Journal of Geochemical Exploration 258 (2024) 107385

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Exploring geochemical data using compositional techniques: A practical guide

Juan José Egozcue [a], Caterina Gozzi [b,*], Antonella Buccianti [b], Vera Pawlowsky-Glahn [c]

[a] Department of Civil and Environmental Engineering, Technical University of Catalonia – BarcelonaTech, C/ Jordi Girona, 1-3, Barcelona 08034, Spain
[b] Department of Earth Sciences, University of Firenze, Via G. La Pira 4, Firenze 50121, Italy
[c] Department of Computer Sciences, Applied Mathematics and Statistics, University of Girona, C/ de la Universitat de Girona, 6, Girona 17003, Spain

ARTICLE INFO

Keywords:
Compositional data
CoDa-biplot
Isometric log-ratio
CoDa-dendrogram
Principal balance coordinates
Index of proportionality

ABSTRACT

John Aitchison revolutionised in 1982 our way of approaching geochemical data focusing on their relative nature. In this perspective, the investigation of single variables is meaningless due to the entangled structure that links all the parts of a composition. Starting from that time, several developments have characterized the debate within the scientific community, both from the applied and the theoretical point of view. The consequence was that the number of papers where compositional data are consistently and coherently managed increased exponentially. The exploratory phase of compositional data is a very important step in data analysis and modeling. It not only helps to clarify the available sample data structure but also determines the base to develop models to predict time and space changes. Real chemical data along the course of the river Tevere (Tiber) (Italy) and its tributaries are taken to illustrate how compositional techniques help explore compositions and detect patterns and outliers in the data.

1. Introduction

Scientists confronted with the need to investigate environmental issues or the geological evolution of a certain region proceed in general to design a sampling campaign that involves mainly geochemical elements. Soon they have to handle the fact that this type of data is not suitable for standard statistical analysis, as many techniques designed for real random variables, especially those involving correlations or covariances, are prone to deliver spurious results. Here we pretend to guide the reader through all the initial steps in a statistical analysis of compositional data.

For illustration, standard compositional techniques are used to explore geochemical water data sampled along the Tevere River (Tiber) and its tributaries in central Italy. The catchment drains an area of 17,375 km^2 and presents high hydrogeological and morphological heterogeneity associated with significant anthropogenic inputs. The question to be addressed is to determine geochemical changes in the water along the catchment divided into four sections or sub-basins (Gozzi et al., 2020, 2021). The choice of the sub-basins was guided by the drainage network structure and by the different outcropping lithology.

The very basics of compositional data analysis are assumed to be known by the reader, including the Euclidean space structure of the simplex that gives rise to the Aitchison geometry as defined in Pawlowsky-Glahn and Egozcue (2001). Concepts like compositional perturbation and powering, closure operation, and subcomposition are used in the manuscript. Their definitions and interpretations can be found in many papers and books. Here we recommend Pawlowsky-Glahn et al. (2015); Boogaart and Tolosana-Delgado (2013); Egozcue and Pawlowsky-Glahn (2019); Filzmoser et al. (2012) among many others.

2. The data and research questions

The data consist of $N = 222$ samples, made of $N_{HT} = 96$ corresponding to the high Tevere (HT), $N_{MT} = 33$ to the medium Tevere (MT), $N_{NE} = 42$ to the Nera sub-basin, and $N_{LT} = 51$ to the lower Tevere (LT). The data were collected during two campaigns in 2017 (160 samples) and 2018 (62 samples). A detailed geological context of the Tevere-Nera basin and a study of the 2017 data are presented in Gozzi et al. (2019). Fig. 1 shows the sampling locations within the 4 sub-basins.

Beyond the metadata, each observation consists of a composition of $D = 9$ dissolved ions in mg/L (HCO_3^-, F^-, Cl^-, NO_3^-, SO_4^{2-}, Ca^{2+},

* Corresponding author at: Department of Earth Sciences, University of Firenze, Via G. La Pira 4, Firenze 50121, Italy.
E-mail address: caterina.gozzi@unifi.it (C. Gozzi).

https://doi.org/10.1016/j.gexplo.2024.107385
Received 13 October 2023; Received in revised form 12 December 2023; Accepted 18 December 2023
Available online 10 January 2024

Mg^{2+}, Na^{+}, K^{+}). These ions are denoted without the charge signs along the remainder of the article to simplify the notation. Additionally, the following geochemical parameters are also reported: height in m over sea level (h); water temperature in degrees Celsius (T); pH; conductivity (CND, μS/cm); redox potential (Eh, mV).

