_{i1}$ and $\tilde{p}_{i2}$. If $0.5 < \tilde{p}_{i1} \leq 1.0$, then field sample i is associated with pdf 1. All field samples that satisfy this criterion constitute the cluster for pdf 1, which we call "cluster 1." It is helpful to indicate the strength of the association within the cluster. Consequently, the interval between 0.5 and 1.0 is divided into two parts: $0.5 < \tilde{p}_{i1} \leq 0.9$ and $0.9 < \tilde{p}_{i1} \leq 1.0$, for which the strength of association is deemed moderate and strong, respectively. An analogous procedure defines the cluster for pdf 2, which we call "cluster 2."

Clusters 1 and 2, including the strength of association within the clusters, are plotted as maps (Fig. 4e and f for the first and second modes, respectively) showing the location of the field samples for each of the two clusters in relation to the five major geologic regions for Colorado (Fig. 1). It is readily apparent that the clusters show a distinct spatial correlation with three of these geologic regions. Within the Great Plains, almost all field samples are from cluster 2. Within the Southern Rocky Mountains, almost all field samples are from cluster 1. Within the Wyoming Basin, most field samples are from cluster 2. Within the Colorado Plateau, there is a mixture of field samples from clusters 1 and 2. Within the Middle Rocky Mountains, there are too few samples to make any inferences.

Within the Southern Rocky Mountains geologic region, the soil parent material is most commonly the underlying bedrock, which includes plutonic and volcanic igneous rocks of felsic to intermediate composition. Another soil parent material is the sediments derived from the underlying bedrock. In addition, this area includes the Colorado Mineral Belt, which contains deposits of Pb, Cu, Zn, Mo, Au, and Ag. As shown in Fig. 4a, cluster 1 is characterized by a general enrichment of Zn, In, Mn, Mo, Cu, Pb, Bi, and S. These elements are commonly associated with the base- and precious-metal deposits found within the Southern Rocky Mountains. In addition, cluster 1 is generally enriched in Fe, Sc, Co, Mg, P, and V. These elements are commonly associated with ferromagnesian minerals within igneous rocks of intermediate to mafic composition as found in this geologic region. Within the Great Plains geologic region, the soil parent material is mostly sedimentary deposits including sandstone, gravel, alluvium, colluvium, claystone, mudstone, and shale. Many of the soils in this geologic region also contain a significant eolian component composed of mostly quartz and potassium feldspar. Cluster 2 is generally enriched in K, Rb, Ba, Tl (Fig. 4a), all closely associated with potassium feldspar. In addition, cluster 2 is also enriched in Y, Nb, La, Ce, Th, and Be. These elements are commonly associated with felsic rocks from which many of the sedimentary rocks in the Great Plains were derived.

