Research Article Environmetrics

Received: 5 July 2012, Revised: 18 March 2013, Accepted: 15 April 2013, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/env.2214

Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy)

M. Gallo[a]* and A. Buccianti[b]

Data collected for the investigation of the environmental and ecological characteristics of a river basin are often in the form of a large three-way array; hence, a particular version of the Tucker model could be applied to gather more information contained in such complex geochemical systems. Indeed, when the data are in compositional form, more attention must be given to the analysis of the numerical data. Recently, the Tucker3 model has been proposed to analyze compositional data characterized by a three-way structure. In this work, a particular version of the Tucker model, known as the weighted principal component analysis, was used to analyze water samples collected from the Arno river (Tuscany, central Italy) in order to evaluate the method's effectiveness. Several graphical displays have been developed to allow an accurate and complete interpretation of results. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: compositional data; log ratios; Aitchison geometry; Tucker models; three-way arrays

1. INTRODUCTION

Compositional data (CoDa) consist of vectors of positive values summing to a unit or, in general, to some fixed constant for all vectors. They appear as proportions, percentages, concentrations, parts per million (ppm), or parts per billion. The CoDa are commonly present in all experimental fields, and the methods used to describe the chemical composition of the river water (mg L^{-1} or ppm) may be cited in this context.

The sample space for CoDa is the simplex with its natural geometry coherent with the concept of distance (Aitchison, 1986; Egozcue *et al*. 2003; Buccianti *et al*., 2006; Buccianti, 2011; Buccianti and Magli, 2011). Consequently, CoDa have important and particular properties that preclude the application of standard statistical techniques to such data in their raw form (Pawlowsky-Glahn and Buccianti, 2002; Buccianti and Pawlowsky-Glahn, 2005). Thus, several multidimensional techniques have been adapted to analyze CoDa, including principal component analysis (PCA; Aitchison, 1983), partial least squares (Hinkle and Rayens, 1995; Gallo, 2003), discriminant partial least squares (Gallo, 2010), and hierarchical clustering (Martin-Fernandez et al., 1998), which are only some of the multivariate techniques proposed in literature. Sometimes, the CoDa are arranged into three-way arrays, for instance, when the aim is to consider the changes in the composition of river water chemistry monitored in space and time. In these cases, three-mode models such as the Tucker models (Kroonenberg and de Leeuw, 1980; Tucker, 1966; Lombardo *et al*., 1996) can be proposed to explore the interrelations between the elements but only if the particular properties of CoDa are taken into account, that is, scale invariance and subcompositional coherence. Here, the scale invariance merely reinforces the intuitive idea that CoDa provide the same information when a full composition or some of its subcompositions are investigated, thus showing the same relations within the common parts. Gallo (2012a) has shown that centered log ratios are an adequate preprocessing solution for analyzing CoDa by the Tucker3 model. Through these developments, it is possible to correlate the results already obtained by means of other types of log-ratio transformations proposed in literature, such as the additive and isometric log-ratio transformations. Furthermore, several versions of joint biplots can be suggested to represent data and to facilitate the interpretation.

Starting from Gallo's (2012a, 2013) study, the aim of this paper was to analyze CoDa through a special version of the Tucker model that is known as weighted PCA (wPCA; Ten Berge *et al*., 1987) and by using some graphical procedures such as one-mode plots, *clr*-joint biplot, and trajectory plots to visualize data. To this purpose, an extensive matrix related to a sampling carried out along the Arno river (central-northern Apennines, Italy) between May 2002 and October 2003 (Nisi *et al*., 2008b; Nisi *et al*., 2008a) was analyzed. The chemical composition of surficial water pertaining to the Arno river is affected by the lithological features of the area, but high perturbation as expected is also due to the presence of urban, industrial, and agricultural–zootechnical contributions. To be noted here is that the anthropogenic pres-

* *Correspondence to: M. Gallo, Department of Human and Social Sciences, University of Naples "L'Orientale", Largo San Giovanni Maggiore, 30 I-80134 Naples, Italy. E-mail: mgallo@unior.it*

a *Department of Human and Social Sciences, University of Naples "L'Orientale", Naples, Italy*

b *Department of Earth Sciences, University of Florence, Florence, Italy*

sure of the Arno basin is estimated to be equal to that of 8,500,000 inhabitants. A such complex system affected by several physical–chemical processes, natural or attributable to anthropogenic phenomena, appears to be an adequate case study to evaluate the effectiveness of wPCA.

This work has been developed by considering the following steps. The wPCA for CoDa and its graphical representations are instead in Section 2. Therefore, after describing the physical–chemical processes that affect the geochemistry of the Arno river waters, the results that have been obtained through the application of the wPCA can be seen in Section 3. The paper ends, in Section 4, with a brief summary and some concluding remarks.

2. THEORY

2.1. Compositional data and preprocessing

Let $v'_1, \ldots, v'_J$ be positive quantities with the same measurement scale K, the vector $\mathbf{v}' = v'_1, \ldots, v'_J$ is the basis of CoDa, and $\mathbf{v} = C(\mathbf{v}') = \mathbf{v}'/\sum_{j=1}^{J} v'_j$ is a composition vector, where C is defined as the closure operation. In other words, the closure operation normalizes any vector $\mathbf{v}'$ to a constant sum and defines a transformation $\Re_+^J \to S^J$, where $\Re_+^J$ is the positive orthant of $\Re^J$ defined by $\Re_+^J = \{v'_1, \ldots, v'_J : v'_1 > 0, \ldots, v'_J > 0\}$ and S^J is the simplex space defined by $S^J = \{v_1, \ldots, v_J : v_1 > 0, \ldots, v_J > 0; \sum_j v_j = \kappa\}$, where K is a given positive constant.

The natural sample space for CoDa is the simplex S^J, a subset of a $(J-1)$-dimensional subspace of $\Re^J$. Aitchison (1986) introduced two basic operations on the simplex, namely perturbation and powering, which introduce a real vector space structure on the simplex. This geometric structure, known as the Aitchison geometry, satisfies standard properties, such as compatibility of the distance with perturbation and powering transformation, and subcompositional coherence (Pawlowsky-Glahn and Egozcue, 2001). In the statistical literature, there are two approaches to analyze CoDa: moving from and staying in the simplex. The difference between them is more psychological than mathematical; see Mateu-Figueras (2011) for more details. In this paper, we follow the log-ratio transformation approach. Through this transformation, a direct association between the simplex sample space and the real space is found. Thus, it is possible to work in real space, where it is easier to apply statistical methods, and then, by using inverse functions, the results can be projected back to the simplex space.