The purpose of these campaigns was mainly exploratory to investigate, in a first attempt, water-rock interaction processes in a catchment that had received very little attention in the past. However, research questions are of utmost importance to guide both the sampling design and the consequent analysis (Gozzi and Buccianti, 2022). Recall that a preliminary step in any data analysis should be to identify which is the sample space of the data. This is always an assumption since there are multiple choices. One important criterion for selecting an appropriate sample space is the fact that research questions need to be answered within the framework of the sample space. For instance, if differences (in mean) between sub-basins are of interest, a difference between samples and a distance between them needs to be selected. In the present case, this is readily done by assuming that the concentrations of reported ions are compositions and that the data follow the Aitchison Geometry of the simplex (Pawlowsky-Glahn and Egozcue, 2001). Then, the difference between compositions is assumed to be the negative perturbation (Aitchison, 1982), and the distance the Aitchison's one (Aitchison, 1992). Also, the sample space of co-variables like the height or elevation h of the sample point must be decided. Here, h was taken as a real random variable, that is, the differences are computed by subtraction like in $h_2 - h_1$, and the distance is then $d(h_1, h_2) = |h_2 - h_1|$. However, there are other possibilities; for instance, since heights are positive, it can be assumed that the sample space is the positive half-axis of real space and the distance between two heights is $|\ln(h_2) - \ln(h_1)|$. The consequence of this assumption would be that the difference of 1 cm in the low course of the river is much more important than the difference of 1 cm in the high course of the river.

3. Exploratory analysis

In most statistical analyses, the first step is to have a quick look at the available data. The goals of this are multiple, from detecting errors and missing data, or showing main features, up to distributional characteristics of the variables, including their relationships if any. The particular characteristics of compositional data impose also special ways to explore the data. One important characteristic of CoDa is that the values of the components are not fully meaningful unless some reference is given. In our case, the measurement of an ion like e.g. Ca in mg has a meaning when referred to a liter (L), i.e. when the units are expressed as mg/L. However, note that in most cases the reference liter (L) is fictitious since it was not measured but idealized in a calibration process. In compositional analysis, the possible references are the other components, called parts, of the composition. This is attained by considering the ratios between parts. For instance, the dimensionless ratio HCO_3/Ca does not need the liter as a reference as it cancels out in the ratio. This turns the univariate summary of parts of little use. Even so, a quantile exploration of the parts can be useful for detecting the presence of zeros, errors, and