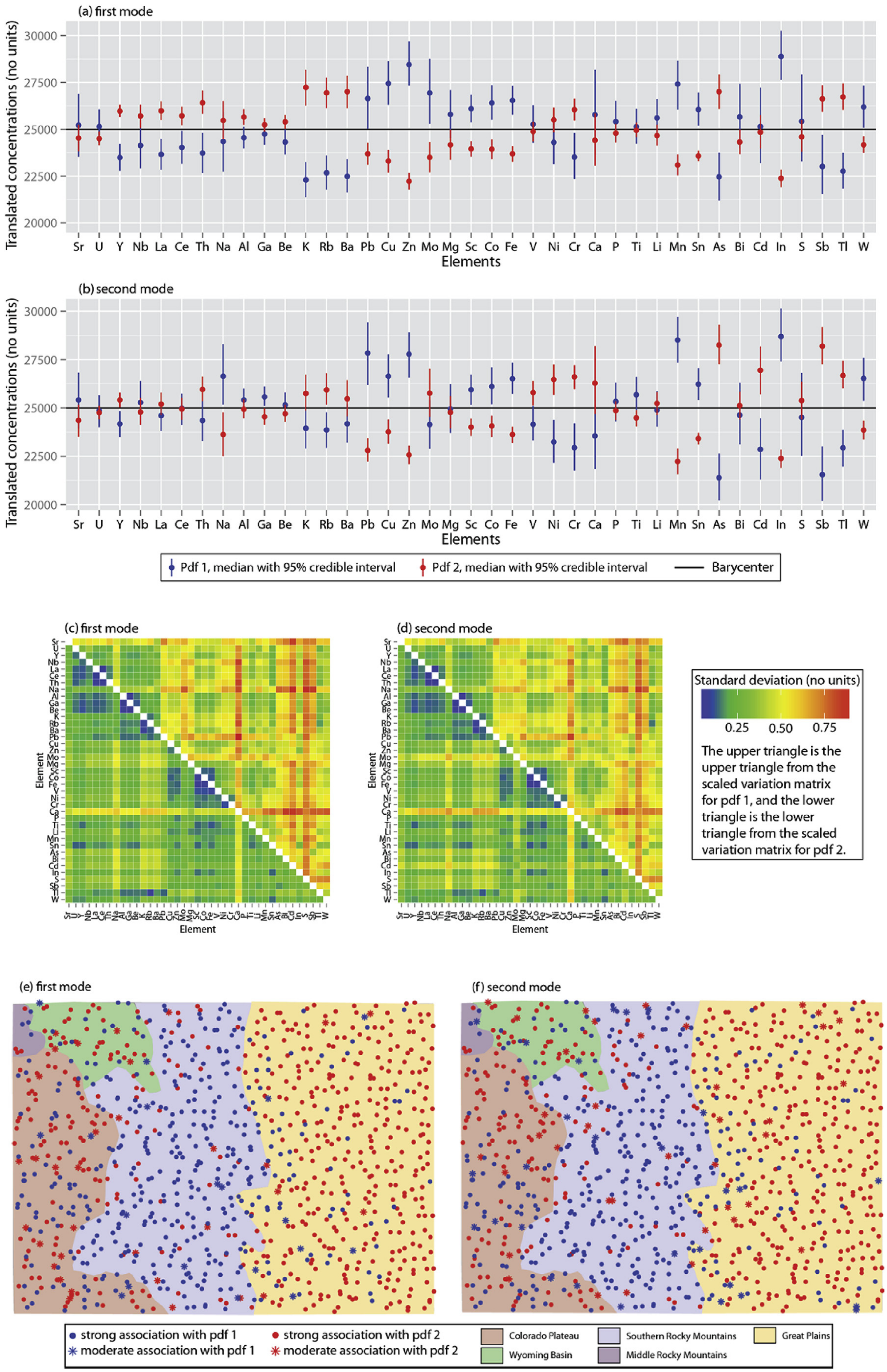
The Southern Rocky Mountains region (Fig. 1) is more geologically heterogeneous than the other four geologic regions are. Consequently, the geochemistry within this region is expected to vary more than it does in the other four regions. This expectation is consistent with the combined, scaled variation matrices for both modes (Fig. 4c and d).

In summary, the cluster maps are consistent with the geologic regions, the relative enrichments of the translated concentrations are consistent with the geology (especially for the second mode), and the combined, scaled variation matrices are consistent with the geology. Therefore, we infer that the finite mixture model fits our independent knowledge of the geology. Because of this inference and the corresponding inference regarding the fit to the data (section "Checking the fit of the model to the data" in the Supplementary Materials), we are confident that the finite mixture model can be used to interpret the geochemical data.

3.5. Partitioning the field samples

In section 3.4, the translated concentrations for the second mode are interpreted to be more consistent with independent knowledge than the translated concentrations for the first mode are. Consequently, the results for the second mode are selected for further analysis and sub-clustering.

The value of the likelihood function indicates how well model predictions fit the data (namely, the principal components). In this case, the value of the likelihood function for the second mode is lower than that for the first mode (Figure S1d in Supplementary Materials). That is, the fit for the second mode is not as good as the fit for the first mode, so the selection of the second mode may seem



The upper triangle is the upper triangle from the scaled variation matrix for pdf 1, and the lower triangle is the lower triangle from the scaled variation matrix for pdf 2.

strange. However, we believe that the primary criterion for selecting a mode for further analysis and sub-clustering should be the consistency with the independent knowledge, not the value of the likelihood function.

