NEW INSIGHTS ON RIVER WATER CHEMISTRY BY USING NON-CENTRED SIMPLICIAL PRINCIPAL COMPONENT ANALYSIS: A CASE STUDY

A. Buccianti[1], O. Vaselli and B. Nisi
Department of Earth Sciences, University of Florence, Flornce (I)

Abstract

The use of perturbation and power transformation operations permits the investigation of linear processes in the simplex as in a vectorial space. When the investigated geochemical processes can be constrained by the use of well-known starting point, the eigenvectors of the covariance matrix of a non-centred principal component analysis allow to model compositional changes compared with a reference point.
The results obtained for the chemistry of water collected in River Arno (central-northern Italy) have open new perspectives for considering relative changes of the analysed variables and to hypothesise the relative effect of different acting physical-chemical processes, thus posing the basis for a quantitative modelling.

Introduction

The chemical composition of surface water depends fundamentally on the minerals that have dissolved, on the ion-exchange reactions, on processes such as precipitation, mixing and dilution, uptake and recycling of nutrient elements, exchange with the gases of the atmosphere and discharge of municipal and industrial waste. Consequently, river chemistry studies represent a fundamental tool to investigate the nature of erosion products of continents that reach the oceans, the features of the biogeochemical cycles of elements, weathering and physical erosion rates and CO_2 consumption by acid degradation of continental rocks. Moreover, the analysis of elemental concentrations in river water may be useful to estimate the whole influx of the chemical elements into the oceans and seas, an investigation that may be difficult, if the variable partitioning between the particulate and the dissolved fractions, which for many elements depends on filter size, is considered. For example elements as Na, Sr, Ca, Rb, Ba, U and B do not generally change concentrations with filter size, implying that they are prevalently soluble and not affected by organic complexation. As a consequence, their abundance can be used for the evaluation of the river discharge.

The investigation of the water of River Arno may be inserted in this geochemical framework with the aim to propose new tools to visualise, quantify and model compositional changes affecting the collected samples. The catchment of River Arno is located in northern Tuscany (central-northern Italy) and has a drainage basin of about 8.228 km^2. The river springs out in the Apennine chain and flows into the Tyrrhenian Sea, after crossing highly urbanised, industrialised and cultivated areas, responsible for spoiling the water quality (Fig. 1). From a chemical point of view, the collected waters (about 60 samples as a function of the source distance) are mainly characterised by a Ca-HCO_3 facies, although exceptions are represented by samples where Na-Cl and Ca-SO_4 enrichments are due to seawater intrusion and mixing with thermal waters and/or leaching of evaporites. The effect of anthropic influence is clear when the concentrations of N-bearing species, such as NH_4, NO_2 and NO_3, and the high TDS (Total Dissolved Salts) values, are considered.

Application of non-centred principal component analysis on Na^+, K^+, Mg^{2+}, Ca^{2+}, HCO_3^{2-} + CO_3^{2-}, Cl^-, SO_4^{2-} and SiO_2 by considering as initial composition those related to the source, allowed us to model compositional lines/processes as a function of the position of the samples in the river path.

Working on the simplex: brief state of the art

Standard descriptive statistics, e.g. arithmetic mean and variance, are not very informative in the case of compositional data since they are defined in the framework of the Euclidean geometry in real space. This geometry is not able to describe changes or differences in compositional (closed) data so that modelling

[1] buccianti@unifi.it

and statistical inferences may lead to serious mistakes. In order to understand this aspect, let to consider the difference between a proportion of 5% and a proportion of 10% and the difference between a proportion of 50% and a proportion of 55%. Even if the Euclidean distance between them is the same (in both cases we have an increase of 5 units), the relative increase is different; in the first case the proportion is doubled, in the second case it is 1.1 times, and the relative increase is only about 10%.