Following this strategy, Aitchison (1982, 1986) proposed three kinds of transformation: pairwise (*plr*), additive (*alr*), and centered (*clr*) log ratios.

Through the *plr*, each element of a compositional vector, such as $\mathbf{v}(1 \times J)$, is transformed into the logarithm of the ratio $v_j/v_{j'}$ $(j < j')$ calculated between all the pairs of elements of the considered vector. In this way, the vector of pairwise log ratios, $\hat{z} = plr(v)$, has a dimension $(1 \times (J-1)J/2)$ with a generic element $\log\left(v_j/v_{j'}\right)$. The vector of additive log ratios, $\widetilde{z} = alr(v)$, has a dimension $(1 \times J - 1)$, where each element is given by $\log(v_j/v_{j*})$; therefore, with $j* = J$, we have $\widetilde{z} = \left[\log\left(v_1/v_j\right), \ldots, \log(v_{J-1}/v_J)\right]$. Finally, the vector of centered log ratios, $z = clr(v)$, has a dimension $(1 \times J)$ and a generic element $z = \log(v_j/g(v))$, with $g(\mathbf{v}) = g(v_1 v_2 \ldots v_J)^{1/J}$.

Egozcue *et al*. (2003) introduced an additional transformation called *isometric* log ratios (*ilr*), which is the only one that can be directly associated with an orthogonal coordinate system in the simplex. Let $\breve{z}(1 \times J - 1)$ be the vector of isometric log ratios, so that $\breve{z} = ilr(v)$, which is given by $z\Psi$, where $\Psi'\Psi = I_{J-1}$ and $\Psi\Psi' = \left(I_J - \hat{1}_J \hat{1}'_J/J\right)$, with I_J being the identity matrix $(J \times J)$ and $\hat{1}_J$ being a unit vector of J dimension. All these transformations have a role to play in the analysis of CoDa, and the choice is determined by the type of analysis that we have to use and the kind of application that we are interested in. Here, the *alr* is not discussed because it is an asymmetric transformation, and multidimensional analysis may fail. On the other hand, *ilr* transformations have very important properties that could also be applied to the multidimensional analysis. Nevertheless, the interpretation of their components is very difficult, and therefore, in this paper, we only take into account their mathematical properties.

2.2. Compositional data in three-way arrays

Let $\underline{V}$ $(I \times J \times K)$ be the three-way array where the I compositions, observed on K different occasions, are arranged in rows, whereas the J columns represent the parts of the compositions. The three-way array can also be seen as a collection of matrices, known as slices. There are three types of slices referred to as $(I \times J)$ frontal V_k with $k = 1, \ldots, K$, $(I \times K)$ vertical V_j with $j = 1, \ldots, J$, and $(K \times J)$ horizontal V_i with $i = 1, , I$. These can be concatenated in several ways as $\mathbf{V}_A = [\mathbf{V}_1 | \ldots | \mathbf{V}_k | \ldots | \mathbf{V}_K]$, $\mathbf{V}_B = [\mathbf{V}_1^t | \ldots | \mathbf{V}_k^t | \ldots | \mathbf{V}_K^t]$, and $\mathbf{V}_C = [\mathbf{V}_1 | \ldots | \mathbf{V}_i | \ldots | \mathbf{V}_I]$, which are, however, only some examples (for more details, see Kroonenberg, 2008; Gallo 2012b).

According to the discussion in Section 2.1, the logarithmic transformation can be applied to a three-way array $\underline{V}$; thus, $\underline{L}$ $(I \times J \times K)$ is an array with a typical element $\log(v_{ijk})$, and L_k is the kth frontal slice $(k = 1, \ldots, K)$. Hence, the kth frontal slice of the *clr* can be written as $L_k P_J^{\perp}$, where $P_J^{\perp} = \left(I_J - \hat{1}_J \hat{1}'_J/J\right)$ is the symmetric and idempotent centering matrix. The kth frontal slice of the *plr* can be written as $L_k \Xi$, where Ξ is a $(J \times J(J-1)/2)$ matrix with 0 s in each column except for 1 and -1 in two rows, because $\Xi\Xi^t = JP_J^{\perp}$. Finally, the kth frontal slice of the *ilr* can be written as $L_k P_J^{\perp} \Psi$. In addition, to ensure that the log ratios are centered with respect to the column means, each frontal slice is premultiplied by the symmetric and idempotent centering matrix $P_I^{\perp} = \left(I_I - \hat{1}_I \hat{1}'_I/I\right)$. So the kth frontal slice of *clr* is $Y_k = P_I^{\perp} L_k P_J^{\perp}$, whereas the kth columnwise-centered frontal slice for *plr* and *ilr* are $\hat{Y}_k = P_I^{\perp} L_k \Xi$ and $\hat{Y}_k = P_I^{\perp} L_k P_I^{\perp} \Psi$, respectively. Therefore, $\underline{Y}(I \times J \times K)$ is the three-way array where the data are *clr* preprocessed and columnwise centered. Like the three-way array $\underline{V}$, it is possible to obtain the

following representations of $\underline{Y}\, Y_A = [Y_1|\cdots|Y_k|\cdots|Y_k]$, $\mathbf{Y}_B = \left[\mathbf{Y}_1^t|\ldots|\mathbf{Y}_k^t|\ldots|\mathbf{Y}_K^t\right]$, and $\mathbf{Y}_C = [\mathbf{Y}_1|\ldots|\mathbf{Y}_i|\ldots|\mathbf{Y}_I]$. The other log-ratio transformed data can be arranged in a similar way.

2.3. The Tucker3 model for compositional data

It is well known that the Tucker analysis is a three-way generalization of PCA where loadings matrices are employed for each mode. A different number of loadings can be used in the different modes, and ($P < I$), ($Q < J$), and ($R < K$) are the number of loadings used to approximate the data for the first, second, and third modes, respectively. Moreover, the relationships between the loadings of each mode are captured by the elements of the three-way array $\underline{G}(P \times Q \times R)$, called *core* array. Specifically, the generic element of the core-array g_{pqr} gives the strength (or weight) of the linkage between the pth, qth, and rth loadings of the first, second, and third modes, respectively. So defining $A(I \times P)$ as the loadings matrix for compositions, B ($J \times Q$) as the loadings matrix for variables or parts, and C ($K \times R$) as the loadings matrix for occasions, the Tucker3 model for the centered log-ratio data can be written as

$$Y_A = AG_A(C \otimes B)^t E_A \tag{1}$$

where $\otimes$ is the Kronecker product; $\mathbf{G}_A = [\mathbf{G}_1|\ldots|\mathbf{G}_r|\ldots|\mathbf{G}_R]$ and $\mathbf{E}_A = [\mathbf{E}_1|\ldots|\mathbf{E}_k|\ldots|\mathbf{E}_K]$ are the juxtaposition of the frontal slices of the core and residuals array $\underline{E}$ ($I \times J \times K$), respectively.