A simple way to partition the field samples is suggested by the cluster map: Those field samples that are strongly associated with cluster 1 become a new data set, which is called “data subset 1.” Likewise, those field samples that are strongly associated with cluster 2 become another new data set, which is called “data subset 2.” Data subsets 1 and 2 comprise, respectively, 407 and 486 field samples. Consequently, of the 959 field samples in the entire data set (section 2.2), 66 field samples did not satisfy the partitioning criterion and are omitted from further analysis. Although this number is relatively small compared to 959, it could be made even smaller by decreasing the threshold that specifies strong association.

3.6. Manual hierarchical clustering

The next step is to apply the clustering procedure to the two data subsets—the results will be the next level of the hierarchy. The results from only data subset 2 are presented because they are enough to show the benefits of manual hierarchical cluster.

Nineteen principal components account for 96.18% of the total variance, so 19 principal components are used in the clustering procedure. The point statistics for selected model parameters indicate that only one mode from the posterior pdf was sampled, and the Monte Carlo sampling converged. The checks regarding the model fit to the data indicate that this fit is very good. The checks regarding the model fit to geologic knowledge are summarized in Fig. 5. The composition that is used for the translation (Fig. 5a) is sample center for the entire geochemical survey.

In cluster 1, the enriched elements include Mo, As, Cd, Sb, S, Bi, Li, Cr, Ni, V, Cu, Zn, Sc, Co, Fe, and P (Fig. 5a). These elements are all commonly enriched in shales and other fine-grained marine sedimentary rocks such as mudstone and claystone. This association is apparent in the map of the cluster locations (Fig. 5c). In cluster 2, the enriched elements include K, Rb, Ba, Th, Na, Al, Tl, Pb, Sr, U, Y, Nb, La, Ce, Ga, and Be (Fig. 5a). Most of these elements are commonly associated with potassium feldspars or felsic rocks. Again, this association is apparent in the map of cluster locations (Fig. 5c)—the field samples are primarily found in areas underlain by sandstone, gravel, alluvium, and colluvium. Also in these areas, many of the soils contain a significant eolian component, and the above element association is consistent with the composition of the eolian and siliciclastic parent material. The combined, scaled variation matrix (Fig. 5b) shows that the concentrations for cluster 1 are slightly more variable than those for cluster 2. In summary, the cluster map is consistent with the geologic formations (i.e., soil parent materials), and the relative enrichments of the translated concentrations are consistent with the geology and geochemistry. Therefore, we infer that the finite mixture model fits our independent knowledge of the geology.

Consider the difference between the results in Fig. 4 (namely, the first level of the hierarchy) and in Fig. 5 (namely, the second level of the hierarchy). The results at the first level are interpreted in terms of the geologic regions—the spatial scale is large. In contrast, the results at the second level are interpreted in terms of geologic formations within geologic regions—the spatial scale is moderately large. This difference is a significant advantage of the manual hierarchical clustering; each level of the hierarchy highlights geologic and geochemical processes occurring at different spatial scales.

4. Discussion

4.1. Comparison to K-means clustering

To assess the value of the Bayesian formulation of finite mixture modeling, its clustering results should be compared to clustering results from other methods. Because there are many other methods (Johnson and Wichern, 2007, p. 671–706), a comprehensive comparison is infeasible for this article. However, a comparison to one method is feasible. The comparison is partial because the clustering results in section 3.6 pertain to only dataset 2.

The chosen method is K-means clustering (Hartigan and Wong, 1979). The reason for choosing this method is that it is used to cluster geochemical data (Reimann et al., 2008, p. 239–240), as well as other types of data. The data for the K-means clustering are the 22 principal components (section 3.1), which are the very same data for the finite mixture model. The number of K-means clusters is four because four clusters were found in the original processing of the data (Ellefsen et al., 2014) and because four clusters correspond to the second level of the hierarchy (section 3.6). The K-means clustering assigns the field samples to the most appropriate cluster and, it estimates four means for the four clusters. The four means, which relate to the principal components, are back transformed to concentrations and then translated to facilitate their interpretation; the procedure is identical to that applied to the results from the finite mixture model (section 3.4).

