MONITORING PROCEDURES IN ENVIRONMENTAL GEOCHEMISTRY AND COMPOSITIONAL DATA ANALYSIS THEORY

A. Buccianti[1], O. Vaselli[1], B. Nisi[1], A. Minissale[2] and F. Tassi[1]
[1]Department of Earth Sciences, University of Florence (I)
[2]CNR-Institute of Geosciences and Earth Resources, Florence (I)

Abstract

First discussion on compositional data analysis is attributable to Karl Pearson, in 1897. However, notwithstanding the recent developments on algebraic structure of the simplex, more than twenty years after Aitchison's idea of log-transformations of closed data, scientific literature is again full of statistical treatments of this type of data by using traditional methodologies. This is particularly true in environmental geochemistry where besides the problem of the closure, the spatial structure (dependence) of the data have to be considered. In this work we propose the use of log-contrast values, obtained by a simplicial principal component analysis, as *indicators* of given environmental conditions. The investigation of the log-constrast frequency distributions allows pointing out the statistical laws able to generate the values and to govern their variability. The changes, if compared, for example, with the mean values of the random variables assumed as models, or other reference parameters, allow defining *monitors* to be used to assess the extent of possible environmental contamination. Case study on running and ground waters from Chiavenna Valley (Northern Italy) by using Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^-, SO_4^{2-} and Cl^- concentrations will be illustrated.

1. Introduction

Environmental Geochemistry is a discipline deeply involved in the investigation of the quality of the environment and in this context (geo)chemical monitoring is used to determine the amount of pollution in relation to natural background levels. The meaning of monitoring, according to the Oxford Dictionary, is related to the maintenance of a regular surveillance in a given area. Generally, when the object of a monitoring program is the determination of water quality as collected from springs, groundwater wells or running waters, data will show a spatial pattern of distribution and geochemical maps represent the tool to visualise and interpret chemical-physical phenomena. However, geochemical mapping can be considered as a tool of monitoring only if spatial analysis of data is repeated at fairly regular time intervals, so that the dynamics of the processes can be evaluated. In this context, the identification of parameters to be considered as *indicators*, whose characteristics are used to point out the presence or absence of given environmental conditions, or as *monitors*, whose changes can be measured to asses the extent of environmental contamination, may be highly useful. Water chemistry studies, under compositional data analysis theory, may allow us to define these types of parameters, particularly when the results of log-contrast analysis may be interpreted as related to specific chemical-physical phenomena and their spatial distribution is valuable. Application examples by considering water chemistry data collected from Chiavenna Valley (Central-Western Alps, Italy) will be shown. The aim is to define and develop the use of tools able to simulate the geochemical behaviour of water components in natural systems in order to: i) monitor the current state of the investigated natural system on a scientific base, under a correct statistical theoretical framework and ii) to estimate and predict direction of changes, particularly when the behaviour of pollutant, e.g. heavy metals, is considered.

2. Compositional data analysis theory and spatial investigation on water chemistry

Different methods of representation of hydrochemical data are used to understanding chemical distribution of species and for assessing the relationships among different water types. All of them are generally used to identify processes involved in the geochemical evolution of running and ground waters. However, graphical representations pose considerable problems, since dimension limitations severely restrict possible illustration of complex phenomena, offering in several cases only partial visions. Several problems affecting also diagrams used for classification purposes, as the square diagram of Langelier-

Ludwig or the Piper ternary diagram. In these cases, the sample space in which data are represented is indeed given by the simplex and we know (see i.e. Aitchison, 1986) that its geometry is not the same of the real space. In other words, discrimination of groups and definition of trends in these types of representations is not possible with standard statistical methodologies.
Distribution maps represent a tool used to analyse the spatial structure of water chemistry data. The most frequently used maps generally depict total dissolved solids or electrical conductivity and give preliminary information on a given water system and on its quality. These maps can be obtained by considering also the concentrations of anions, cations and trace elements (i.e. heavy metals), so that localisation of anomalous values, identification of trends and so on, may have important relapses on the interpretation of the effect of natural and antrophic phenomena and finally on the management of natural resources.