When there is only one loading for the third mode, that is, $R = 1$, Equation (1) can be written as

$$Y_A = AG_1(c \otimes B)^t + E_A \tag{2}$$

where G_1 is the first frontal slice of core array, necessarily diagonal, and $\mathbf{c}$ contains the elements $[c_1, \ldots, c_k, \ldots, c_K]$. Equation (2) defines a very simple generalization of the PCA, known as wPCA (Ten Berge *et al.*, 1987). In fact, the K frontal slices of $\underline{Y}$ are given by the same AG_1B^t but weighted from the different coefficient C_{k1}.

To estimate the parameters of the wPCA, in the same way as for the Tucker3 model, several algorithms can be used. Thus, presuming that the wPCA model is perfectly fitting, the estimation of the parameters can be obtained in subsequent steps. The loadings $\mathbf{A}$ may be determined as the first P left singular vectors of Y_A, whereas the loadings $\mathbf{B}$ and $\mathbf{c}$ can be given by the first Q and *one* left singular vectors of Y_B and Y_C, respectively. And once $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{c}$ are fitted, the core array can subsequently be calculated as $G_1 = A^t Y_A(c \otimes B)$.

The only difference between this model and the traditional ones is that the loadings matrices for the first and second modes must have a column sum equal to zero: $\hat{1}_I A = \hat{1}_J B = \hat{0}$ ($\hat{0}$ is a zero vector). The Tucker3 algorithm previously described automatically respects these additional constraints. In fact, to determine the loadings matrices $\mathbf{A}$ and $\mathbf{B}$, we take the first P left singular vectors of Y_A and the first Q left singular vectors of Y_B, with $P = Q$ in case of wPCA. Now, in singular value decomposition (SVD), a generic matrix $\mathbf{X}$ of size ($I \times J$) with a rank equal to F can be decomposed into $X = U_F \Lambda_F V_F^t$, where U_F, V_F, and Λ_F are matrices containing both the singular vectors and values, $U_F U_F^t = V_F V_F^t = I_F$, and so if $\mathbf{X}$ is double centered $\left(\hat{1}_I X = \hat{0}_J, X\hat{1}_J = \hat{0}_I\right)$, then the result is $\hat{1}_I U = \hat{0}_F$ and $V\hat{1}_J = \hat{0}_F$. In the same way, the SVD of Y_A assures that the matrix of left singular vector $\mathbf{A}$ is orthonormal and column centered. Thus, it is easy to verify that these constraints are automatically respected.

Gallo (2012a) has shown that the loadings matrices of *clr*, *plr*, and *ilr* transformed data are strongly linked. In detail, the loadings matrices for the first and third modes are equivalent for the three transformations. In contrast, let $\hat{B}$ and $\breve{B}$ be the loadings matrices for the second mode in case of *plr* and *ilr* preprocessed data. It has been shown that $\widehat{B} = \Xi^t B$ and $\breve{B} = \Psi^t B$, where the matrices Ξ and Ψ have been previously defined. Hence, all the Tucker3 results on pairwise log-ratio data can be obtained by the analysis of smaller three-way arrays of centered log-ratio data. In the same way, it is possible to obtain the Tucker3 results on the isometric log-ratio data by the loadings matrices of the *clr* preprocessed data.

2.4. Procedures for displaying results

The Tucker3 results are usually given in the form of tables or plots (Kiers, 2000). Here, for the representation of the Tucker3 results of log-ratio data, we propose using one-mode plots, *clr*-joint biplots (Gallo, 2012a), and trajectory plots.

We are referring to one-mode plots when the components of a single mode are plotted against each other in scatter plots, that is, the first and second columns of $\mathbf{A}$. To provide some adequate representation of the higher-dimensional space of log-ratio data, it is necessary to choose carefully the low-dimensional space axes with respect to the compositions, parts, or occasions that are to be visualized and to compute coordinates with respect to these axes. Let $W_A = (c \otimes B)G_1^t$ be the base of $\mathbf{A}$, W_A is the orthonormal base for the principal coordinates of compositions $\mathbf{A}$. So, in this form, plotting the elements of $\mathbf{A}$ assures that the high-dimensional Aitchison distances are correctly represented in the low-dimensional plot within the accuracy of the approximation. Of course, $W_B = (c \otimes A)G_1^t$ is the orthonormal base for the principal coordinates of part $\mathbf{B}$ because of the symmetry of the model. Then, given the accuracy of the approximation, the distances between the parts in the low-dimensional plot of $\Re^{IK}$ indicate the relative variation.

In one-mode plots, only the relationships between the entities in the same mode can be investigated. The joint biplot can be used when the aim is to investigate both the relationship between the entities in the same mode and the relationship between the entities from the different modes.

Joint biplots are constructed by displaying simultaneously the components of the first and second modes for the component of the third mode. For the Tucker3 results, the matrix G_r can be decomposed by an SVD into $U_r \Lambda_r V_r^t$ so that Λ_r, U_r, and V_r are the matrices of the singular values and the orthonormal left and right singular vectors, respectively (Kroonenberg, 2008). Then the joint biplot for rth occasions is obtained by plotting the compositions as points, with coordinates given by the rows of $A'_r = AU_r\Lambda_r^a$, and the parts as rays, with coordinates

given by the rows of $B'_r = BV_r\Lambda_r^{1-a}$ with $\alpha \in [0,1]$. To assure a symmetric scaling of the loadings, α is chosen equal to 0.5, but different values can also be considered.

Given the joint biplot for centered log ratios, it is easy to obtain the joint biplot for pairwise log ratios. In fact, the only difference is that the rays have the coordinates $B'_r = BV_r\Lambda_r^{1-a}$. If the joint biplots for *clr* and *plr* are built following defined rules, then they can lead to a correct interpretation of the results (Gallo, 2012a). In case of wPCA, the matrix G_1 is diagonal. So the coordinates for the compositions are given by $A' = AG_1^{\alpha}$, and the coordinates for the parts are given by $B' = BG_1^{1-\alpha}$ with $\alpha \in [0,1]$.