The translated compositional centers appear in Fig. 6a. The centers for clusters 1 and 2 are similar to the centers for pdfs 1 and 2 from the finite mixture modeling (Fig. 5a). Consequently, their interpretation is the same. The locations of the field samples for cluster 1 and cluster 2 are plotted on the map of the aggregated geologic formations (Fig. 6b). The locations for cluster 1 generally correspond to the shale, claystone, and mudstone. Likewise, locations for cluster 2 generally correspond to the sandstone, gravel, alluvium, and so on. Thus, these map associations are consistent with interpretation of the translated compositional centers.

The cluster maps in Figs. 5c and 6b are similar but not identical. In other words, the field samples are classified differently. For example, in the southeastern corner of the state, many field samples are associated with pdf 2 (Fig. 5c) but relatively few with cluster 2 (Fig. 6b). Another example is in the north-central part of the state—this region does not have the rock types in the map of aggregated geologic formations, so it is white. In this region there is only 1 field sample associated with pdf 2 (Fig. 5c) but 27 field samples associated with cluster 2 (Fig. 6b).

Of course, different clustering methods have different advantages and disadvantages. The advantages of the K-means clustering include speed and simplicity. One disadvantage is the lack of information regarding uncertainty in the translated compositional centers and in the assignments to the clusters. Another disadvantage is there are no variation matrices. These matrices are important because they provide information that helps geochemists interpret clustering results (section 3.4). These disadvantages, as well as our concerns regarding the classification of the field samples, cause us to prefer manual hierarchical clustering with a Bayesian finite mixture model.

Fig. 4. Fit of model to independent geologic knowledge for the entire data set. (a and b) Translated compositional centers for the first and the second modes. (c and d) Combined, scaled variation matrices for the first and the second modes. (e and f) Cluster maps (of the State of Colorado) for the first and second modes. Area and scale are the same as in Fig. 1; latitudes and longitudes have been omitted to improve clarity.