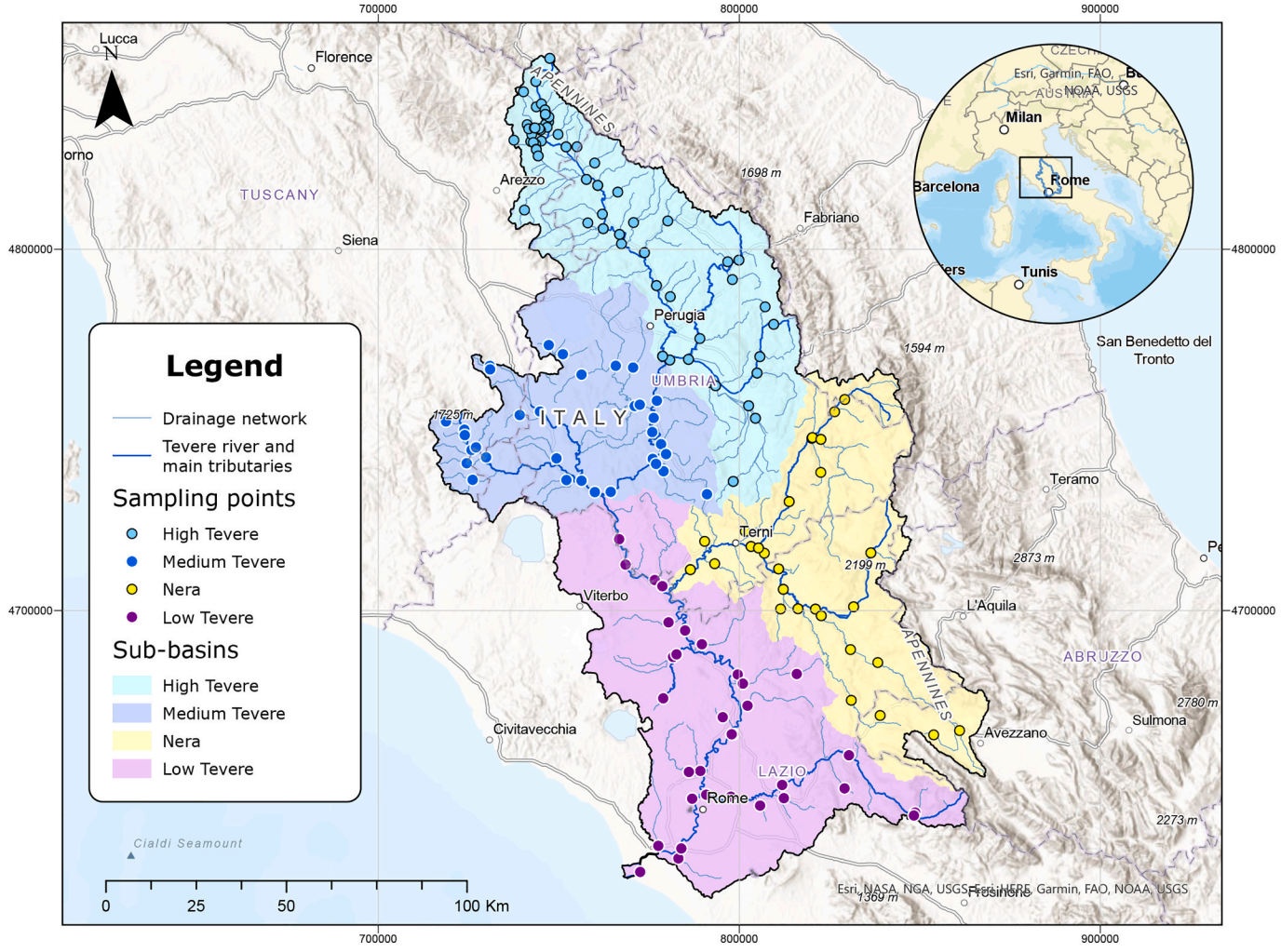

Fig. 1. Map showing the Tevere River basin along with sampling points locations within the 4 sub-basins.

missings. Table 1 shows a quantile table for the water composition (four sub-basins together).

In the standard multivariate exploration of real random variables, it is customary to compute the arithmetic average, variance, standard deviation, and correlation of the variables. For CoDa, these parameters are almost meaningless. As mentioned, relevant information is in ratios between parts. After averaging parts, a ratio of averages may have an undesirable behavior. Additionally, the correlations between parts have been recognized as spurious since they change in an uncontrolled way when changing subcomposition (Chayes, 1971; Aitchison, 1986).

These problems with the standard exploratory tools for real multivariate data require specific procedures for CoDa. The first tools needed are the compositional alternatives to the arithmetic averages, dispersion measures, and correlation for real variables which are discussed in Section 3.1. Section 3.2 examines the compositional sample of ion concentrations to detect outliers and treat them statistically.

3.1. Center, variation array and PIP-table

The compositional alternative to the sample mean or average for real variables is the center of the data. It is estimated as the compositional average

$$\mathrm{Cen}[X] = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \mathbf{x}_i = \mathscr{C}\left(\mathrm{g_m}(X_1), \mathrm{g_m}(X_2), \ldots, \mathrm{g_m}(X_D)\right), \tag{1}$$

where $\mathbf{x}_i$ are the compositional observations (rows), and N is the sample size. The $\oplus$ denotes repeated compositional perturbation along the sample and $\odot$ denotes powering. The columns of the data set, called parts, are denoted by X_j. Taking closure, $\mathscr{C}$, is optional as proportional compositions are equivalent (Aitchison, 1992; Barceló-Vidal and Martín-Fernández, 2016). In the case where the parts are concentrations in mg/L, suppressing the closure gives the center in mg/L; alternatively, taking closure gives the mean proportions of ions considered. Table 2 shows the center of the sample and its change when 6 outliers are removed (see Section 3.2). Table 2 allows us to compare centers in different sub-basins. Some differences are apparent. For instance, LT exhibits the highest concentrations of HCO_3, NO_3, Cl, and K, thus suggesting agricultural activity and alteration of rocks as corresponds to the low course of a river in a populated region. However, comparisons of centers expressed in mg/L are not easy. It is better to compute the Aitchison distances between compositional centers of sub-basins (Table 3). It is clear that the center of LT (cenLT) is the most distant from the overall center of the sample (cenallb) and also from other sub-basins. On the other hand, the centers of MT and NE (cenMT, cenNE) appear as equally distant from the cenallb although they are different themselves. Aitchison distances depend on the subcomposition where they are computed, but they are dominant. The smaller the subcomposition, the smaller the interdistances (Aitchison et al., 2000); in the present example comparisons may change, for instance, if HCO_3 and Cl are not included.