Regionalised compositions, similar to all compositions, present the problem of spurious spatial correlation so that a regionalised composition is a co-regionalisation characterised by the fact that the random variables describing it have a constant sum at each point of the sampled region. When this type of observations is used, auto- and cross-covariance functions are subjected to non-stochastical controls and may present distortions thus leading to a misinterpretation of the phenomenon under study (Pawlowsky-Glahn and Burger, 1992). In this work, an insight on this item is presented with the aim to focus the attention on the possible use of log-contrast values to map natural phenomena thus giving parameters useful in environmental monitoring programs that are not influenced by closure.

Our observations are given by running and ground waters from Chiavenna Valley, a NNW-SSE elongated basin with a surface of about 700 km^2, located in the central-western sector of the Alpine chain. From a geological point of view the area consists of schistose to crystalline rocks, sedimentary formations and a relatively well-developed Quaternary cover (Fig. 1). Starting from 1998, in the framework of the Chiavenna Project, financed by CNR and CARIPLO, more than 230 running and ground water samples have been collected and analysed for major, minor and trace components (Vaselli et al., 1999).

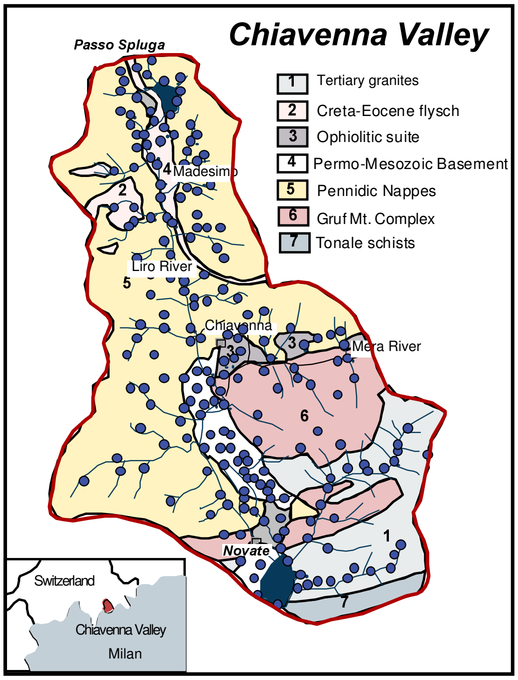

Figure 1. Schematic geological map of Chiavenna Valley and sampling location sites.

Considering a region $\Omega \subset \Re^n$ and an D-dimensional random vector function or co-regionalisation (Matheron, 1971)

$$z(x) = \left(z_1(x), z_2(x), \ldots, z_D(x)\right)'$$

(the prime standing for transpose), defined at each point $x \in \Omega$; $z(x)$ can be considered a D-part regionalised composition if: (1) all its components are strictly positive: $z_i(x) > 0,\ i = 1,2,\ldots,D$, and (2) for each point $x \in \Omega$ the sum of all its components is a constant c. The natural sample space for $\mathbf{z}(x)$ is the d-dimensional simplex

$$S^d = \{\mathbf{z}(x)\colon\ z_i(x) > 0,\ i = 1,2,\ldots,D;\ \mathbf{j}'\mathbf{z}(x) = c\}, \qquad (d = D - 1)$$

where $\mathbf{j} = (1,1,\ldots,1)'$ and c is a constant.

When compositional data are analysed

$$\gamma_j(h) = -\sum_{i \neq j} \gamma_{ij}(h), \qquad j = 1,2,\ldots,D$$

and

$$C_j(h) = -\sum_{i \neq j} C_{ij}(h), \qquad j = 1,2,\ldots D$$

where γ indicates the auto and cross-semivariograms and C the auto or cross-covariances (Pawlowsky-Glahn, 1984; Pawlowsky-Glahn, and Burger, 1992). Thus, if in regionalised compositions $\gamma_j(h)$ (for $h \neq 0$) and $C_j(h)$ are strictly positive, cross-covariances and cross-semivariograms show the same problems that compositional data show for covariances: (1) necessarily non-zero values, (2) bias toward negative values, (3) singularity of associated matrices.