In cases where there is a sequence of measurements, it could be of interest to visualize the location of combinations of the entities of two modes, that is, compositions and occasions. In these cases, let $B' = BG_1^{1-\alpha}$ be the coordinates for the parts, and then the coordinates of the compositions observed on the k different occasions $A'_k = c_kAG_1^{\alpha}$ ($k = 1, \ldots, K$) are plotted together to see how the location of the compositions in the space spanned by the parts changes over the occasions.

3. RESULTS

3.1. Case study

By considering the impact of water pollution on the environmental and ecological characteristics of a river, it is important to model the situation in space–time coordinates with the aim of having a reference framework to monitor the change in progress. Thus, to study the complex system affected by several physical–chemical processes, natural or attributable to anthropogenic phenomena present along the Arno river (central-northern Apennines, Italy), an extensive three-way array related to a sampling carried out between May 2002 and October 2003 (Nisi *et al*., 2008b; Nisi *et al*., 2008a) was analyzed.

The hydrographic catchment area of the Arno river basin, covering a surface of 8228 km^2 with an average elevation of 353 m, is entirely located in Tuscany (central Italy, http://www.arno.autoritadibacino.it). The Arno river, 242 km long, springs from the Northern Apennines at an elevation of 1650 m and flows into the Ligurian Sea, 10 km west from Pisa and 110 km from Florence (Figure 1, Nisi *et al*., 2008b). The annual rainfall pattern is typical of the Mediterranean area, with low regime in summer and two peaks of precipitation in winter (December and February). Mean annual rainfall values range from 600 mm, mainly in the low lands, up to 3000 mm on the Apennine ridge.

The outcropping rocks in the basin are predominantly sedimentary folded and faulted Mesozoic and Tertiary units resulting from the formation of the Apennine chain. The subsequent extensional tectonic phase has produced a NW–SE-oriented horst and graben system, made of Cretaceous to Paleogene allochthonous units, belonging to the Ligurian, sub-Ligurian, and Tuscan domains, being overthrusted mostly in the Early Miocene (e.g., Carmignani and Kligfield, 1990; Abbate *et al*., 1992; Carmignani *et al*., 1994; Moretti, 1994). The drainage network of the Arno river follows these NW–SE trending structures by six main sub-basins (Figure 1), from east to west: (i) Casentino (CA), (ii) Chiana Valley (CH), (iii) Sieve (SI), (iv) Upper Valdarno (UV), (v) Middle Valdarno (MV), and (vi) Lower Valdarno (LV).

Data are expressed in mg L^{-1}, a measure unit equivalent to ppm if data are multiplied for the density of the water (g cm^{-3}). When the saline content of water is low, mg L^{-1} and ppm are also numerically equal units, because the water density is practically equal to 1, the value

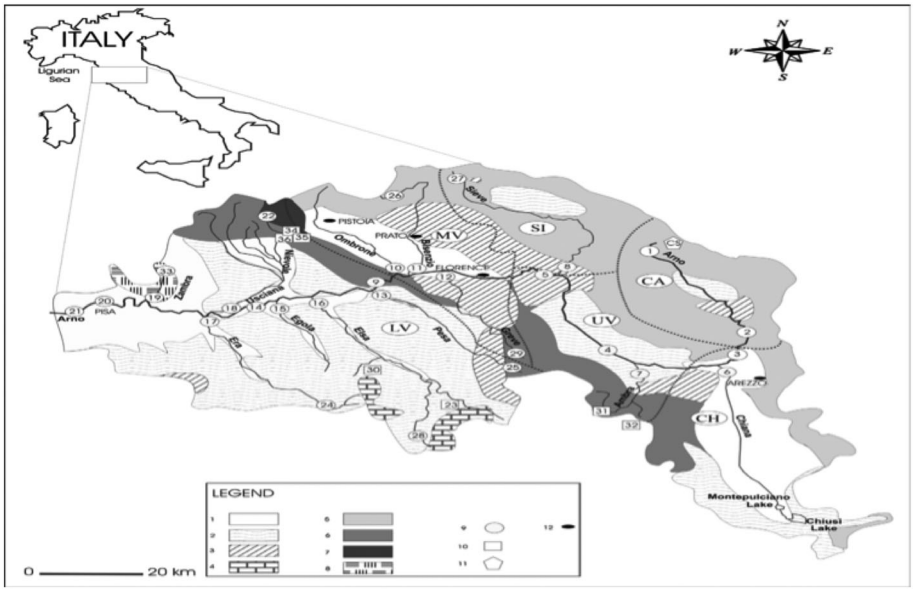

Figure 1. Schematic geological map of the Arno basin with the locations of the sampling sites along the river and in the whole basin. The catchment is divided into six sub-basins by dashed lines, as follows: Casentino (CA), Chiana Valley (CH), Sieve (SI), Upper Valdarno (UP), Middle Valdarno (MV), and Lower Valdarno (LV). Legend: 1, alluvial deposits; 2, clays, sands, and conglomerates of continental, lacustrine, and marine environments; 3, chaotic shaly rocks with calcareous and ophiolitic olistolitus; 4, dolomitic and evaporitic rocks; 5, Cervarola formation: turbiditic sandstones; 6, Macigno formation: turbiditic sandstones; 7, Modino formation: turbiditic sandstones; 8, Paleozoic formations; 9, running water samples; 10, thermal water samples; 11, rock samples; and 12, towns (Nisi *et al*., 2008b)

of pure water. Thus, from the compositional point of view, these measure units represent equivalent classes in the simplex geometry (Buccianti and Pawlowsky-Glahn, 2005; Buccianti, 2013).

Temperature and pH were measured in the field. SO_4^{2-}, Cl^-, and NO_3^- concentrations were determined by ion chromatography (Dionex DX100) and HCO_3^- by titration with 0.01 M HCl using methyl orange as indicator. Major cations (Ca^{2+}, Mg^{2+}, Na^+, and K^+) were determined by atomic absorption spectrometry (Perkin–Elmer AAnalyst 100). Dissolved NH_4^+ and NO_2^- were measured by molecular spectrophotometry (Hach DR2001). Analytical errors were generally <3% for the main components and 5–10% for trace species (Nisi *et al*., 2008b).

Values of pH varied from 7.29 to 8.63, indicating an alkaline nature typical of most surficial waters worldwide. The least mineralized site (upstream Zambra tributary) presented a total dissolved solids of 83 mg L^{-1}, whereas the most saline (Arno river near to the mouth) presented a value of 5344 mg L^{-1}, because of seawater intrusion (Cortecci *et al*., 2002; Sbolci *et al*., 2003). Anions were mostly represented by HCO_3^- contributing to 63% of the charge balance, whereas Cl^- and SO_4^{2-} only accounted for 13% and 7%, respectively. Among the cations, Ca^{2+} was the main dissolved species (57% of charge balance), followed by Na^+ (23%), Mg^{2+} (19%), and K^+ (1%). The concentrations of the nitrogenated species varied from <0.01 to 3.68 mg L^{-1} for NH_4^+, 0.003 to 3.36 mg L^{-1} for NO_2^-, and <0.01 to 17 mg L^{-1} for NO_3^-. According to nitrogen and oxygen isotope in nitrates, the surficial waters appear to be mainly affected by mineralized fertilizer, soil-organic nitrogen, manure, and septic waste (Nisi *et al*., 2005).