In order to compute distances between compositions, like the mentioned centers, there are several ways. One of them is to compute the centered log-ratio of the D-part compositions $\mathbf{x} = (x_1, x_2, \ldots, x_D)$

$$\mathbf{y} = \mathrm{clr}(\mathbf{x}) = \left(\ln\frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \ln\frac{x_2}{\mathrm{g_m}(\mathbf{x})}, \ldots, \ln\frac{x_D}{\mathrm{g_m}(\mathbf{x})}\right),$$

and then, the Euclidean distance between the clr's, which is equal to the Aitchison distance. For instance, the Aitchison distance between cenHT and cenLT is

$$\mathrm{d}_a(\mathrm{cenHT}, \mathrm{cenLT}) = \mathrm{d}_e(\mathrm{clr}(\mathrm{cenHT}), \mathrm{clr}(\mathrm{cenLT})),$$

where d_e is the ordinary Euclidean distance in the real space. The clr transformation of compositions is useful for several CoDa computations. This is the case of the centers (Eq. 1). The clr of the center is the arithmetic average of the clr's of the sample. Recovering the center from its clr is achieved taking exponential,

$$\mathbf{x} = \mathrm{clr}^{-1}(\mathbf{y}) = \mathscr{C}\exp(\mathbf{y}),$$

where the closure is again optional. Particularly, if all the compositions are given as mg/L, then the result of the exponential is also in mg/L.

The statistical analysis of these centers is based on the Aitchison distances and leads to analyses as presented in Section 4. Also comparisons of mean ilr-coordinates are useful for understanding which are the differences of centers between sub-basins (see Section 3.4).

The standard exploratory analysis of multivariate data includes a description of the co-variability of variables. This is done through the covariance and/or correlation matrices. In Compositional Data Analysis (CoDA) these tools are known to be *spurious* (Aitchison, 1986) since the entries of these matrices change depending on the subcomposition considered or the units in which the composition is expressed.

In Aitchison (1986) the variation matrix is proposed as a representation of the second-order variability of the compositional sample. For a D-part composition, the (i,j)-th entry of the variation (D,D)-matrix, T, is

$$\tau_{ij} = \mathrm{Var}\left(\ln\frac{X_i}{X_j}\right), \quad i,j = 1,2,\ldots,D,$$

where X_i, and X_j are the respective compositional variables, here called parts. The variation matrix T is symmetric and has zeros in its diagonal. The smaller is τ_{ij}, the closer to proportionality are the parts X_i and X_j. Conversely, if τ_{ij} is large, the log-ratio $\ln(X_i/X_j)$ contributes largely to the variability of the sample. Related to these facts, the total variance of the compositional sample is defined as

$$\mathrm{TotVar}(X) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\tau_{ij} = \sum_{j=1}^{D}\mathrm{Var}\left(\mathrm{clr}_j(X)\right), \tag{2}$$

where X_j is the j-th part (column) of the compositional sample X. Table 4 shows the upper triangle of T for the ion concentrations of the Tevere basin. The minimum entry is $\tau_{\mathrm{HCO_3,Ca}} = 0.08$ thus suggesting some association or proportionality between HCO_3 and Ca that finds a geochemical explanation in the common source of these species from carbonatic lithology.

Table 1
Univariate quantiles (mg/L) for ion concentration (parts). There is no zero concentration. Maxima can take values as large as 100 times the median value (e.g. Cl), suggesting the presence of outliers, special sampling conditions, or even errors. There are no missing values.

q-prob.	min	0.01	0.05	0.25	median	0.75	0.95	0.99	max
HCO_3	161.04	172.15	205.40	256.20	307.75	356.91	497.27	860.16	3101.24
F	0.04	0.05	0.07	0.11	0.18	0.34	1.01	2.28	2.90
Cl	3.57	4.30	5.97	9.53	15.58	27.55	118.42	530.94	1173.28
NO_3	0.01	0.02	0.07	0.95	3.09	6.33	15.05	28.46	33.16
SO_4	3.16	4.24	7.50	25.35	42.39	66.03	215.76	514.96	1569.29
Ca	38.24	43.52	55.62	75.83	90.36	110.53	164.86	354.63	585.69
Mg	1.12	1.87	4.97	11.94	16.93	22.60	45.28	84.76	420.72
Na	2.16	2.34	3.50	8.88	18.56	33.04	94.72	427.64	747.64
K	0.01	0.01	0.87	1.58	2.42	4.26	19.12	46.99	90.63

Table 2
Centers of ion composition (mg/L). First row: complete sample before removal of outliers. Subsequent rows, all sub-basins jointly, and sub-basins separately after removal of outliers.

ion (mg/L)	HCO_3	F	Cl	NO_3	SO_4	Ca	Mg	Na	*K*
compl.sample	312.06	0.21	18.31	2.14	43.69	94.07	16.18	17.89	2.57
after removal outliers	HCO_3	F	Cl	NO_3	SO_4	Ca	Mg	Na	*K*
all sub-basins	311.38	0.21	18.73	2.18	43.56	94.30	16.08	18.22	3.00
High Tevere (HT)	303.39	0.14	13.63	1.53	35.19	87.18	15.78	15.07	2.12
Medium Tevere (MT)	295.92	0.24	24.82	2.24	90.14	108.89	19.82	28.98	3.79
Nera (NE)	296.27	0.17	12.34	1.39	29.07	92.53	11.59	8.470	1.80
Low Tevere (LT)	350.69	0.50	38.47	5.82	54.91	100.44	18.79	34.67	7.23