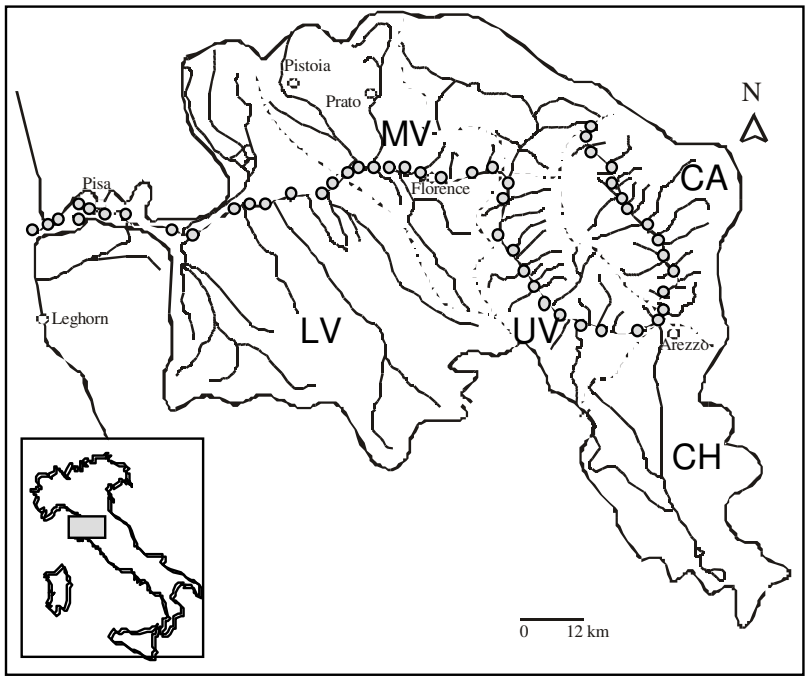

Figure 1. The Arno catchment and sampling points.

Main sub-basins:
Ca = Casentino;
CH = Chiana;
UV = Upper Valdarno;
MV = Medium Valdarno;
LW = Lower Valdarno.

It is clear that to capture changes in compositional data a sensible geometry is needed and now several tools are available to work in the simplex S^d, the sample space of compositional data with d components of a composition (the variables), as in the real space. In order to give to the simplex a vector space structure some basic operations have to be defined as for example closure, perturbation and power transformation. Closure of $\mathbf{x} = (x_1, \ldots, x_d) \in \Re_+^d$, with $x_i \geq 0, \mathrm{i} = 1, \ldots d$, is defined as:

$$C(x) = \left(\frac{\kappa\, x_1}{\sum_{1=1}^{d} x_i}, \ldots, \frac{\kappa\, x_d}{\sum_{1=1}^{d} x_i} \right),$$

and it is used to transform any vector with positive components in a composition. The perturbation of a composition $\mathbf{x} \in \mathbf{S^d}$ by a composition $\mathbf{y} \in \mathbf{S^d}$ is given by

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, \ldots, x_d y_d)$$

while the power transformation of a composition $\mathbf{x} \in \mathbf{S^d}$ by a constant $\alpha \in \Re$ is given by

$$\alpha \otimes \mathbf{x} = C(x_1^{\alpha}, \ldots x_d^{\alpha}).$$

The perturbation operation is analogous to addition in real space, while power transformation is analogous to multiplication by a scalar value. Both require in their definition the closure operation thus

leading to a composition. If an inner product, used to check for orthogonality and to determine angles between vectors, and a norm, used to determine the length of vectors, are further introduced, a distance definition is possible, and a linear vector space structure (a geometry), obtained (see i.e. Barceló-Vidal et al. 2001, Pawlowsky-Glahn and Egozcue, 2002).

Within this framework compositional lines can be defined in S^d to model changes in compositional data as

$$\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \otimes \mathbf{x})$$

where $\mathbf{x}_0$ is the starting point and $\mathbf{x}$ is the leading vector, both elements of S^d, while α varies in $\Re$.

Aitchison (1986) realised that for compositional data size is irrelevant since the interest is in the relative proportions of the measured components. He introduced two transformations based on ratios, the additive log-ratio transformation (*alr*) and the centered log-ratio transformation (*clr*). Classical statistical analysis methodologies can be applied to the transformed observations, taking care to use *alr* for modelling and *clr* for techniques based on a metric. This is due to the fact that *alr* transformation does not preserve distances while *clr* preserves it, but leads to a singular covariance matrix. Recent developments are giving an algebraic geometric foundation to the Aitchison approach (*i.e*. Egozcue et al., 2003).