Over the last few years, a new approach to statistical analysis of compositional data has been proposed by Aitchison (1986) based on transformations to leave the simplex as sample space in order to avoid problems induced by the constant-sum constraint. The transformations lead to the definition of an appropriate spatial covariance structure that can be used to analyse, describe and interpret spatial relationships of regionalised compositions.

In this work a new procedure is adopted in order to avoid the closure constraint and to develop a way to consider the dynamical nature of geochemical processes as described by changes in water composition. Our proposal is based on the following steps: (1) choice of the anions and cations normally used to plot data in the Langelier-Ludwig (1942) square diagram and classical interpretation of the points pattern; (2) application of log-contrast analysis for compositional data on the same variables and interpretation of the log-contrasts (Aitchison, 1997); (3) analysis of the frequency distribution of the log-contrast values, considered as random variables; (4) variograms of the log-contrast values and identification of anisotropies and trends in data variability for further mapping. Details on log-contrast analysis may be found in Aitchison, 1986.

3. Case study

In Figure 2 the Langelier-Ludwig square diagram is reported; as we can see, samples can chemically be classified as Ca(Mg)-HCO_3 waters, even if few observations have a Ca(Mg)-SO_4 composition; however, as known to put points in this diagram groups of variables reported on the two axes have to be closed to 50 and the interpretation of patterns/trends, as well as of the distance among observations, is not usable from an inferential statistical point of view.

Results of the log-contrast analysis applied on the cations and anions of the square diagram indicate that if we consider four log-contrasts, more than the 90% of the data variability is explained, about 55% attributable to the first one, 13% to the second one and about 12% to the other two ones.

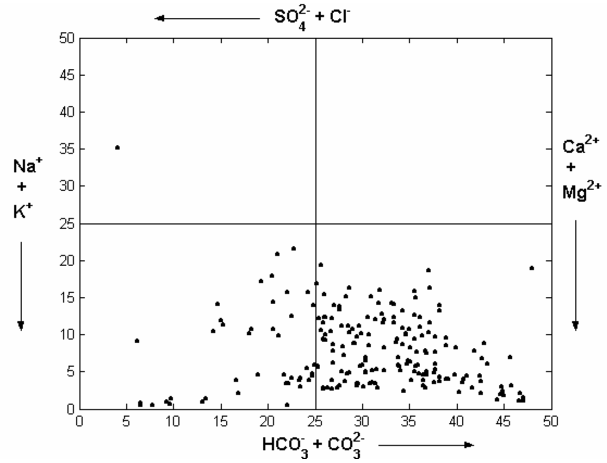

Figure 2. Langelier-Ludwig diagrams for running and groundwaters samples of Chiavenna Valley

Each log-contrast can be written as:

$$[0.46\log(Cl)+0.29\log(Na)+0.29\log(K)+0.21\log(Mg)]-[0.61\log(HCO_3)+0.32\log(Ca)+0.32\log(SO_4)]=k_1$$

$$[0.46\log(Na)+0.41\log(HCO_3)+0.30\log(K)]-[0.53\log(Mg)+0.48\log(SO_4)+0.07\log(Cl)+0.07\log(Ca)]=k_2$$

$$[0.66\log(SO_4)+0.30\log(Na)+0.04\log(Cl)+0.002\log(K)]-[0.59\log(Mg)+0.34\log(HCO_3)+0.08\log(Ca)]=k_3$$

$$[0.79\log(Cl)+0.19\log(HCO_3)+0.08\log(Ca)]-[0.43\log(K)+0.30\log(Mg)+0.22\log(Na)+0.11\log(SO_4)]=k_4$$

or, if the exponential form is considered:

$$\frac{(Cl)^{0.46}(Na)^{0.29}(K)^{0.29}(Mg)^{0.21}}{(HCO_3)^{0.61}(Ca)^{0.32}(SO_4)^{0.32}}=k_1, \qquad \frac{(Na)^{0.46}(HCO_3)^{0.41}(K)^{0.30}}{(Mg)^{0.53}(SO_4)^{0.48}(Cl)^{0.07}(Ca)^{0.07}}=k_2,$$