Table 3
Aitchison distances between the center of all sub-basins (cenallb) and the sub-basins HT, MT, NE, LT, after removal of outliers.

	cenallb	cenHT	cenMT	cenNE
cenHT	0.43			
cenMT	0.66	0.81		
cenNE	0.66	0.62	1.16	
cenLT	1.04	1.45	1.29	1.52

From their introduction in Aitchison (1986) the entries of the variation array have been recognized as measures of association between parts, a sort of alternative to correlation. However, normalization seems to be convenient to maintain the values within $(0, 1)$ and to remove the effect of the total variance of the sample. Several indexes of association between parts have been proposed, but most of them depend on the selected subcomposition, thus preventing their use to examine the relation between parts (in our case chemical elements) (Egozcue and Pawlowsky-Glahn, 2023). In the latter reference, a proportionality index of parts (PIP) is proposed. It is based on the variation array but removes the effect of the total variance and maintains the subcompositional invariance and the range of values within $(0, 1)$. Its expression is

$$\mathrm{PIP}_{ij} = \frac{1}{1+\sqrt{\tau_{ij}}} \ .$$

The values of the PIP matrix are shown in the lower triangle of Table 4. The larger PIPs are in boldface on pale-red background thus pointing out which are the largest associations. The largest PIP corresponds to HCO_3 and Ca, although the value 0.78 suggests a weak relation. Moreover, HCO_3 appears also weakly linked to Mg (PIP 0.65), and Mg and Ca (PIP 0.64) seem also weakly associated, suggesting an association between the three elements. Geochemical experience indicates that this result can be due to the alteration of carbonate rocks, such as limestones or dolostones, where the cited elements are contained in minerals, often in a proportional way. On the other side, the values of PIP under discussion come from the overall Tevere basin. Since sub-basins have different behaviors, the mentioned associations could be stronger for individual sub-basins. This is not the case, only in the LT sub-basin does the PIP between HCO_3 and Ca rise up to 0.82, which is still a weak association. This confirms that the weathering of carbonate rocks is a fundamental process affecting the dynamic of the water of the catchment.

Table 4
Upper triangle: variation matrix for all sub-basins after removing outliers. Lower triangle: PIP for all sub-basins after removing outliers. No strong association is detected, although weak associations of HCO_3 with Ca and Mg, and Cl with Na are suggested (colored cell and bold case).

	HCO_3	F	Cl	NO_3	SO_4	Ca	Mg	Na	K
HCO_3	0	0.77	0.71	2.57	0.88	0.08	0.29	0.80	0.80
F	0.53	0	0.87	2.12	0.89	0.72	1.02	1.14	0.54
Cl	0.54	0.52	0	2.59	0.69	0.64	0.70	0.17	0.76
NO_3	0.38	0.41	0.38	0	3.07	2.65	3.09	2.92	2.20
SO_4	0.52	0.51	0.55	0.36	0	0.63	0.58	0.81	1.14
Ca	**0.78**	0.54	0.55	0.38	0.56	0	0.33	0.82	0.91
Mg	**0.65**	0.50	0.55	0.36	0.57	**0.64**	0	0.69	0.92
Na	0.53	0.48	**0.71**	0.37	0.53	0.53	0.55	0	0.68
K	0.53	0.58	0.53	0.40	0.48	0.51	0.51	0.55	0

3.2. Outliers

The univariate analysis of parts of a composition, in our case concentrations of dissolved ions, does not give reliable clues about outliers since CoDa is multivariate by nature. In Fig. S.1 univariate candidate outliers appear out of the boxplot whiskers. The maxima of all ion concentrations correspond to multivariate candidate outliers. However, the second maxima, which are also univariate candidates, were not detected using multivariate techniques. For instance, the second maximum of Cl and also that of Na, are not considered multivariate outliers. Recall that the compositional information is in the ratios between the parts and they involve more than a single ion. The detection of outliers can follow two different ways: one based on metadata and sampling characteristics and the statistical detection and treatment. The second way is generally based on the assumed multivariate distribution of data and leads to robust statistics. The first campaign of our data was analyzed using robust statistics in Gozzi et al. (2019) and the interested reader is redirected to this reference. Here, the analysis is a mixture of data analysis and the corresponding metadata.