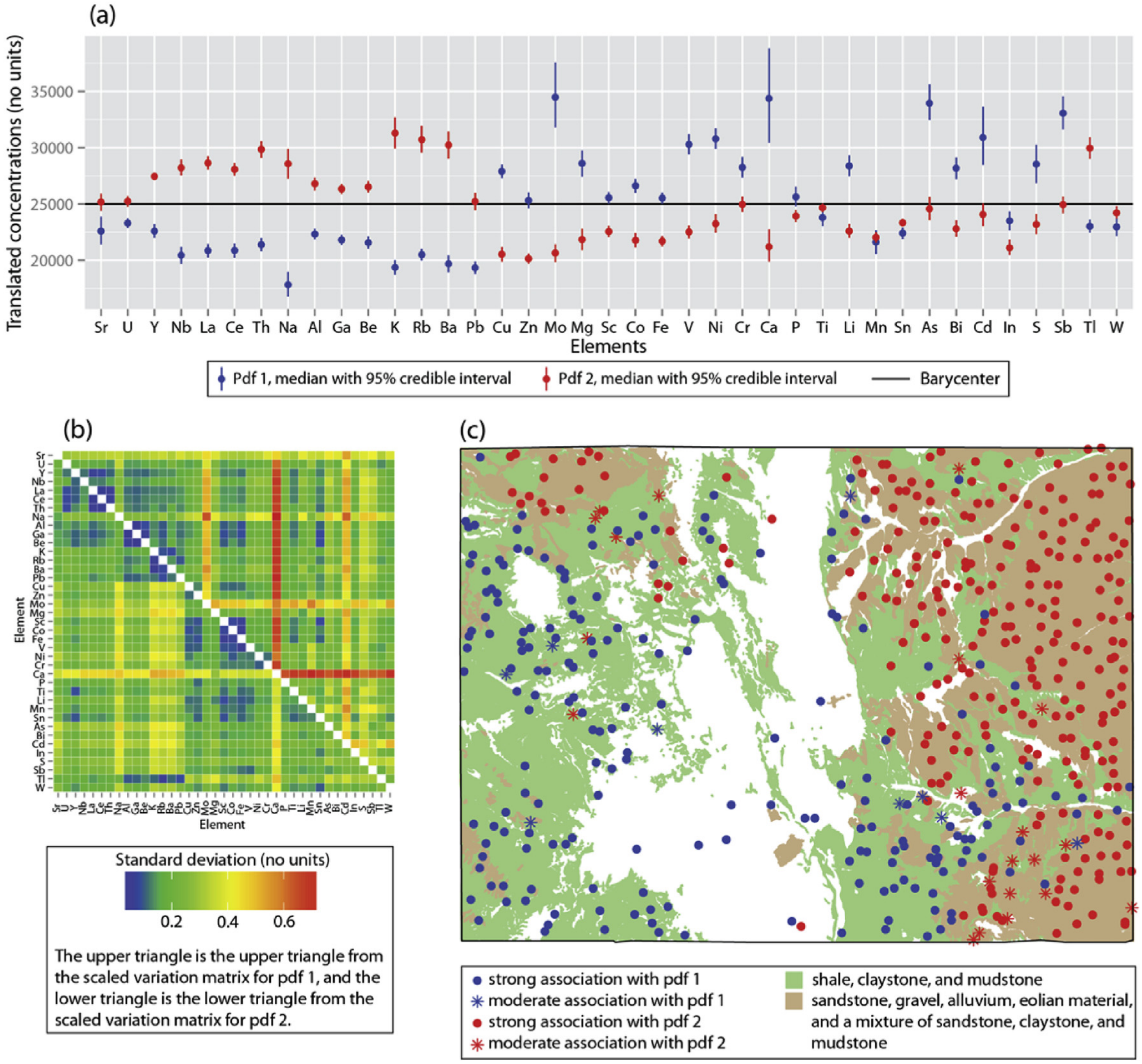


Fig. 5. Fit of model to independent geologic knowledge for data subset 2. (a) Translated compositional centers. (b) Combined, scaled variation matrix. (c) Cluster map (of the State of Colorado), overlying map of aggregated geologic formations. Area and scale are the same as in Fig. 1; latitudes and longitudes have been omitted from the map to improve clarity.

4.2. Miscellaneous issues

An assumption of the finite mixture model is that each vector $\boldsymbol{y}_i$ is probabilistically independent of the other vectors. This assumption is not satisfied because the field samples that are close together are spatially correlated. Currently, we do not know how the finite mixture model is affected by violating this assumption. There has been some related research by Gibergans-Baguena et al. (2011), who accounted for spatial correlation in rainfall frequencies, using a Mahalanobis distance calculated from multivariate variograms. Despite this research advance, there are some outstanding research issues. For example, one issue is measuring distances between sample locations in non-contiguous regions. Another issue is incorporating the spatial correlation in the finite mixture model for each principal component. Such research issues must be resolved before spatial correlation can be incorporated into the finite mixture model.

There are different ways to perform the ilr transformation (Pawlowsky-Glahn et al., 2015, p. 38–42). Data that have undergone different ilr transformations (but not the principal component transformation) will yield different clustering results. We suspect that the differences will be small, but we have not investigated this issue. Such different clustering results are very undesirable, but this problem can be overcome by applying the principal component transformation. That is, whatever ilr transformation is applied, the subsequent principal component transformation will always yield the same transformed data, to within the finite precision of the computer. Consequently, the clustering results will be the same. In this way, the principal component transformation contributes to reproducibility.

A pertinent topic for future research is evaluating how well different methods are able to cluster different data sets; this research would build on similar previous research conducted by Templ et al. (2008). One data set should be geochemical concentrations of field samples; another data set should be geochemical concentrations simulated on a computer. The advantage of the field data set is that it includes complexities that are difficult to simulate; the advantage of the simulated data set is that the clustering is known—this known clustering can be used to evaluate the clustering results from the different methods. Because there are many different methods and many different ways to prepare the data for clustering, this research would require a lot of work by many different investigators.