$$\frac{(SO_4)^{0.66}(Na)^{0.30}(Cl)^{0.04}(K)^{0.002}}{(Mg)^{0.59}(HCO_3)^{0.34}(Ca)^{0.08}}=k_3, \qquad \frac{(Cl)^{0.79}(HCO_3)^{0.19}(Ca)^{0.08}}{(K)^{0.43}(Mg)^{0.30}(Na)^{0.22}(SO_4)^{0.11}}=k_4.$$

and represent, from a general point of view, the relationships among the involved variables in four different directions of the multidimensional space. The coefficients of the log-contrasts, their signs and sizes, indicate how the variables are added to, or subtracted from, the water, and their correlations, helping us to understand the action of possible geochemical processes. For example, in the first log-contrast a possible balance between Cl, Na, K and Mg versus HCO$_3$, Ca and SO$_4$ may be evidenced, so that the contribution of the weathering of silicate rocks is balanced by those attributable to carbonate rocks and gypsum; to be noticed in this case is the same size of the coefficients of Ca and SO$_4$, thus representing their relationship in gypsum. This component explains the highest variability and consequently represents a very general phenomenon, as possibly those described.

The other components apparently describe local phenomena primarily depending from lithology. For example the second component is characterised by the absence of a role played by Ca and Cl while Na, HCO_3 and K are balanced by Mg and SO_4, thus representing weathering of silicate and carbonate minerals, against local phenomena involving Mg and SO_4. The third component is characterised by a clear association between SO_4 and Na compared with Mg and HCO_3 (K, Ca and Cl are not here important). Finally, the fourth component is particularly related to Cl, compared with Na, K and Mg representing, in this case too, probable local phenomena.

The k parameter of the log-contrasts changes by considering the values of the variables for each observation and represents the effects of geochemical processes that link the variables; the investigation of the frequency distribution of k values, considered as extracted from random variables, allows to verify how these data have statistically been generated, a tool useful to interpret in a subsequent phase, their meaning from a natural point of view (Fig. 3).

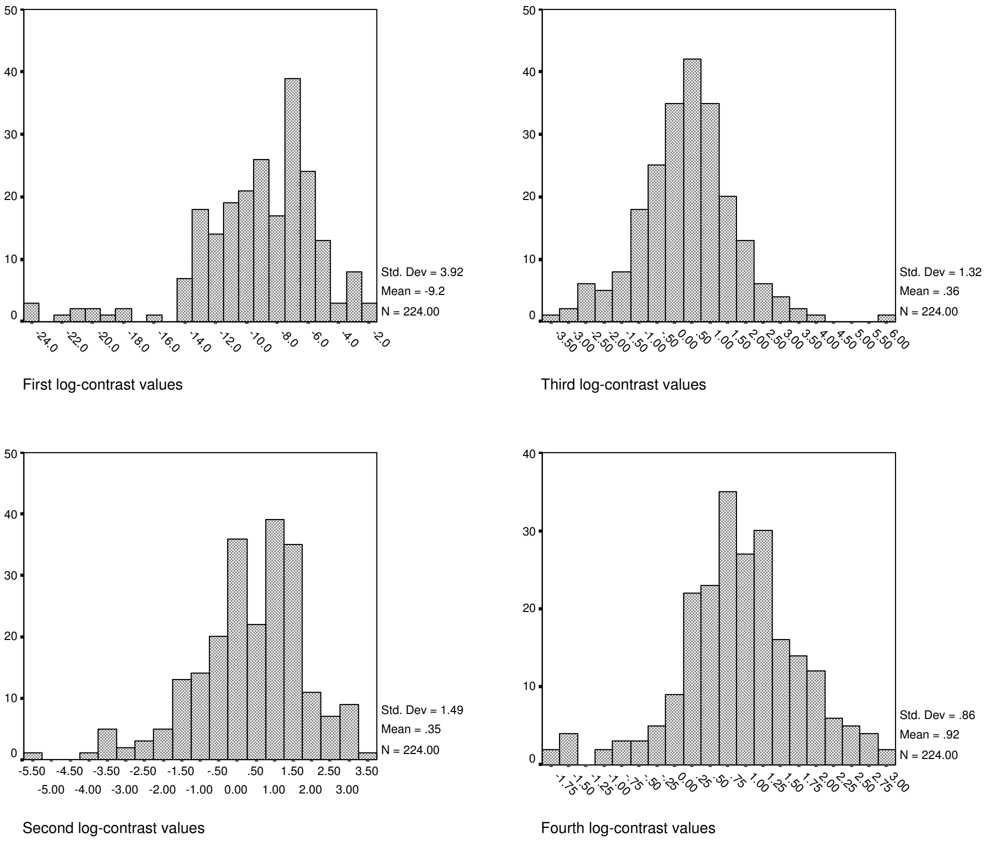

Figure 3. Frequency distributions of the k log-contrast values.