A preliminary inspection for multivariate outliers was carried out by computing Mahalanobis distance to the center of the data (see e.g. Filzmoser et al., 2018, ch. 5). Table S.1 shows the position of the data points considered candidates to be multivariate outliers, their Mahalanobis distance to the center, and the χ^2 probability assuming a multivariate normal distribution for the coordinates (see Sections 3.3 and 3.4).

In the case of the Tevere basin, 6 outliers were considered. The initial clue to detect them was the form biplot (see Section 3.3) in Fig. S.2 (Supplementary material) where those 6 data points appeared separated from the cloud of the sample. Examining the data set, the mentioned data points correspond to samples at the upper part of the river, placed in sub-basins HT and NE. They are characterized by low ratios of K to other ions. These data points were then diagnosed as potential outliers by using the Mahalanobis distance to the center. They are identified in Table S.1 of the Supplementary material. These data points have been removed from the analysis and are only used in Fig. S.2.

3.3. Form and covariance biplots

In multivariate real data analysis, Principal Component Analysis (PCA) (e.g. Jolliffe (2002)) plays an important role. It allows the determination of orthogonal directions of maximum variation such that

the corresponding coordinates are uncorrelated. These techniques are not directly applicable to CoDa as the variables at play are not real random variables. Their sample space is not the whole real space but a subset of the same. In our case, concentrations in mg/L cannot be negative, the information is in the ratios between them, and the sum of ratios is almost meaningless. Attending to the characteristics of CoDa, Aitchison (1983) applied PCA to the clr thus proposing what is known as CoDa-PCA. It provides three important tools for the analysis of CoDa: principal components which are real orthogonal coordinates for CoDa (ilr/olr, isometric and orthonormal log-ratio coordinates); an analysis of the variability in the sample; and a method to orthogonally project data in a lower dimension (dimension reduction) like in biplots (Gabriel, 1971; Gower and Hand, 1996). The CoDa-biplot takes advantage of the latter to project the ilr coordinates into two (or three) dimensions, jointly with the clr variables.

The procedure to obtain CoDa-PCA and its corresponding biplots is the following. As a first step, the clr of the compositional data is computed and then centered. This is equivalent to a double centering of the logarithm of data. Denote this matrix cclr(X). The second step is the singular value decomposition (svd) of cclr(X). It consists of the matrix decomposition

$$\text{cclr}(X) = U\Lambda V^\top, V^\top V = I_D, UU^\top = I_N, \tag{3}$$

where Λ is a (D,D) diagonal matrix containing the singular values $(\lambda_1, \lambda_2, \ldots, \lambda_{D-1}, \lambda_D)$, the last of them being zero, i.e. $\lambda_D = 0$. The (N,D) matrix U contains the standardized coordinates of the data points, frequently called scores. The (D,D) matrix V, is an orthogonal matrix, as indicated in Eq. (3). The rows of the matrix cclr(X) add to zero since they are clr of the centered data. This causes that $\lambda_D = 0$. The consequence of this is that the three matrices in Eq. 3 can be reduced by one dimension, so that U, Λ and V have dimensions $(N, D-1)$, $(D-1, D-1)$, $(D, D-1)$, respectively, after removing the singular value $\lambda_D = 0$. The notation for these reduced versions is maintained as U, Λ, and V.

The interpretation of Eq. (3) is relevant. The columns of V, frequently called loadings, are the clr of $D-1$ orthogonal compositions which constitute an orthogonal basis of the simplex. They are ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{D-1}$. Consequently, the $(N, D-1)$ matrix $U\Lambda$ are the orthogonal coordinates of the data points (ilr) expressed in the basis defined by V and ordered in decreasing variance. In Section 3.4 these kinds of coordinates are called ilr/olr (isometric/orthogonal log-ratio coordinates). The svd of cclr(X) is a change of representation of the observed compositions, from a clr representation (non-orthogonal) to a new one where the coordinates are orthogonal, i.e. in right angles.

Another important point of Eq. (3) is that it is perturbation invariant, that is U, Λ, V do not change when the compositional observations are perturbed, for instance, by changing the units from mg/L to meq/L, mol/L or to proportions.

After Eq. (3), the orthogonal projection of observations into a reduced dimension space is possible. In fact, retaining only two (or three) columns of V, the observed compositions are projected into a plane (or a three-dimensional space) and the retained coordinates can be plotted accordingly. Also, the unitary vectors in the columns of V, projected as the observations, can be plotted in the same graph. This kind of plot is called *form biplot* (Aitchison and Greenacre, 2002). Fig. 2 shows the form biplot of observed data (outliers removed; see Fig. S.2 including removed outliers). The main indicator of the quality of the projection is the fraction of the total variance retained by the two (or three) components. If only two principal components are retained, the proportion of explained variance (in percent) is

$$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{i=1}^{D-1} \lambda_i^2} \cdot 100 \ .$$

The origin of the red rays represents the data center. The red rays correspond to the compositions of the orthonormal basis (norm equal to 1). If these vectors were perfectly projected on the biplot they would have unitary length, that is the rays would reach the unitary circle (in grey). A short ray points out that the corresponding composition of the basis is poorly captured by the projection which is dominated by the longer rays. In Fig. 2 the ray labeled F (hardly visible) is the worst represented vector and the ray labeled NO_3 is the best represented. The projection explains only 66.2 % of the total variance and, consequently, our observations can be considered as suggestions but not concluding.