3.2. Preprocessing and analysis

The three-way compositional array (main chemical composition of water, distance from the spring, and time of collection) was given only for the samples from the main stream of the Arno river, 92 compositions related to four occasions, May 2002 (May02), January 2003 (Jan03), May 2003 (May03), and October 2003 (Oct03). The 23 compositions, associated to different spatial coordinates, indicating distance from the spring, were arranged by rows, the 11 main chemical compositions (parts) by columns and the four occasions by tubes.

In accordance with the approach discussed in Section 2.1, the *clr* transformation was applied to the three-way array, and subsequently, the data were centered across the second mode; thus, the data were logarithm-scaled so that the average of each row and each column was zero.

A Tucker3 analysis was performed on these preprocessed data by the use of algorithm developed for the package R 2.11.1 GUI 1.34 Leopard. With the aim of choosing the best dimensionality of the model, the Timmerman and Kiers (2000) procedure was also applied. It suggests a dimensionality-selection procedure analogous to Cattell's scree plot for two-mode component analysis.

The results suggested a choice of the model that had, proportionally, the highest fitted sum of squares SS_T. The selected model has the same total number of components $T = P + Q + R$, where P, Q, and R are the number of factors for each mode. In Figure 2, the residual sum of squares was plotted versus the number of components to select the model that had the smallest residual sum of squares within the class of the Tucker3 model with the same total number of components. Instead, to compare classes with a different number of components, the difference $dif_T = SS_T - SS_{T-1}$ is then determined, and only those that are sequentially higher were chosen. Finally, the model for which b_s had the highest value was selected, with b_s being the ratio between dif_T and the next highest value after dif_T. The results (reported in Table 1) suggested the choice of a model with two dimensions on the first and second modes and only one for the third mode with a total of five components and with a 60.02% of variability explained.

3.3. Results and discussion

For an in-depth interpretation of wPCA, we have used one-mode plots for the first two modes and *clr*-joint biplots to study the relationship between the elements of the two different modes.

In Figure 2(a), it is possible to interpret the relationships among the variables, where the principal coordinates are the same across all the occasions. The length of the arrows estimates the standard deviation of the variables rationed to the geometric mean on a logarithmic scale. In terms of centered log ratios, the distance between the tips of the arrows, known as links, estimates the standard deviation of the ratios between chemical species. Finally, the angle between the arrows estimates the correlation between the variables.

As we can see, Cl^- (and Na^+) showed the most important variation, followed by NO_2^- and NH_4^+. Cl^- as a product of silicate weathering and carbonate rocks is extremely low in river water, whereas a contribution of about 13% of all the sources on a global scale is expected from atmospheric cyclic salts (Berner and Bernes, 1996). Consequently, a high variation has to be attributable to different phenomena such as the presence

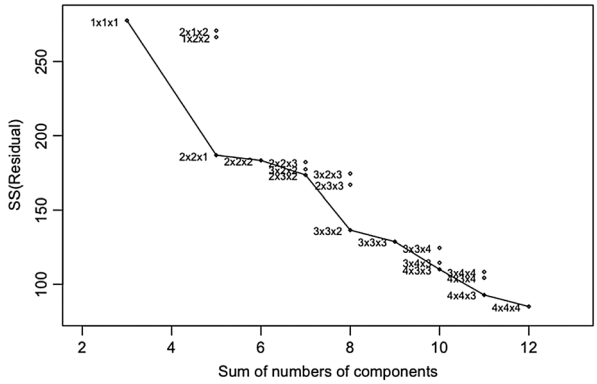

Figure 2. Scree plot of the residual sum of squares versus the number of components

Table 1. Statistical results of the application of several Tucker models

Tucker3 model	T	SS_T	dif_T	b_s
$1 \times 1 \times 1$	3	0.4058960	0.405896034	2.0886537
$2 \times 2 \times 1$	5	0.6002298	0.194333814	25.7807679
$2 \times 2 \times 2$	6	0.6077678	0.007537937	0.3584267
$2 \times 3 \times 2$	7	0.6287984	0.021030623	0.2658147
$3 \times 3 \times 2$	8	0.7079160	0.079117593	4.7591872
$3 \times 3 \times 3$	9	0.7245402	0.016624182	0.4152881
$4 \times 3 \times 3$	10	0.7645707	0.040030478	1.0841270
$4 \times 4 \times 3$	11	0.8014948	0.036924158	2.2172972
$4 \times 4 \times 4$	12	0.8181476	0.016652778	—

T is the total number of components, SS_T is the fitted sum of squares, dif_T is computed by the difference between SS_T and SS_{T-1}, and b_s is a crucial value equal to $b_s = dif_T/dif_{T*}$, where dif_{T*} has the next highest value after dif_T.

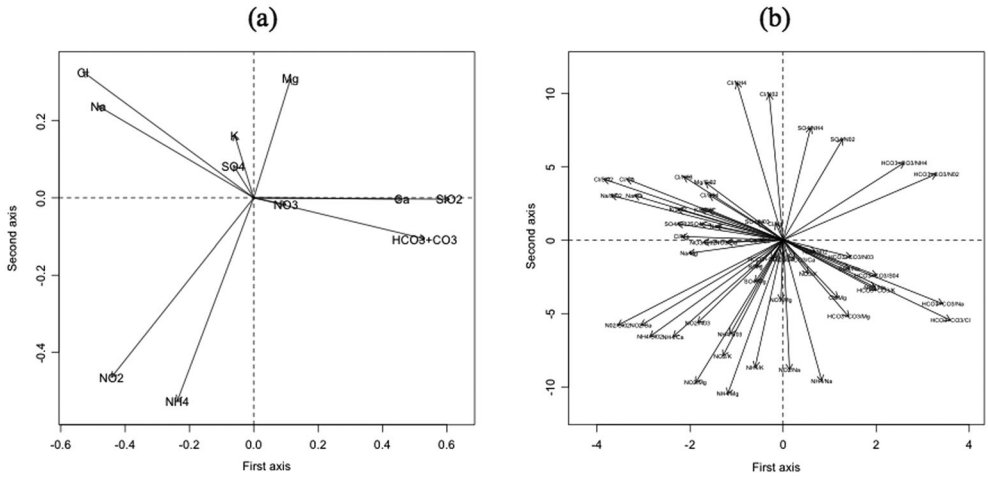

Figure 3. Plot of the relative variation of the parts (variables): (a) centered log-ratio transformed nd (b) pairwise log-ratio transformed. Variables and relative labels: sulfate (SO_4), chlorine (Cl), nitrate (NO_3), nitrite (NO_2), sodium (Na), ammonium (NH_4), carbonates ($HCO_3 + CO_3$), magnesium (Mg), potassium (K), silica (SiO_2), and calcium (Ca)