As verified by using the Kolmogorv-Smirnov test ($\alpha = 0.05$), the frequency distribution of the k values for the four log-contrasts can be considered as normal. This implies that the mean of log-contrast values is significant and that the probability to find higher or lower values of this barycentre is the same and also that about 68% of the data lies within one standard deviation from it. In other words, the phenomena affecting the chemistry of the investigated waters and that act with positive and negative contributions respect to the mean values, follow the Gauss Law and, consequently, appear to be in a general equilibrium state (Aitchison, 1999). This situation tends to arise naturally when many continuous independent variables are added together and indicates us that the life of chemical species is mainly influenced by solution/precipitation events and by diffusion/dispersion phenomena and that the net result is a balance. This result may confirm that the investigated area is not affected by antrophic effects able to disturb in a specific way this type of situation. However, some anomalous samples in the tails of the distributions can be recognised and are responsible, for example in the case of the first log-contrast, for generating the most negative values, to which a particular contribution of HCO_3, SO_4 and Ca has to be considered. Thus, following our procedure, at this point k values can be investigated from a spatial point of view and the variographic analysis is a fundamental step of further developments.

4. Spatial variability of the log-contrasts: anisotropies in geochemical processes

The variographical analysis of k parameters as *indicators* of given environmental conditions, allows to evidence anisotropies in the geochemical processes. As an example of explorative investigation, in Figure 4 the variograms in three main directions, N-S, E-W and 45° have been reported for all the four log-contrasts. On the x and y-axes the lag distance and $\gamma(h)$, the values of the semivariograms, are reported, respectively.

From a general point of view, we can see that the squared differences in the values of couples of samples up to a lag distance of about 5000 m are higher for the first log-contrast if compared with the others. Moreover, in the case of the first and second log-contrast values, a clear difference between the E-W and N-S directions is evidenced, with the higher values attributable to the E-W direction. Differences registered at 45° appear to follow in each case the behaviour of those related to the E-W direction. In all the variograms the *nugget effect* indicates the presence of a short-range variability due to geological features with correlations ranges shorter than the sampling resolution. Furthermore, the presence of fluctuations (possible *hole effect*, *i.e.* k_4 values) is representative of periodic phenomena at the investigated scale (same water/rocks interaction processes?).

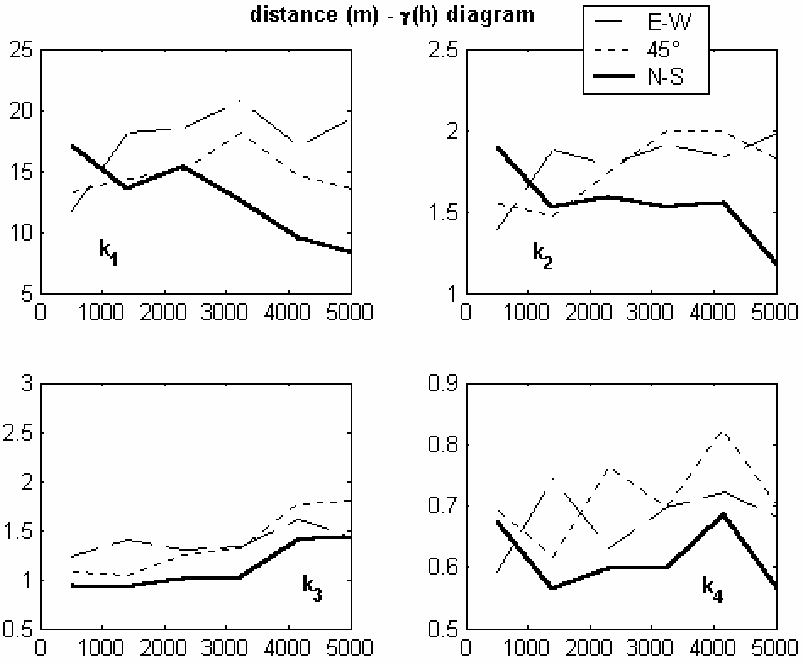

Figure 4. Variograms of the k_i log-contrast values by considering three main directions.