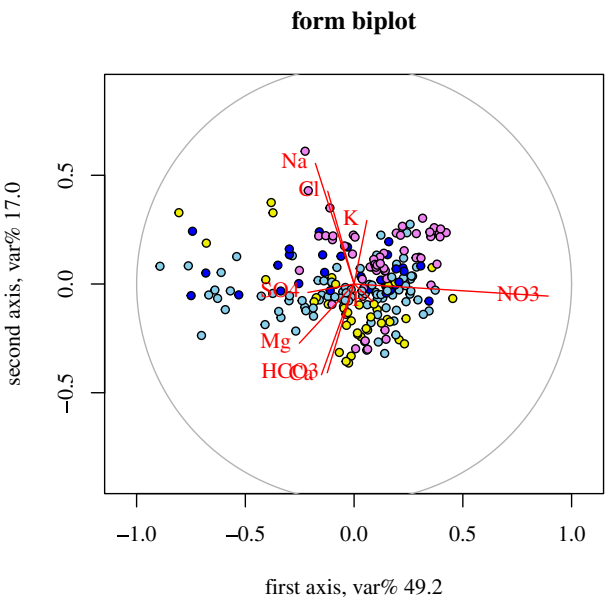

Fig. 2. Form biplot of observed data. Colors indicate the sub-basins, HT sky-blue, MT blue, NE yellow, LT violet. The grey circle is the unit circle. For interpretation of principal axes see Table 5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Moreover, since the points are the orthogonal projections of the observations onto the biplot, this is the best way of looking at the points in two dimensions. Also, the distances between points are the best approaches to the Aitchison distance between points. Although biplots are not a tool for distinguishing clusters or populations, Fig. 2 shows that the sub-basin LT (violet) is shifted to the upper-right of the plot relative to the other sub-basins, thus suggesting that the center of LT can be considered different to that of other sub-basins.

The form-biplot is an orthogonal projection of the data illustrated with the rays indicating the loadings of the elements of the basis. However, when one is interested in the relations between elements, it is preferable to use the alternative normalization of the biplot called *covariance-biplot* (Aitchison and Greenacre, 2002). In this representation, the points are directly the coordinates in the score matrix U and the rays are the projections of the compositions whose clr are the columns $V\Lambda$, that is the vector of the basis elongated by the singular values which are proportional to the standard deviation of the corresponding clr component.

However, the main elements in the CoDa-biplot are the segments that link two vertices of rays, called *links* for simplicity. The length of the links is proportional, up to the projection, to the standard deviation of the log-ratio of the two parts involved. Therefore, again up to the projection, these lengths are proportional to the square roots of the elements in the variation matrix. In Fig. 3, the links between HCO_3-Ca and Cl-Na are the shortest ones, coinciding with the smallest values in the variation matrix (Table 4, upper triangle). As mentioned previously, low values in the variation matrix or large values in the PIP indicate possible proportionality between the parts.

Another important feature in the covariance-biplot is the long links

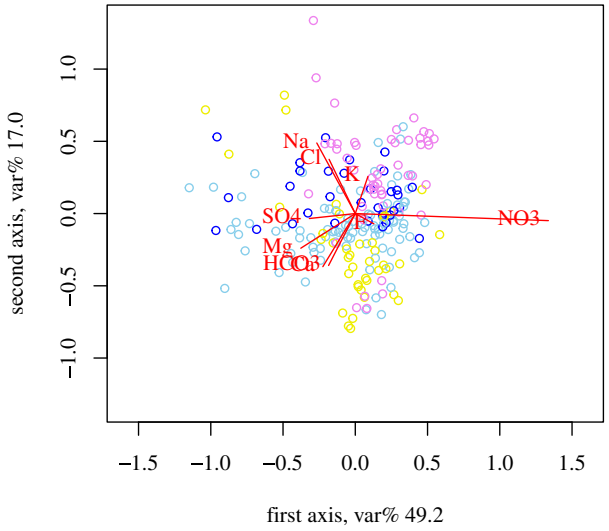

Fig. 3. Covariance biplot of observed data. Colors indicate the sub-basins, HT sky-blue, MT blue, NE yellow, LT violet. The vertices of rays represent the clr-variables. For interpretation of principal axes see Table 5. Attention should be paid to the links between rays of clr-variables in covariance biplots. Points, although colored, are neither filled nor profiled, thus focusing on the rays and links. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