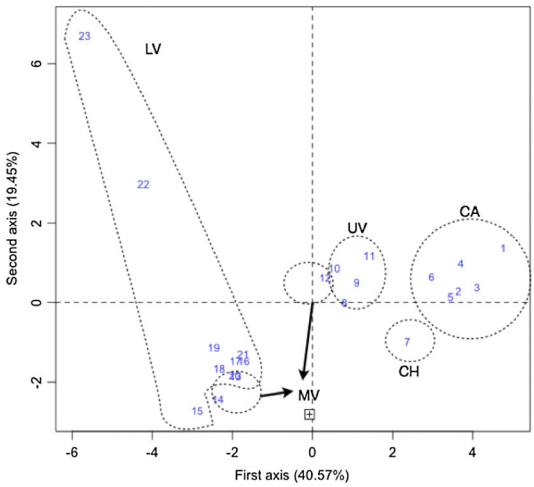

Figure 4. Object plot of the sampling sites. The catchment is divided into six sub-basins by dashed lines, as follows. Casentino (CA): La Casina (1), Pratovecchio (2), Ponte a Poppi (3), Rassina (4), Subbiano (5), and Buon Riposo (6). Chiana Valley (CH): Ponte a Buriano (7). Upper Valdarno (UV): Ponte del Romito (8), S. Giovanni Valdarno (9), Incisa (10), and Rosano (11). Middle Valdarno (MV): Ponte San Niccolò Firenze (12), Ponte a Signa (13), and Camaioni (14). Lower Valdarno (LV): Montelupo Fiorentino (15), Empoli (16), Colle Alberti (17), Castelfranco (18), Calcinaia (19), San Giovanni Alla Vena (20), Caprona (21), Pisa (Ponte Solferino) (22), and Arno Vecchio (23)

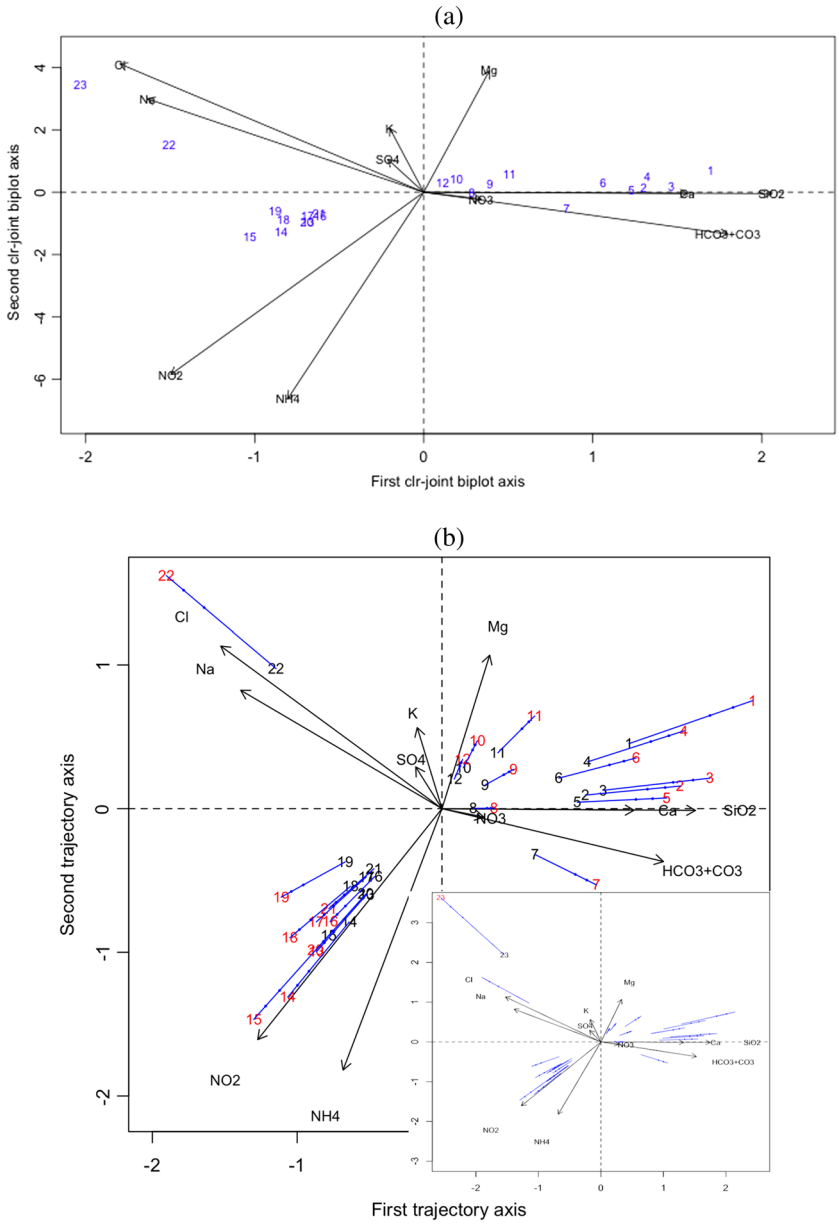

Figure 5. *clr*-joint biplot of the compositions and the parts (a). Trajectory plot for centered log ratios (b): the compositions of January 2003 are in black and the compositions of May 2003 are in red. Labels for compositions: La Casina (1), Pratovecchio (2), Ponte a Poppi (3), Rassina (4), Subbiano (5), Buon Riposo (6), Ponte a Buriano (7), Ponte del Romito (8), San Giovanni Valdarno (9), Incisa (10), Rosano (11), Ponte San Niccolò Firenze (12), Ponte a Signa (13), Camaioni (14), Montelupo Fiorentino (15), Empoli (16), Colle Alberti (17), Castelfranco (18), Calcinaia (19), San Giovanni Alla Vena (20), Caprona (21), Pisa (Ponte Solferino) (22), and Arno Vecchio (23)

of evaporation, pollution, or, as in our case, to the contribution of sea water intrusion perturbing the samples near the mouth of the river, all contributions that suffer the effect of seasonality. Na^+ showed the same pattern even if a major contribution from silicates weathering (22% on a global scale) was generally expected. High variation of the nitrogen species, NO_2^- and NH_4^+, was attributable to anthropogenic sources, changing in time (e.g., seasonality and water availability) and space (e.g., soil use), and to the chemical reactions that transformed the reduced forms into the more stable species, NO_3, in oxidized environment. The variation affecting the SiO_2, Ca^{2+}, $HCO_3^- + CO_3^{2-}$, and Mg^{2+}, which were mainly related to the weathering of the sedimentary rocks characterizing the basin, appeared to suffer the effect of time, whereas the behavior of SO_4^{2-} and K^+ indicated that their sources (evaporation and pollution for the first and mainly silicates for the second one) were not changed.