On the whole, the results indicate that variograms in the E-W and 45° directions for k_1 and k_2 values show a more or less classical form indicating that these paths are characterised by higher continuity if compared with the behaviour in the N-S direction. From a geochemical point of view this indicates that data appear to be more continuous transversally with respect to the N-S main drainage system. Continuity at the scale here considered (up to 5000 m) could be dominated both by lithology and secondary drainage systems.

In the case of the k_3 log-contrast values, no big differences characterise the behaviour in the three considered directions while k_4 behaviour shows differences between E-W and 45° directions, compared with the N-S one. In this case remarkable fluctuations (*hole effect?)* are also recognised.

A display that is able to reveal and summarise directional anisotropies is a contour map of the sample variogram surface. In calculating the variogram value for pairs of points separated by the vector $\mathbf{h} = (h_x, h_y)$, all pairs whose separation in the x direction is $h_x \pm \Delta x$ and whose separation in the y direction is $h_y \pm \Delta y$ are grouped together. As an example, Figure 5 shows the variogram surface calculations of k_1 values for pairs whose separation distance in the east-west direction, h_x, is less than 10 km and whose separation distance in the north-south direction, h_y, is less than 30 km. Data pairs have been grouped into 64 lags, with a lag increment of 12.8 km in E-W direction (lag tolerance ± 6.4) and of 37.6 km in N-S direction (lag tolerance ± 18.8).

