Available online at www.sciencedirect.com
SCIENCE DIRECT

Journal of Geochemical Exploration 87 (2005) 83–91

JOURNAL OF GEOCHEMICAL EXPLORATION
www.elsevier.com/locate/jgeoexp

Recognition of areas of anomalous concentration of potentially hazardous elements by means of a subcatchment-based discriminant analysis of stream sediments

M. Spadoni *, M. Voltaggio, G. Cavarretta

Consiglio Nazionale delle Ricerche, Istituto di Geologia Ambientale e Geoingegneria, via Bolognola 7, 00138 Roma, Italy

Received 5 August 2004; accepted 12 August 2005
Available online 27 September 2005

Abstract

The statistical technique of discriminant analysis associated with the calculation of an *information coefficient* has been applied to the concentrations of 37 chemical elements for calculating the mixing of stream sediments of different origin in the Mignone river basin.

Discriminant analysis has been based on *sample catchment basins* (SCBs), defined as the part of the drainage basin between two consecutive sampling points along the same stream branch, and on the identification of 4 different litho-geochemical groups. This approach, has been used to define the membership probability values for every sample by applying Bayes' rule and calculating posterior probability. The grade of uncertainty for each group assignment has been evaluated by using an *information coefficient*, based on a *classification entropy index*, and running a procedure analogue to that used for processing membership values in fuzzy analysis. The maximum theoretical concentration that can be expected in soils near the sampling point (*enhanced concentration*) has then been calculated from both the measured and the membership values by introducing a specific *enhancement function*. Theoretical background concentrations at every sampling point have been also calculated by weighting the average value of concentration in each group with the membership values for each sample. These have been successively compared with the measured and enhanced concentrations to identify anomalous areas.

The distribution maps of Arsenic and Vanadium in the Mignone River basin (central Italy) have been drawn accordingly to this technique, leading to the identification of areas of potential risk for human health.

Keywords: Stream sediments; Discriminant analysis; Information coefficient; Geochemical mapping

1. Introduction

Active sediments collected along a stream (stream sediments, Salminen et al., 1998) can be genetically considered as a mixing between grains and particles of different nature originated from erosive processes within a catchment basin. Consequently, the geochemical characteristics of each sample can be considered as a function of the composition of different geological materials and sediments of anthropogenic origin transported along the hydrographical network (Meyer et al., 1979; Bölviken et al., 1986; IGS, 1978; Webb et al., 1978; Lahermo et al., 1996). Their composition can be described by means of weighted averages of different factors such as geological setting and history, slope, vegetation, pedogenesis, industrial activities. These factors are hardly quantifiable in detail, even when an

* Corresponding author. Tel.: +39 0649914660; fax: +39 064468632.
E-mail address: massimo.spadoni@igag.cnr.it (M. Spadoni).

0375-6742/$ - see front matter
doi:10.1016/j.gexplo.2005.08.001

accurate knowledge of the natural processes and economic activities in the whole drainage basin is available.

Understanding the complex nature of stream sediments and estimating the geochemical inputs rising from different sources is of fundamental importance in environmental geochemistry, especially when investigating the background concentrations and the presence of natural and man-induced anomalies. For this purpose multivariate statistics approaches, such as discriminant and cluster analysis, can help in identifying subtended environmental processes, assigning each sample to different statistic 'groups.' Among these techniques, fuzzy c-means cluster analysis has recently spread in the field of environmental sciences for its capability to give answers in terms of membership, allowing to appreciate gradual changes among clusters (i.e. Hanesch et al., 2001; Kramar, 1995). It has also been successfully applied to stream sediments to identify local enrichments of geochemical concentrations (Rantitsch, 2000). In general, fuzzy c-means studies are directly based on the algorithm given by Bezdek et al. (1984) and are used to give a membership value, variable between 0 and 1 for each case, in order that the sum of all membership values is 1.

As an alternative to the fuzzy c-means analysis, in this paper we suggest the use of a simple discriminant analysis associated with the measurement of the conditional and posterior probability, measured with the Bayes' theorem, to identify areas of anomalous enrichment of potentially harmful elements. We also show how this information can be used to draw operative maps in a GIS in order to better circumscribe areas of risk for the health of population and to plan further detailed surveys.

2. Data processing

Sample Catchment Basins (SCB) can be described as the part of the drainage basin between two consecutive sampling points along the same stream branch (Spadoni et al., 2004). The sediments are expected to be more homogeneous when SCBs are of low rank (close to the watershed line) and, above all, when a homogeneous geological setting (i.e. same bedrock) is recognizable in them. In this kind of catchments, named 'index SCB' ($_i$SCB), we can reasonably associate the measured concentrations with specific and well defined geochemical sources. However, even when considering different catchments extended over the same bedrock, i.e. two $_i$SCBs extended over the same geological unit, a natural geochemical variability has to be expected as well.

Considering different $_i$SCBs inside a wider catchment system, their geochemical variability can be efficiently modellized thanks to techniques of multivariate statistics. Cluster analysis is one of the traditional statistical techniques used in geochemistry to group multivariate objects (SCBs in our case) of similar characteristics. This approach demonstrates its effectiveness especially when litho-geochemical features of the surveyed areas are poorly known. On the contrary, when an accurate knowledge of the geological setting is available, as in our case, statistical analysis can be better oriented by establishing a priori the belonging of the objects to specific groups characterized by a specific field of variability. Discriminant analysis is the statistical technique that can be used: (1) to identify the geochemical variables and their linear combination that better characterize the differences between 'groups'; (2) to assign ungrouped objects to the recognized groups at specified probability levels. This statistical technique requires that the predictor variables have multivariate normal distributions. Since geochemical variables usually show very skewed distributions (Reinmann and Filzmoser, 2000), the lognormal transformation should be used to fulfil the required assumption. Discriminant analysis approach allows to assign every sample to one group with a certain grade of probability. For each sample, the probability level to belong to the generic group n can be expressed in terms of *posterior probability* (P_n) that can be estimated, using Bayes' rule, from the *conditional probability* (C_n) and the *prior probability* (G_n):

$$P_n = \frac{C_n \cdot G_n}{\sum_{n=1}^{f} C_n \cdot G_n} \tag{1}$$

where the G_n is an estimate of the belonging of a sample to a specific group when no information is available and the C_n is the probability associated to the membership score for the same sample when it is assumed to belong to a particular group.

The grade of uncertainty of each group assignment can be calculated, considering the associated posterior probabilities, by means of a *classification entropy index* (h, from Brown, 1998), calculated adapting the equation that was originally developed for managing fuzzy membership values:

$$h_x = \left| \frac{1}{\ln f} \sum_{n=1}^{f} P_{nx} \cdot \ln(P_{nx}) \right| \tag{2}$$

where P_{nx} is the posterior probability value for sample x in group n and f is the total number of groups.