Computers & Geosciences 72 (2014) 122–133

Contents lists available at ScienceDirect

Computers & Geosciences

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

Fuzzy hierarchical cross-clustering of data from abandoned mine site contaminated with heavy metals

A. Pourjabbar [a,*], C. Sârbu [b], K. Kostarelos [c], J.W. Einax [d], G. Büchel [a]

[a] Institute of Geosciences, Friedrich Schiller University of Jena, Burgweg 11, 07749 Jena, Germany
[b] Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, Arany Janos Str. No. 11, RO-400028 Cluj Napoca, Romania
[c] Subsurface Research Laboratory, University of Cyprus, Kallipoleos Str. 75, 1678 Nicosia, Cyprus
[d] Institute of Inorganic and Analytical Chemistry, Friedrich Schiller University of Jena, Lessingstr 8, 07743 Jena, Germany

ARTICLE INFO

Article history:
Received 8 October 2013
Received in revised form
30 June 2014
Accepted 1 July 2014
Available online 10 July 2014

Keywords:
Fuzzy clustering
Fuzzy cross-clustering
Acid mine drainage
Abandoned mine
Heavy metals

ABSTRACT

The characteristics of pore water and slate samples are critically analyzed using fuzzy hierarchical cross-clustering statistical techniques. The main aim of this study was to investigate the source of contamination near an abandoned uranium mine in Germany. The mining activities were abandoned in 1990 the site was closed, and the surrounding area was remediated. However, heavy metal contamination is still detectable in water, soil and plants today. Hence, investigating the source of the current contamination is an important task. In order to achieve the goal, results from chemical analysis of both pore water samples and leachates from slate samples were initially analyzed using hard (classical) hierarchical clustering algorithms that did not provide meaningful results. By using two fuzzy clustering algorithms, Fuzzy Divisive Hierarchical Clustering (FDHC) and Fuzzy Hierarchical Cross-Clustering (FHCC), a relationship between the leachate from Ordovician-Silurian slate samples (10 samples collected from the test site and the surrounding area) and pore water samples (53 samples collected from 3 locations within the test site at 3 depths over the course of 4 years) was identified. The leachate data formed a cluster which was statistically similar to the cluster formed by the pore water samples collected from two of three locations. In addition, the fuzzy cross-clustering approach allowed for the identification of the characteristics (qualitative and quantitative) responsible for the observed similarities between all the samples. We conclude that the fuzzy algorithms were a better tool for the analysis and interpretation of geological/hydrogeological data where the data sets have an inherent vagueness/uncertainty.

1. Introduction

The main focus of the current study is the determination of the contamination sources of an area in close proximity to a former leaching heap, the Gessenhalde, of an abounded mine in east Thuringia, Germany. This leaching heap was composed of Silurian slates. Initially, sulfuric acid and later acid mine drainage (AMD) were used to leach uranium from the excavated slates. When the mine was closed, up to 10 m of the leaching heap were removed and placed within the open pit mine (Lichtenberg) in order to remediate the area. However, the surrounding area is still contaminated by heavy metals. Special conditions at this area, namely low pH and high concentration of heavy metals, lead to a high potential for a contaminated environment. The degree of contamination was measured and reported for various media such as groundwater, pore water, surface water, soil, and plants (Carlsson and Büchel, 2005; Grawunder, 2010; Hafeburg, 2007; Horn, 2003; Lonschinski, 2009; Lorenz, 2009; Mirgorodsky et al., 2010). Pourjabbar (2012) determined that the remediated top soil has become contaminated by fluctuating contaminated groundwater levels during high precipitation events. However, due to the large quantity of the slate material still presents in the surrounding area the role of slates as a potential source of contamination needs to be determined. The hypothesis of this study is that leachates from the Ordovician and Silurian slates are the source of the environmental contamination (heavy metals including rare earth elements (REE)) in this area. In order to determine whether a relationship exists between these slates and the contamination, chemical analyses of the pore water and leachate from the slate samples (leached with water) were studied using statistical methods. Previous studies performed on pore water and leachate from slate samples collected from the study area focused on the

* Corresponding author. Tel.: +357 2289 3527.
E-mail address: pourjabbar.anahita@ucy.ac.cy (A. Pourjabbar).

http://dx.doi.org/10.1016/j.cageo.2014.07.004
0098-3004/

chemical analyses (Lorenz, 2009; Pasalic, 2011; Wagner, 2010). However, there was no effort made to apply clustering techniques in order to reveal a relationship between the data sets. With regard to geo-environmental studies, cluster analyses are used in order to find and make visible structures within observations and variables. Templ et al. (2006) suggested using cluster analysis as an "exploratory data analysis tool" to better understand the multivariate behavior of a data set.

Other statistical methods which are used to identify the relationship between parameters include factor analysis and principal component analysis. Factor analysis is a technique which reduces a large set of data into a smaller dimension or component. This method is used when variables are highly correlated. Principal component analysis is applied as an initial step of any multivariate analysis to obtain a 'first look' at the structure of the data (Wold et al., 1987), to help to identify outliers, to delineate clusters, *etc*. The advantage of using this technique for a large data set is the ability to reduce the dimension of the data set without losing the information. Considering the goal of this study, clustering analysis is a relevant method to be applied.