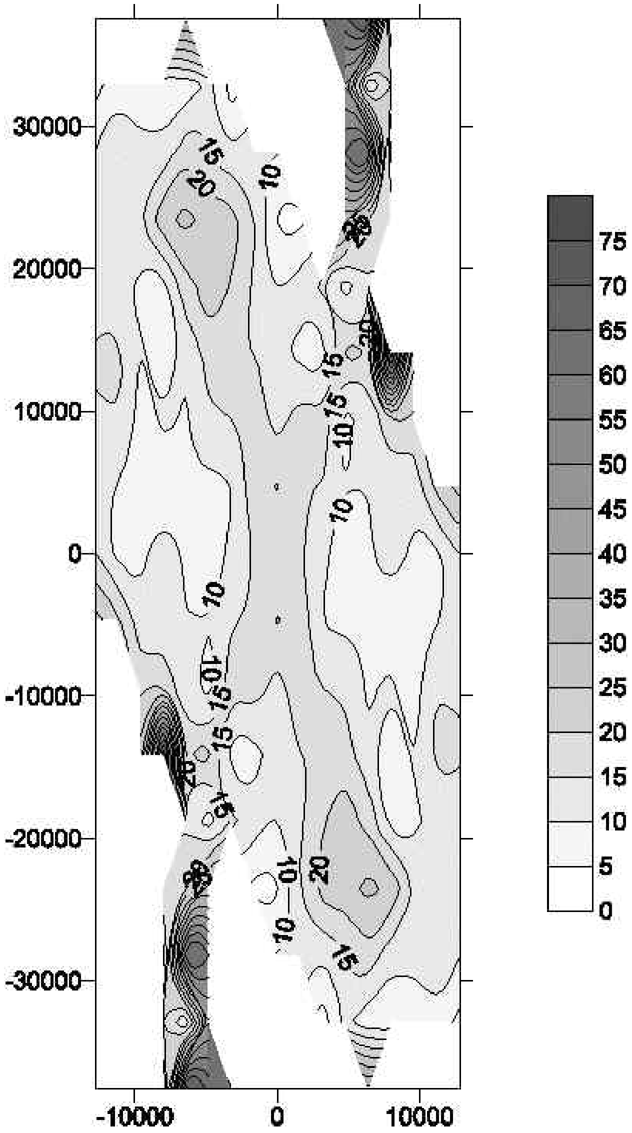

Figure 5. Variogram surface of the k_1 log-contrast values

The center of the map corresponds to the origin of the semivariogram and any cross section appears as a traditional one-dimensional variogram. Semivariogram values are small near the origin and increase with distance from this point. When the variation is isotropic the increase is fairly similar in every direction and the map shows concentric contour lines. On the contrary, geometric anisotropy appears as elliptical contour lines whose major axis indicates the direction of maximum continuity. In the case of k_1 log-contrast values a better long-range spatial continuity (smaller semivariogram values) is related more or less with the N120°E direction. Consequently if the first log-contrast is associated to the contribution of rocks weathering, both carbonate and silicate ones, this direction is related to minor changes, compared with the perpendicular one.

Similar results can be obtained for the mapping of the variogram surface of the other log-contrast values and may be used to understand the spatial behaviour of the relationships among the investigated variables, thus representing a useful tool to monitor complex phenomena that cannot be described by using a variable at a time. Furthermore, k_i values are random variables whose values are not constrained and standard geostatistical methodologies can be properly used.

Conclusions

In this work a new procedure to investigate from a spatial point of view geochemical processes has been proposed with the aim to capture the dynamics of natural environmental changes affecting the chemical abundance of elements. The approach is based on the theory developed in recent years for compositional data to take into account the nature of their sample space so that descriptive and inferential statistical methodologies can be applied without errors and, consequently, misinterpretations. In our case, the modelling of the log-contrast frequency distributions obtained by a simplicial principal component analysis offer tools to be used as indicators of given environmental conditions. When log-contrast values, treated as random variables, follow a normal (gaussian) distribution, a general equilibrium state can be deduced so that geochemical processes able to add or subtract elements to water tend to balance each other.

The spatial behaviour of log-contrast values can be investigated by variographical analysis based on variograms and on the map of the variogram surface so that anisotropies in the geochemical processes can be revealed. Further studies in this direction are in progress, since the proposed methodology can be implemented by means of the mapping of k in order to have a tool to be related to the features of the investigated area. Knowledge about the spatial scale of geochemical processes characterising a certain environmental domain is fundamental when pollution problems and remedial programs have to be respectively investigated and planned. In the case study of Chiavenna Valley, a relatively pollution-free area, the results can be considered as a starting point to follow the evolution of natural geochemical precipitation/solution and diffusion/dispersion processes in time and to evaluate the impact of possible human effects in the future.

References

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, Ltd., London, 416 p.

Aitchison, J., 1997, The one hour course in compositional data analysis, or Compositional data analysis is easy: Pawlowsky-Glahn ed., Proceedings of IAMG'98, the Third Annual Conference of the International Association for Mathematical Geology: CIMNE, Barcelona, p. 3-35.

Aitchison, J., 1999, Logratios and natural laws in compositional data analysis: Mathematical Geology, 31, 5, p. 563-580.

Langelier, W., and Ludwig, H., 1942, Graphical methods for indicating the mineral character of natural waters: J. Am. Water Ass., 34, p. 335-352.

Matheron, G., 1971, The theory of regionalised variables and its applications: C-5, Centre de géostatistique et de Morphologie Mathématique, Fointainebleau, France.

Pawlowsky-Glahn, V., 1984, On spurious spatial covariance between variables of constant sum: Sci. De La Terre, Ser. Inf., n. 21, Fontainebleau, France, p. 107-113.

Pawlowsky-Glahn, V., and Burger, H., 1992, Spatial structure analysis of regionalised compositions: Mathematical Geology, 24, 6, p. 675-691.

Vaselli, O., Caviglia, A., Tassi, F., and Gallorini, L., 1999, Running and ground water geochemistry from Chiavenna Valley, Northern Italy, Armannsson (Ed.), Geochemistry of the Earth's Surface: Balkema, Rotterdam, ISBN 9058090736, p. 555-558.

Webster, R., and Oliver, M. A., 1992, Sample adequately to estimate variograms of soil properties: Journal of Soil Science, 43, p. 177-192.