Considering the angle between the arrows, a high correlation could be derived for Na and Cl, stressing their similar geochemical source in our case, and between variables contributed by the weathering of sedimentary rocks (e.g., Ca^{2+}, SiO_2, and $HCO_3^- + CO_3^{2-}$). The high variability/correlation of NO_2^- and NH_4^+ if compared with NO_3^- described well the instability of the first two species in the superficial environment and the natural passage from reduced to oxidized (stable) forms.

Unfortunately, these approaches do not preserve subcompositional coherence, so if we need to maintain this property, it would be appropriate to examine the links in Figure 3(a) or plot the pairwise log-ratio coordinates (Figure 3(b)).

For example, in Figure 3(a), considering the distance between the tips of the rays of Na^+ and Cl^-, it was possible to evaluate that the ratio Na^+/Cl^- had maintained its values constant in time, which was an indication of the preservation of the environmental conditions affecting the behavior of the variables. A similar conclusion could be drawn for SO_4^{2-} and K^+ and for SiO_2 and Ca^{2+}. As can be seen, the points where the links intersect each other were very distant from the origin of the plot. Thus, it was not possible to study the correlations between the ratios of parts. In this case, we would consider the values of the cosine of the angles between the rays in Figure 3(b), which are equivalent to the cosine of the angles between the links in centered log-ratio plots.

In Figure 4, the results obtained by the investigation of spatial coordinates are shown. Considering all the covered period, and consequently the effect of seasonality (sampling in May 2002 and January, May, and October 2003), most of the samples were located on a linear pattern that in part corresponded to the distance from the source and to the different sub-basins that occurred along the river's path. Marked deviations characterized only two localities where the closeness to the mouth (influence of sea water intrusion) strongly modified the chemical composition. It is noteworthy here that the samples pertaining to the Casentino (CA) sub-basin, when compared with the others (Figure 1), were characterized by a lower variability in time, which was an expected result because here the chemical composition of waters was mainly affected by natural weathering phenomena due to the pristine nature of the area. More scattering appeared to perturb data of the other sub-basins, indicating that time (seasonality) could have had an effect on natural and anthropogenic processes (for example, use of soils and pressure of cities and industrialized areas).

On the diagram, cases 1 (La Casina, CA), 11 (Rosano, UV), 22 (Pisa, Ponte Solferino, LV), and 23 (Arno Vecchio, LV) were located in the opposite part with respect to cases 14 (Camaioni, MV) and 15 (Montelupo Fiorentino, LV). The last cases were characterized by an important contribution of NH_4^+ along the river's path, indicating that these places suffered pollution in the time of sampling when compared with all the others.

The one-mode plots give information only on each single mode, but there is no relationship between the elements plotted in different one-mode plots. To investigate the relationship between the elements of different modes, the simultaneous plotting of the compositions and parts may highlight the relation between them across the occasions. Figure 5(a) shows the *clr*-joint biplot where the variables are plotted in standard coordinates and the compositions in principal coordinates. Moreover, they are regulated by factor scales $[(I \times K)/J]^{75}$ and $[(J \times K)/I]^{25}$, respectively. In Figure 5(b), it was possible to analyze the trajectories for some compositions as they move across the space. The *clr*-joint biplot of the compositions, the parts (Figure 5(a)), and the trajectories for centered log ratios allowed us to identify sites where ratios between variables present different behaviors in time and space. For example, in the case of samples 23 (Arno Vecchio) and 22 (Pisa, Ponte Solferino), the Mg^{2+}/NH_4^+ ratio was maintained constant, indicating that potential environmental changes had affected both the chemical species during the considered interval of time. On the other hand, changes had also influenced the Cl^-/Na^+ ratio. Here, it is important to note that both the samples were collected near the mouth of the river. Sample 1 (La Casina) was characterized by an increase in the ratio Mg^{2+}/NH_4^+, revealing for the same variables a possible perturbation due to different sources (increase of Mg^{2+} or decrease of NH_4^+, perhaps due to seasonality). The sample was collected at the spring of the river. Thus, for different spatial positions, a different behavior of two rationed variables can be pointed out.

On the whole, most of the samples were characterized by changes in time that affected the variables related to pollution, such as NO_2^- and NH_4^+, or concerned the main components of the water chemistry, that is, Ca^{2+}, SiO_2, and $HCO_3^- + CO_3^{2-}$, typically related to water/rock interactions in the considered basin lithology. It is evident from this analysis that the samples related to the two extreme conditions (spring and mouth) tend to show some chemical features that distinguish their behavior with regard to all the others.

4. Conclusions

A particular version of the Tucker models for CoDa was proposed to analyze the chemical composition of surficial waters pertaining to the Arno river. Gallo (2012a) has shown how it is possible to apply the Tucker3 model for CoDa by discussing the results obtained when different kinds of preprocessing are used. Here, this approach was proposed for wPCA model, and a full interpretation of the results is given. Moreover, the use of some graphical methods in regard to CoDa such as one-mode and trajectory plots was considered.

The results encourage the application of these kinds of tools to display the information concerning the chemical composition of surficial water when data are collected in different sampling locations and periods. In fact, by employing the one-mode plot for the first mode, it was possible to detect the differences between the six sub-basins of the Arno river. By using the one-mode plot for parts (*clr* and *plr* version), it was possible to extract useful information on the correlations between the ratios of parts and the variability of parts across the samplings. Finally, on the *clr*-joint plot, the behavior of the compositional variables for the different ratios could be read as well as the trajectories of the sampling locations across the seasonality on trajectory plots. On the whole, the approach appeared to be able to give an exhaustive and interpretable framework of a complex geochemical system whose equilibrium moving in space and time.

