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.

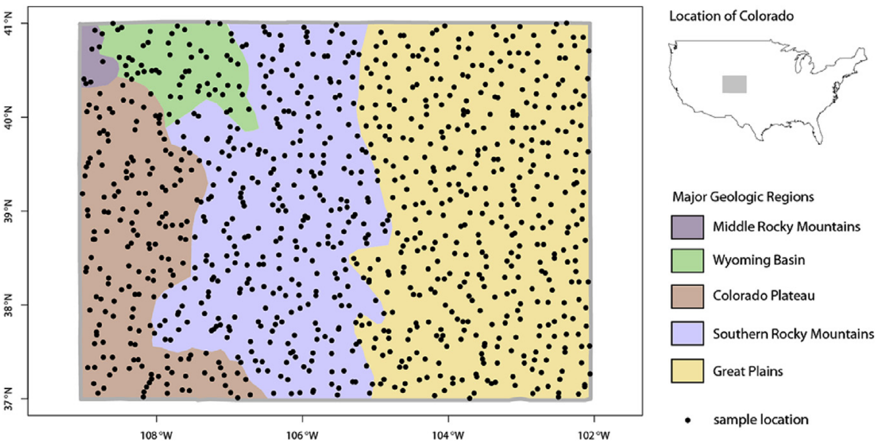

Fig. 1. Major geologic regions within the State of Colorado and sample locations.