Although this article focuses on clustering using finite mixture

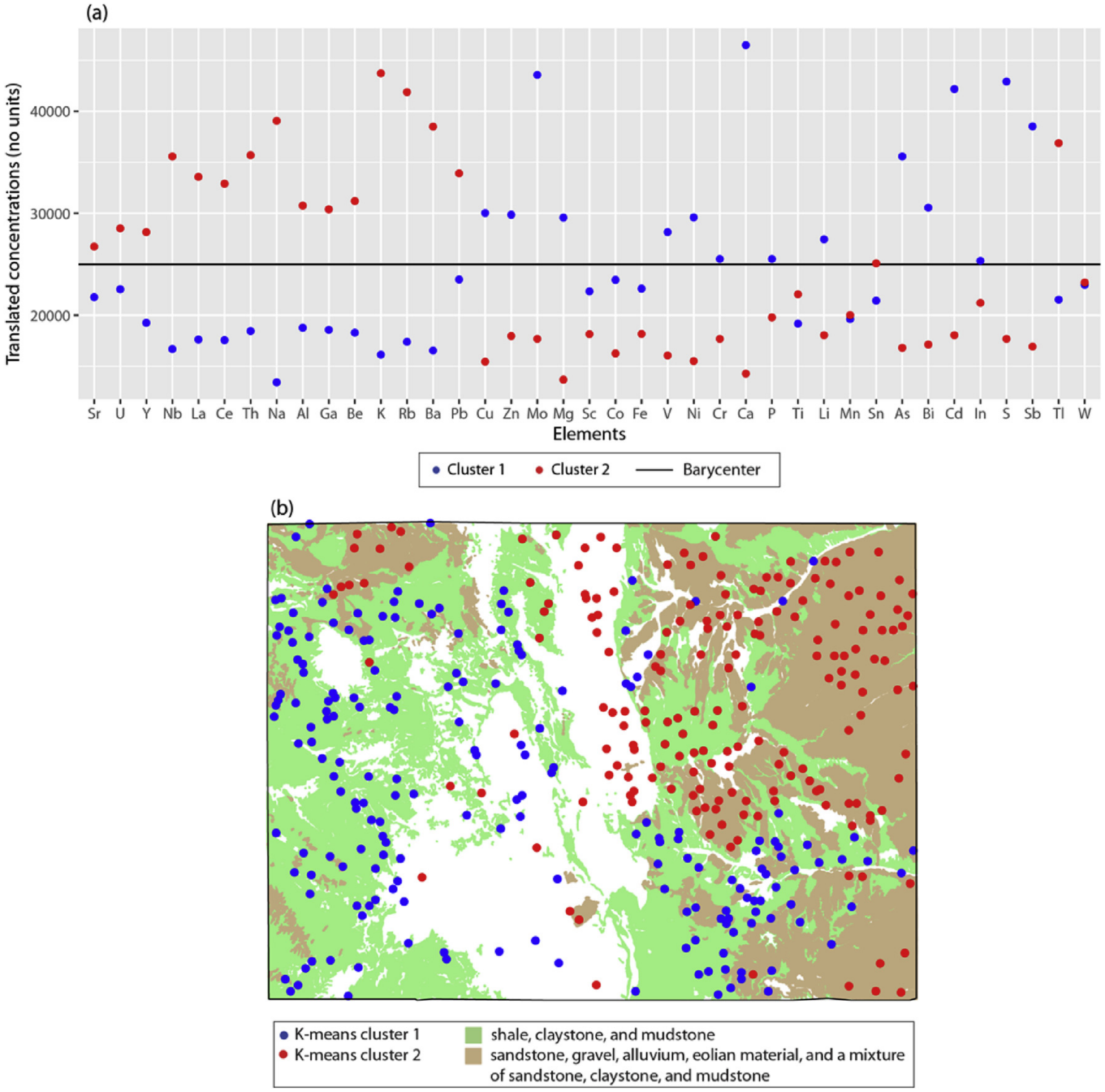


Fig. 6. Results of K-means clustering for clusters 1 and 2. (a) Translated compositional centers. (b) Cluster map (of the State of Colorado) overlying map of aggregated geologic formations. Area and scale are the same as in Fig. 1; latitudes and longitudes have been omitted from the map to improve clarity.

modeling, recall that techniques for compositional data analysis (Pawlowsky-Glahn et al., 2015) are part of the clustering procedure. The geochemical data are mapped with the ilr transformation to coordinates where the clustering is performed. Using the inverse ilr transformation, the results are mapped back to concentrations where they are interpreted. This interpretation requires statistics and algebraic operations that are part of compositional data analysis—the compositional center, the variation matrix, and perturbation difference. Thus, clustering using finite mixture modeling is possible only because of the modern techniques for compositional data analysis.