Non-centred principal component analysis

Perturbation and power transformation can be combined to define compositional lines on the simplex S^d, lines that can be used to describe trends of compositional observations. Any trend can be determined by an initial composition $\mathbf{x}_0$ and a unitary composition $\mathbf{p}$ that defines the direction of the trend. Consequently, any composition $\mathbf{y}$ on the linear trend is completely determined by a scalar α such that

$$\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \otimes \mathbf{p}) = C(x_{01} p_1^{\alpha}, x_{02} p_2^{\alpha}, \ldots, x_{0d} p_d^{\alpha}).$$

This compositional linear trend, originated in $\mathbf{x}_0$ and with direction of changes given by $\mathbf{p}$ can be symbolised by $L_{x_0}(\mathbf{x}_0; \mathbf{p})$. If a linear trend has to be adjusted to a set of observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, or compositions in S^d, having as a starting point the composition $\mathbf{x}_0$, an axis going through $\mathbf{x}_0$ explaining as much as possible of the total simplicial variability of all the n compositions with respect to $\mathbf{x}_0$ has to be found. This problem can be solved applying non-central principal component analysis to the $n \times d$ matrix where the n rows are given by the vectors $clr\,\mathbf{x}_1 - clr\,\mathbf{x}_0, \ldots, clr\,\mathbf{x}_n - clr\,\mathbf{x}_0$. If the eigenvalues $\lambda_1, \ldots, \lambda_d$ of the covariance matrix of this transformed data set are considered in decreasing order of magnitude, the vectors $\mathbf{e}_1 = clr^{-1}\mathbf{v}_1, \ldots, \mathbf{e}_{d-1} = clr^{-1}\mathbf{v}_{d-1}$ give an ordered orthonormal basis of the vector space S^d, and explain the variability of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to $\mathbf{x}_0$ in decreasing order. As a consequence the linear trend starting from $\mathbf{x}_0$ which better adjusts the set of compositions $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is given by $L_{x_0}(\mathbf{x}_0; clr^{-1}\mathbf{v}_1)$ and the proportion of the total variability of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to $\mathbf{x}$ retained by this trend is obtained by considering the ratio $\lambda_1 / (\lambda_1 + \ldots + \lambda_{d-1})$, a parameter useful to estimate the quality of the fit.

The orthogonal projection of a compositional observation $\mathbf{x}_i$ onto the linear trend $L_{x_0}(\mathbf{x}_0; clr^{-1}\mathbf{v}_1)$ is given by the composition $\mathbf{y} = \mathbf{x}_0 \oplus (\alpha_i \otimes \mathbf{p})$, where $\alpha_i = \langle \mathbf{x}_i \oplus \mathbf{x}_0^{-1}, \mathbf{p} \rangle$ and the symbol $\langle\ \rangle$ indicates the inner product. In this contest, the real number α_i indicates *how many times* the initial composition $\mathbf{x}_0$ has to be perturbed by $\mathbf{p}$ to arrive in $\mathbf{y}$. A fundamental consequence of this approach is related to the interpretation of the process represented by $\mathbf{p}$ that determines the changes in the compositions $\mathbf{x}_1, \ldots, \mathbf{x}_n$; in this context, the α_i values can provide a measure of the amount or intensity of the process (precipitation rates, dilution, residence time and so on), between $\mathbf{x}_0$ and $\mathbf{x}_i$. An example of application to the chemical weathering of granitoid rocks can be found in von Eynatten et al., 2003 and another concerning fluid geochemistry in Pawlowsky-Glahn and Buccianti, 2002.

Case study on rivers water chemistry

In order to model and quantify compositional changes characterising water river chemistry, the composition of about 60 samples collected in the River Arno, from the source to the mouth, has been considered. The components of the composition are given by the concentrations of Na^+, K^+, Ca^{2+}, Mg^{2+}, $HCO_3^- + CO_3^{2-}$, SO_4^{2-}, Cl^-, SiO_2. Even if the number and type of variables are not exhaustive for understanding the effects on abundance of all natural and anthropic phenomena, some preliminary indications on the data structure can be obtained. In Table 1 the coefficients of the first three eigenvectors, explaining about the 86% of the total variability, are reported; they have been obtained on the original data modified by using the transformation $clr\,\mathbf{x}_1 - clr\,\mathbf{x}_0, \ldots, clr\,\mathbf{x}_n - clr\,\mathbf{x}_0$, where $\mathbf{x}_0$ is the composition of the source of the river. The proportion of the total variability of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to $\mathbf{x}_0$ captured by the compositional linear trend of the first eigenvector is 51% and we will focus our attention on the modelling of the processes that it may represent.

variables	I component	II component	III component
$HCO_3^- + CO_3^{2-}$	-0.18	0.55	-0.16
Cl^-	0.38	-0.37	0.24
SO_4^-	0.08	0.003	0.008
Na^+	0.45	-0.14	0.17
K^+	-0.15	-0.46	-0.78
Ca^{2+}	-0.16	0.38	-0.09
Mg^{2+}	0.28	0.33	0.10
SiO_2	-0.70	-0.29	0.51

Table 1. Coefficients of the first three eigenvectors calculated from the variables measured along the Arno river.