Acknowledgements

Comments and suggestions by the editor and two anonymous reviewers have greatly improved an early version of the manuscript. This work was financially supported by ex-60% 2011 funds of the University of Naples "L'Orientale" (I) and by ex-60% 2012 funds of the University of Florence (I).

REFERENCES

Abbate R, Castellucci P, Ferrini GL, Pandeli E. 1992. I dintorni di Firenze. Società Geologica Italiana. In Guide Geology Regulatory **4**: 214–223.
Aitchison J 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) **44**(2): 139–177.
Aitchison J 1983. Principal component analysis of compositional data. Biometrika **70**(1): 57–65.
Aitchison J 1986. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd: London.
Berner EK, Bernes RA. 1996. Global Environment. Water, Air, and Geochemical Cycles. Prentice Hall, Upper Saddle River: New Jersey.

Buccianti A, Tassi F, Vaselli O. 2006. Compositional changes in a fumarolic field, Vulcano Island, Italy: a statistical case study. In *Compositional Data Analysis in the Geosciences: from theory to practice*, Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds.), Geological Society: London, Special Publications, **264**: 67–77.

Buccianti A. 2011. Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. In Compositional Data Analysis. Theory and Applications, Pawlowsky-Glahn V, Buccianti A (eds.), John Wiley & Sons, Ltd: Chichester, West Sussex, UK; 255–266.

Buccianti A, Magli R. 2011. Metric concepts and implications in describing compositional changes for world river's water chemistry. Computers and Geosciences **37**(5): 670–676.

Buccianti A. 2013. Is compositional data analysis a way to see beyond the illusion? Computers & Geosciences **50**(2013): 165–173.

Buccianti A, Pawlowsky-Glahn V. 2005. New perspectives on water chemistry and compositional data analysis. Mathematical Geology **37**(7): 703–727.

Carmignani L, Kligfield R. 1990. Crustal extension in Northern Apennines: the transition from compression to extension in the Alpi Apuane core complex. Tectonics **9**: 1275–1303.

Carmignani L, Decandia FA, Fantozzi PL, Lazzarotto A, Liotta D, Meccheri M. 1994. Tertiary extensional tectonics in Tuscany (northern Apennines Italy). Tectonophysics **23**(8): 295–315.

Cortecci G, Dinelli E, Bencini A, Adorni Braccesi A, La Ruffa G. 2002. Natural and anthropogenic SO_4 sources in the Arno river catchment, Northern Tuscany, Italy: a chemical and isotopic reconnaissance. Applied Geochemistry **17**: 79–92.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology **35**(3): 279–300.

Gallo M 2003. Partial least squares for compositional data: an approach based on the splines. Italian Journal of Applied Statistics **15**: 349–358.

Gallo M, 2010. Discriminant partial least squares analysis on compositional data. Statistical Modelling **10**(1): 41–56.

Gallo M. 2012a. Tucker3 model for compositional data. Communications in Statistics – Theory and Methods (in press).

Gallo M 2012b. CoDa in three-way arrays and relative sample spaces. Electronic Journal of Applied Statistical Analysis **5**: 401–406.

Gallo M 2013. Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data. In Advanced Dynamic Modeling of Economic and Social Systems, Proto AN, Squillante M, Kacpzyk J (eds.), Springer-Verlag: Berlin Heidelberg, **448**: 209–221.

Hinkle J, Rayens W. 1995. Partial least squares and compositional data: problems and alternatives. Chemometrics and Intelligent Laboratory Systems **30**: 159–172.

Kiers HAL. 2000. Some procedures for displaying results from three-way methods. Journal of Chemometrics **14**(3): 151–170.

Kroonenberg PM. 2008. Applied Multiway Data Analysis. Wiley: New Jersey.

Kroonenberg PM, De Leeuw J. 1980. Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrike **45**: 69–97.

Lombardo R, Carlier A, D'Ambra L. 1996. Non-symmetric correspondence analysis for three-way contingency tables. Methodologica **4**: 59–80.

Martin-Fernandez JA, Barcelo-Vidal C, Pawlowsky-Glahn V. 1998. Measures of difference for compositional data and hierarchical clustering. In *IAMG'98*, Buccianti A, Nardi G, Potenza R. (eds.), De Frede: Napoli.

Mateu-Figueras G 2011. Elements of simplicial linear algebra and geometry. In *Compositional Data Analysis: Theory and Applications*, Pawlowsky-Glahn V, Buccianti A (eds.), Wiley: Chichester.

Moretti S 1994. The Northern Apennines. *Proceeding 76th Summer Meeting of the Italian Geological Society* **3**: 739–956.

Nisi B, Vaselli O, Buccianti A, Silva SR. 2005. Sources of nitrate in the Arno river waters: constraints $d^{15}N$ and $d^{18}O$. GeoActa **4**: 13–24.

Nisi B, Vaselli O, Buccianti A, Minissale A, Delgado-Huertas A, Tassi F, Montegrossi G. 2008a. Geochemical and isotopic investigation of the dissolved load in the running waters from the Arno valley: evaluation of the natural and anthropogenic input. In *Memorie Descrittive della Carta Geologica d'Italia*, Nisi (eds.),LXXIX; 157--.

Nisi B, Buccianti A, Vaselli O, Perini G, Tassi F, Minissale A, Montegrossi G. 2008b. Hydrogeochemistry and strontium isotopes in the Arno river basin (Tuscany, Italy): constraints on natural controls by statistical modeling. Journal of Hydrology **360**: 166–183.

Pawlowsky-Glahn V, Buccianti A. 2002. Visualization and modeling of sub-populations of compositional data: statistical methods illustrated by means of geochemical data from fumarolic fluids, 2002. International Journal of Earth Sciences **91**(2): 357–368.

Pawlowsky-Glahn V, Egozcue JJ. 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment **15**(5): 384–398.

Sbolci N, Bencini A, Malesani P. 2003. A study on salt intrusion in Arno river bed and surrounding plain in Tuscany coastal area, central Italy. In Congress on Tecnologia de la intrusion de agua de mar en acuiferos costieros: Paises Mediterraneos. IGME: Madrid; 11–18.

Ten Berge JMF, De Leeuw J, Kroonenberg PM. 1987. Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms. Psychometrika **52**: 183–191.

Timmerman ME, Kiers HAL. 2000. Three-mode principal components analysis: choosing the number of components and sensitività to local optima. British Journal of Mathematical and Statistical Psychology **53**: 1–16.

Tucker LR. 1966. Some mathematical notes on three-mode factor analysis. Psychometrika **31**: 279–311.