In general, clustering techniques can be used to find relationships within a set of observations (*e.g.*, samples) and variables (*e.g.*, chemical composition), which can be presented in a visual format. The principal aim of clustering is to partition multivariate observations into a number of meaningful, homogeneous groups that share common characteristics; *i.e*., the greatest similarities within the cluster, and greatest dissimilarities between the clusters. A *meaningful* cluster is one that conforms or elucidates geological/geochemical evidence for example it is well-known that REEs have similar geochemical properties and are expected to be placed together within the same cluster. This result would be considered meaningful. The term "*partitioning*" is commonly used for supervised clustering methods, in which the number of clusters are defined prior to data clustering. In this article we use the term "*partitioning" to mean a division of a data set.*

Clustering algorithms are divided into two main categories: hard (or classical) and fuzzy methods. Hard clustering methods are those for which each observation belongs to a single cluster only, while fuzzy methods allow one observation to belong to each cluster to a varying degree. The choice of which algorithm to use depends on the goals of the study as well as the size and structure of the data set (Templ et al., 2006; Davis, 2002; Derde and Massart, 1982; Einax et al., 1997; Höppner et al., 1999; Mechelen et al., 2004).

Hard, hierarchical cluster analyses are being used to cluster various environmental/geological data sets with various aims. However, the outcomes of such hard, classical cluster analyses can become disturbed by factors when: (a) there are insufficient numbers of samples; (b) there are inaccuracies in the chemical analyses, or; (c) the sampling methods are not performed carefully. Furthermore, some factors that play a role in geological and environmental studies cannot be reported numerically: such as water-rock interactions, or those factors that have irregular measurement intervals such as rainfall. Fuzzy methods are helpful in such cases because they are flexible and can consider more possible relationships between the parameters (Demicco and Klir, 2003).

Most fuzzy clustering algorithms are objective-function based (Bezdek, 1981; Bezdek et al., 1987), in which each cluster is represented by a cluster centroid. The cluster centroid is computed by the clustering algorithm and may or may not appear in the dataset. The partitioning of the data points into different clusters is a function of the membership degree, which is computed based on the distance of the data points to the cluster centers. The closer a data point lies to the center of a cluster, the higher is its degree of membership to this cluster. Whereas classical methods require finite precision and grouping of terms with absolute membership, so that the membership within the group is either affirmative (with a membership degree of "1") or negative (a membership degree of "0"), fuzzy methods utilize membership functions or degrees of truthfulness and falsehoods that are not only either 0 or 1, but can be any of the numbers between.

Fuzzy clustering algorithms are based on the ideas of fuzzy set theory by Zadeh (1965), and provide a method to formalize reasoning when dealing with vague terms. The basic algorithms of fuzzy clustering have been developed by Dunn (1973), and improved by Bezdek (1981). In geology fuzzy logic offers an alternative to classical statistical modeling in geology that is more computationally efficient and more intuitive for geologists than complicated numerical models consisting of a few sets of differential equations.

Fuzzy *hierarchical* clustering is a method that has been used for geological/geochemical data. Fuzzy hierarchical *cross-clustering* (FHCC) is an improved version of fuzzy clustering that has been used for a limited number of studies in general (Sârbu et al., 1993; Sârbu and Pop, 2000); few applications have focused on geological/geochemical studies. Data of Greek mud characteristics were clustered with data of other, well-known beautifying muds using FHCC in order to determine whether the Greek muds could be considered similar and thus beneficial for therapeutic use (Dumitrescu et al., 1995). Mineral water samples from well-known sources found in Germany were compared to those from lesser-known sources in Romania using FHCC (Zwanziger and Sârbu, 1999). The physico-chemical properties of coal samples collected from America and Romania were also studied using fuzzy hierarchical cross-clustering (Sârbu and Pop, 2001).

This study relied on FHCC as well as the fuzzy divisive hierarchical clustering (FDHC) algorithms to cluster the data of pore water composition and leachate composition. FDHC uses only cluster centers and a Euclidean distance function to compute the membership degree of the data point to the cluster center. In comparison, the FHCC algorithm produces both a fuzzy partition of the compositions, as well as a fuzzy partition of the samples: the algorithm considers whether or not a group of objects form a cluster, as well as the membership degree to the cluster (Zadeh, 1965). In summary, fuzzy algorithms were used to identify the most characteristic parameters of sample clusters, and the outcome confirms the robustness of the applied fuzzy methods for geological, geochemical, and environmental data.

In the following sections, a description of the site where the samples were collected is presented, along with the details of the chemical analyses that were performed. The techniques that were used to analyze the data are described: these include Hard hierarchical clustering, fuzzy divisive hierarchical clustering and fuzzy hierarchical cross-clustering. Finally the results of the data analyses are presented and compared.

2. Experimental details

2.1. Site description

The former uranium mining site district in eastern Thuringia, Germany with more than 113,000 ton of mined uranium, was the third-largest uranium producer in the world (Jakubick et al., 2002; Lange, 1995). Mining activities began in 1949 and ended in 1990 after the re-unification of Germany. The remnants of the mining activities included a large, open pit mine called "Lichtenberg", with a depth of more than 200 m, 1.6 km length and 0.6 km width, and an underground mining system going down to 900 m depth with 3,000 km of underground galleries, and several waste rock piles or heaps (Wismut, 1994). Among the several heaps in the area, the