If we consider the sign and size of the coefficients of the first eigenvector, we can see that an important role is played by SiO_2, a variable due to the weathering of silicate minerals that adding silicic acid to water. A seconday role appears to be played by Cl and Na, that reflect both their conservative behaviour and the effect of the interaction with seawater upstream 15 km from the mouth of the river. However, even if the size of the coefficients of Na and Cl is similar, an additional contribution of Na, not related to Cl (*i.e.*, incongruent dissolution of Na-bearing silicates, e.g. plagioclase, to form kaolinite or some sort of pollution) has to be considered. By taking into account these observations, we can attribute to the first eigenvector the capacity to represent the weathering of silicate minerals compared with the influence attributable to the seawater wedge of the mouth. When the first phenomenon decreases, for example for dilution, the other tends to increase, as indicated by the signs of the coefficients.

The relative variation diagram of variables normalised to Ca^{2+} of the compositional linear trend $L_{x_0}(\mathbf{x}_0; \mathbf{p})$ adjusted to the Arno data set, by considering the composition of the source as starting point, is reported in Figure 2. The *y*-axis represents the relative change of the variable/Ca^{2+} ratio at a given α, compared to the initial variable/Ca^{2+} ratio in $\mathbf{x}_0$.

As we can see, SO_4^{2-}, Mg^{2+}, Cl^- and Na^+ are enriched relative to Ca^{2+} (they increase in a relatively fast way) while SiO_2 is depleted. In this contest, compared with Ca^{2+}, both K^+ and $HCO_3^- + CO_3^{2-}$ do not change significantly when compared with the starting point (the source in $\alpha = 0$ where $p_j / p_{Ca} = 1$ and $j = 1, \ldots, 8$ indicate the considered variables). In order to quantify our observations, consider that when $\alpha = 2$ the ratio SiO_2/Ca (curve labelled with SiO_2) is less than a half of the initial ratio in $\mathbf{x}_0$, while the ratios (HCO_3+CO_3)/Ca and K/Ca do not change in a significant way. On the other hand, the ratios SO_4/Ca and Mg/Ca are, respectively, 1.5 and 2 times the original value while for the Cl/Ca and Na/Cl the increase may be also three times.

From a geochemical point of view we have to consider that for HCO_3+CO_3 and K no significant processes are influencing the abundance in the river or, in other words, a sort of balance between possible input and output processes, able to compensate dilution and concentration phenomena, is acting. If we consider that carbonate species in solution derive from the dissolution of atmospheric CO_2 and carbonate

and silicate minerals, saturation in calcite is frequent in fluvial environment. Thus, dissolution/saturation processes may account for the absence of any significant trends.

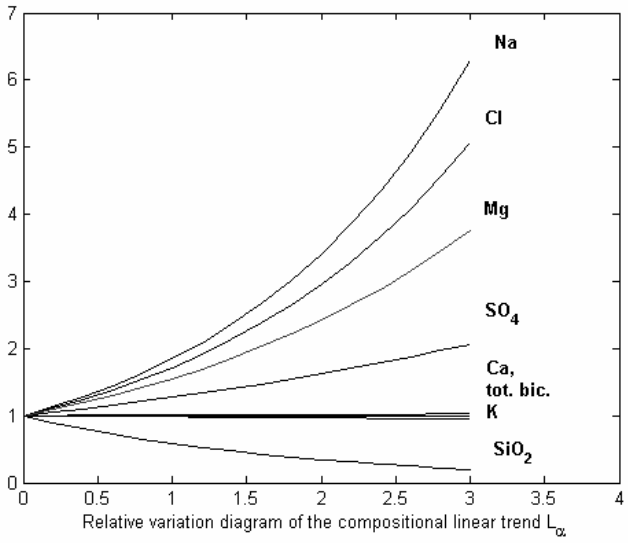

Figure 2. Relative variation diagram of variables normalised to Ca^{2+} and to the initial variable/Ca^{2+} ratio.

By considering that K is released by the weathering of silicate minerals or by the use of fertilizers, a possible balance of this element compared to Ca^{2+} indicates its possible extensive use by vegetation or the effect of sorption on clays and organic materials. The behaviour of SiO_2 is different, and processes able to explain the decreasing are needed. From a general point of view, SiO_2 is added to water as H_4SiO_4, and the solubility of Si-rich phases (particularly amorphous silica), increases with temperature. In our case the depletion of SiO_2 compared with Ca and with the ratio in the source has to be attributable to some process able to capture SiO_2 from water. A simple mechanism may be due to dilution, since the geology of the river basin, controlling the differences in silica concentrations does not change in a significant way. In other words, the flow rate may play an important role if the contribution of silica is due to a single source. However, possible use of silica by diatoms (tiny floating organisms) as a bioorganic element has to be verified.