The Supplementary Materials includes discussions of some details regarding manual hierarchical clustering.

5. Conclusions

Manual hierarchical clustering with a Bayesian finite mixture model has both disadvantages and advantages. For large data sets, the clustering at just one level of the hierarchy requires several hours on a workstation with multiple processing cores. However, compared to the amount of time required to collect the samples and chemically analyze them, a few hours is negligible. The method requires some knowledge of Monte Carlo sampling, especially knowledge about convergence. Further investigation is needed for some details of the procedure, including the appropriate stopping point. The outstanding advantage of manual, hierarchical clustering is that it reveals geochemical processes occurring at different spatial scales—this information is crucial to the interpretation. The procedure quantifies the uncertainty in the model parameters, which is needed for the interpretation. Lastly, the clustering procedure overcomes the significant problems due to multiple modes in the posterior pdf.

We believe that the disadvantages are relatively minor compared to the advantages. Consequently, we are confident that the clustering procedure is useful for interpreting regional geochemical data. Furthermore, the method is directly applicable to related types of earth-science data, such as regional soil surveys of mineral concentrations.

Acknowledgments and disclaimers

This work was funded by the Geochemical Landscapes Project within the Mineral Resources Program of the U.S. Geological Survey. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government. J. Morrison and two anonymous reviewers provide helpful suggestions that improved this manuscript.

Software and reproducibility

The hierarchical clustering that is presented in this manuscript was carried out using a software package called "GcClust," which is written with the R statistical programming language. This public-domain package is available at Ellefsen and Smith (2016). The package includes the geochemical data that were processed for this manuscript. The accompanying software documentation includes the R language scripts that geochemists can execute to reproduce the results in this manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apgeochem.2016.05.016.

References

Caritat, P. de, Cooper, M., 2011. National Geochemical Survey of Australia: the Geochemical Atlas of Australia, vol. 2. Geoscience Australia, Record 2011/20.

Ellefsen, K.J., Smith, D.B., Horton, J.D., 2014. A modified procedure for mixture-model clustering of regional geochemical data: Appl. Geochem. 51, 315–326. http://dx.doi.org/10.1016/j.apgeochem.2014.10.011.

Ellefsen, K.J., Smith, D.B., 2016. User's Guide for GcClust—an R Package for Clustering of Regional Geochemical Data. U.S. Geological Survey report Techniques and Methods 7–C13, p. 21p. http://dx.doi.org/10.3133/tm7c13.

Fauth, H., Hindel, R., Siewers, U., Zinner, J., 1985. Geochemischer Atlas Bundesrepublik Deutschland. Bundesanstalt für Geowissenschaften und Rohstoffe and Schweizerbart'sche Verlagsbuchhandlung.

Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics 20, 621–632.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97 (458), 611–631.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data Analysis, third ed. CRC Press.

Gelman, A., King, G., 1990. Estimating the electoral consequences of legislative redistricting. J. Am. Stat. Assoc. 85 (410), 274–282.

Gibergans-Baguena, J., Ortego, M.I., Tolosana-Delgado, R., 2011. Pluviometric regionalization of Catalunya–A compositional data methodology. In: Egozcue, J.J., Tolosana-Delgado, R., Ortego, M.I. (Eds.), Compositional Data Analysis Workshop – CoDaWork '11, Proceedings: Saint Feliu de Guixols, Girona. http://congress.cimne.com/codawork11/frontal/Home.asp. last accessed April 2016.

Grunsky, E.C., 2010. The interpretation of geochemical survey data: geochemistry–Exploration,. Environ. Anal. 10, 27–74. http://dx.doi.org/10.1144/1467-7873/09-210.