Finally, the fast increase in SO_4^{2-}, Mg^{2+}, Cl^-, and Na^+ may be explained by their "*conservative*" behaviour and in the last portion of the river path by the contribution due to the seawater wedge, although inputs from dissolution of evaporitic rocks by sinister tributaries and contributions, though at a very low extent, of thermal waters discharges have to be considered.

To calculate the appropriate α value for each composition, the relationship $\alpha_i = \langle \mathbf{x}_i \oplus \mathbf{x}_0^{-1}, \mathbf{p} \rangle$ has been used. The plot of the obtained α values as a function of the distance from the source is reported in the left sub-plot (Fig. 3). As we can see, the trend is similar to those shown by Na/Ca, Cl/Ca, Mg/Ca and SO_4/Ca ratios plotted in the previous variation diagram (Fig. 2). If the changes in α values are considered (sub-plot of the right part of the diagram), as a function of the source distance, the increase in the intensity of the acting processes is evident near the mouth, even if fluctuations of minor intensity characterise also other parts of the river path, possibly reflecting the inputs of the tributaries. In other words, the amount of the process leading from the source $\mathbf{x}_0$ to $\mathbf{x}_1$ and then from $\mathbf{x}_1$ to $\mathbf{x}_2$ and so on is not constant for the entire river course, since $\Delta\alpha$ is not constant. As the fluctuations hardly cluster around the zero value, and their amplitude increases towards the mouth, we can hypothesise that a balance for Na^+, Mg^{2+}, SO_4^{2-}, Cl^-, and SiO_2 is not reached and processes able to add or subtract these ions are acting along the entire river path, with a final positive or negative cumulative effect.

After this discussion, a possible simple model as those reported in Figure 4 for the analysed data may be constructed; even if not exhaustive (we have analysed only a compositional linear trend by using the first component explaining about 50% of the total variability), it may be used as a guide for further investigation on the processes affecting the geochemistry of water of River Arno. From a general point of view, samples collected from the source along the Casentino and Chiana basins are characterised by an increase of twice of SiO_2, while SO_4, Mg, Na and Cl are more or less a half of the source content. From the Chiana samples towards the mouth, the relationships among the compositions change abruptly and

SiO_2 is now less than a half, if compared with the source, while SO_4 is 1.5 times and Mg, Cl and Na more than twice. Both for SiO_2 and the group of variables given by SO_4, Mg, Cl and Na, a negative or positive cumulative effect is registered from the Upper Valdarno toward the mouth.
The relative quantitative budget here proposed may be implemented if these results are weighted by considering the water flow data.

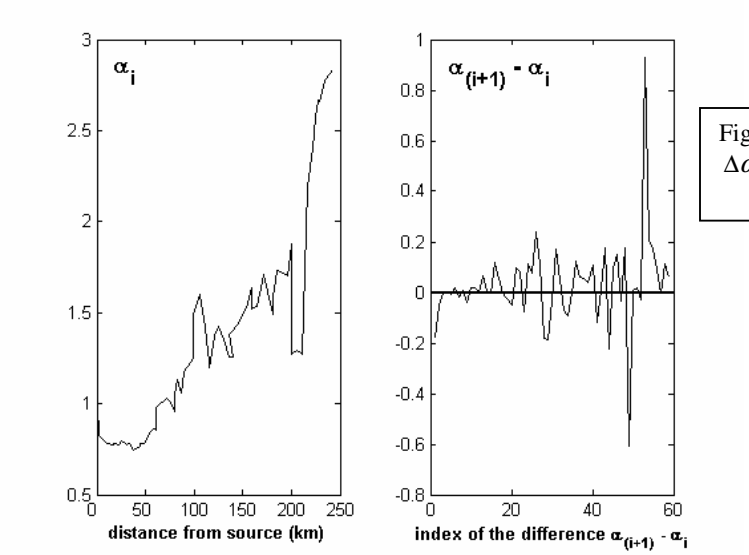

Figure 3. Dependence of α and $\Delta\alpha$ values from the source distance

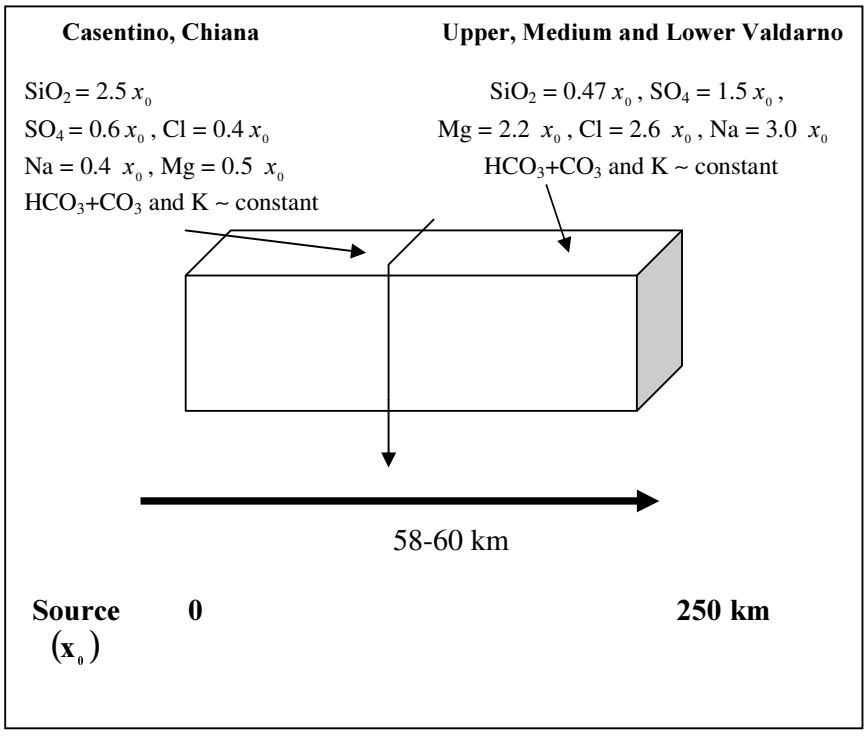

Figure 4.
Relative budget of the variable/Ca ratio and comparison with the source.
The numerical data are related to an average value of the two individuated sub-areas, given by Casentino + Chiana and Upper + Medium + Lower Valdarno.

Conclusions

Chemical changes in water chemistry of rivers can be modelled by using compositional linear trends of the type $L_{\mathbf{x}_0}(\mathbf{x}_0;\mathbf{p})$, where $\mathbf{x}_0$ represents the initial composition (i.e. the composition of the source) and $\mathbf{p}$ represents the unitary perturbation vector obtained from non-centred principal component analysis using $\mathbf{x}_0$ as origin. This approach has allowed the modelling of the relative changes in the chemistry of River Arno water and permitted to define the basis for a quantitative modelling. Even if the considered members of the composition, given by the concentration of Na^+, K^+, Ca^{2+}, Mg^{2+}, HCO_3^- + CO_3^{2-}, SO_4^{2-}, Cl^-, SiO_2 are not exhaustive to understand all the physical-chemical phenomena acting in this type of environment, our results indicate a clear cumulative effect in the abundance of some variables and consequently the absence of some sort of balance for SiO_2, SO_4^{2-}, Mg^{2+}, Cl^- and Na^+. Moreover, the behaviour of the variables appears to be different from a spatial point of view so that a discrimination of two areas, one given by Casentino and Chiana, the other by Upper, Medium and Lower Valdarno sub-basins can be recognised. Further developments of the modelling require the use of flow data and the work is in progress.

References

Aitchison, J., 1986, The statistical analysis of compositional data: Methuen, New York, 416 pp.

Barceló-Vidal, C., Martín-Fernández, J.A.,,and Pawlowsky-Glahn, V., 2001, Mathematical foundations for compositional data analysis: Proc. IAMG'01, CD-Rom, Cancun (Mexico), 20 pp.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, and G., Barceló-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis: Math. Geol., 35, 3, p. 279-300.

Pawlowsky-Glahn, V., and Buccianti, A., 2002, Visualization and modeling of sub-populations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids: Int. J. Earth Sci., 91, p. 357-368.

Pawlowsky-Glahn, V. and Egozcue, J. J., 2002, About BLU estimators and compositional data: Math. Geol., 34, p. 259-274.

Von Eynatten, H., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2003, Modelling compositional change: the example of chemical weathering of granitoid rocks: Math. Geol., 35, 3, p. 231-251.