Hartigan, J.A., Wong, M.A., 1979. A K-means clustering algorithm. Appl. Stat. 28, 100–108.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Science+Business Media, LLC.

Hron, K., Filzmoser, P., 2015. Exploring compositional data with the robust compositional biplot. In: Carpita, M., Brentari, E., Qannari, M. (Eds.), Advances in Latent Variables–Methods, Models and Applications. Springer, p, pp. 219–226.

Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis. Pearson Education, Inc.

Lewandowski, D., Kurowicka, D., Joe, J., 2009. Generating random correlation matrix based on vines and extended onion method. J. Multivar. Anal. 100, 1989–2001. http://dx.doi.org/10.1016/j.jmva.2009.04.008.

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The principle of working on coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis – Theory and Applications. John Wiley and Sons, Ltd., pp. 31–42

McGrath, S.P., Loveland, P.J., 1992. The Soil Geochemical Atlas of England and Wales. Blackie Academic Professional, Glasgow, U.K.

Morrison, J.M., Goldhaber, M.B., Ellefsen, K.J., Mills, C.T., 2011. Cluster analysis of a regional-scale geochemical dataset in northern California. Appl. Geochem. 26, S105–S107. http://dx.doi.org/10.1016/j.apgeochem.2011.03.041.

Palarea-Albaladejo, J., Martin-Fernandez, J.A., Buccianti, A., 2014. Compositional methods for estimating elemental concentrations below the limit of detection in practice using R. J. Geochem. Explor. 141, 71–77. http://dx.doi.org/10.1016/j.gexplo.2013.09.003.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. John Wiley and Sons, Ltd.

Reimann, C., Siewers, U., Tarvainen, T., Bityukova, L., Eriksson, J., Gilucis, A., Gregorauskiene, V., Lukashev, V.K., Matinian, N.N., Pasieczna, A., 2003. Agricultural Soils of Northern Europe: a Geochemical Atlas. Schweizerbart'sche Verlagsbuchhandlung.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained–Applied Environmental Statistics with R. John Wiley & Sons, Ltd.

Rousseeuw, P.J., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator: Technometrics 41, 212–223.

(Geological Survey of Finland, Espoo). In: Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., DeVos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P.J., Olsson, S., Ottesen, R.T., Petersell, V., Plant, J.A., Reeder, S., Salpeteur, I., Sandström, H., Siewers, U., Steenfeldt, A., Tarvainen, T. (Eds.), 2005. FOREGS Geochemical Atlas of Europe, Part 1–Background Information, Methodology and Maps.

Smith, D.B., Ellefsen, K.J., Kilburn, J.E., 2010. Geochemical Data for Colorado Soils—Results from the 2006 State-scale Geochemical Survey. U.S. Geological Survey, Data Series 520, p. 9p. available at. http://pubs.usgs.gov/ds/520/. last accessed October 2015.

Smith, D.B., Cannon, W.F., Woodruff, L.G., Solano, Federico, Kilburn, J.E., Fey, D.L., 2013. Geochemical and Mineralogical Data for Soils of the Conterminous United States. U.S. Geological Survey Data Series 801. available at. http://pubs.usgs.gov/ds/801/. last accessed 22.02.14.

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data–Problems and possibilities. Appl. Geochem. 23, 2198–2213. http://dx.doi.org/10.1016/j.apgeochem.2008.03.004.

Thalmann, F., Schermann, O., Schroll, E., Hausberger, G., 1989. Geochemischeratlas der Republik Österreich, scale 1:1,000,000 (Geologische Bundesanstalt).

Tweto, O., compiler, 1979. Geologic Map of Colorado: U.S. Geological Survey, Scale 1: 500,000.

Wasserman, L., 2000. Asymptotic inference for mixture models using data-dependent priors. J. R. Stat. Soc. B 62 (1), 159–180.

Webb, J.S., Thornton, I., Thompson, M., Howarth, R.J., Lowenstein, P.L., 1978. The Wolfson Geochemical Atlas of England and Wales. Oxford University Press.