approximating the standard deviation of the corresponding log-ratio, that is, they represent the main sources of variation in the sample. In Fig. 3, attention can be paid to anions like NO_3 and SO_4, whose link is almost parallel to the first principal axis. One can assume that NO_3 is associated with anthropogenic origin whereas SO_4 can also come from geologic sources (e.g., evaporites) (Taussi et al., 2022). Then, this link (and the first principal coordinate) can be interpreted as related to more or less presence of human activities. Also interesting is the link between anions Cl and HCO_3, almost parallel to the second principal coordinate. The relative abundance of anion Cl, quite associated with cation Na, can be attributed to dissolved salt from rocks or more probably to seawater intrusions in the low basins or to rain near the coastal areas. Conversely, HCO_3 is attributable to the alteration of carbonate rocks, more active in the north and east parts of the catchment (HT and especially NE). This is a possible interpretation of the second principal coordinate. The fact that the two links are almost orthogonal can be interpreted as that the process of rock alterations and the process of anthropogenic activities are approximately uncorrelated, meaning that, up to the projection, the log-ratio Cl over HCO_3 is approximately uncorrelated to the log-ratio NO_3 over SO_4.

Many details are difficult to detect in the biplots, especially if the number of parts is large. The study of loadings can elucidate the hidden questions. Table 5 shows the loadings corresponding to the biplot in Fig. 3.

For instance, one can realize that the third principal coordinate is driven by the presence of F and K as opposed to all the others, which is not clear in the two first principal coordinates. This means that the third principal component describes the contribution of volcanic sources along the whole basin. Also, the eighth principal component (PC8), has a low variance, meaning that the log-contrast it represents is approximately constant along the sample. This fact can be due to the characteristics of the sample, but also to some geologic features. For instance, the anion HCO_3 is contrasted with the cation Ca, thus suggesting the stoichiometric equilibrium of the dissolution of minerals contained in carbonate rocks.

Table 5
Loadings corresponding to the CoDa-PCA in Figs. 2 and 3. The last row contains the percent of total variance retained by each principal coordinate. Vectors along axes are linear combinations of logs of parts whose coefficients are the loadings in the respective column. Cells with loadings larger than 0.40 in absolute value are colored. The two first columns (PC1,PC2) define the axes in Figs. 2 and 3.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
HCO_3	-0.12	-0.41	0.05	0.39	-0.14	0.16	-0.11	-0.70
F	0.09	-0.08	-0.67	-0.25	-0.39	-0.39	-0.23	0.02
Cl	-0.12	0.43	0.24	0.00	-0.46	-0.16	0.63	-0.09
NO_3	0.89	-0.06	0.28	-0.04	0.07	-0.02	-0.02	0.02
SO_4	-0.21	-0.04	0.12	-0.80	0.25	0.30	-0.01	-0.19
Ca	-0.15	-0.42	0.09	0.12	-0.29	0.41	0.03	0.64
Mg	-0.25	-0.27	0.19	0.11	0.49	-0.65	0.11	0.17
Na	-0.18	0.55	0.24	0.18	-0.01	0.02	-0.67	0.10
K	0.06	0.29	-0.55	0.29	0.48	0.33	0.28	0.04
% var	49.2	17.0	13.4	11.0	5.6	2.4	1.0	0.3

3.4. Coordinates and CoDa dendrogram

After Pawlowsky-Glahn and Egozcue (2001) and Billheimer et al. (2001) the simplex was structured as a Euclidean space. Therefore, D-part compositions can be represented using Cartesian orthogonal coordinates with respect to an orthonormal basis made of compositions $\mathbf{e}_1$, $\mathbf{e}_2$, …, $\mathbf{e}_{D-1}$, such that $\| \mathbf{e}_i \|_a = 1$ and $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = 0$ when $i \neq j$. The Aitchison norm, $\| \cdot \|_a$, and inner product $\langle \cdot, \cdot \rangle_a$ are easily computable from the clr of the compositions. If $\mathbf{x}$ and $\mathbf{y}$ are D-part compositions, then

$$\| \mathbf{x} \|_a = \| \operatorname{clr}(\mathbf{x}) \|_e, \langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{y}) \rangle_e,$$

where the subscripts e mean ordinary Euclidean functions. Once again the clr representation is a powerful tool to operate within the Aitchison geometry. In this framework any D-part composition $\mathbf{x}$ can be expressed as a linear combination of the elements of the basis

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i, x_i^* = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \langle \operatorname{clr}(\mathbf{x}), \operatorname{clr}(\mathbf{e}_i) \rangle_e.$$

The coefficients x_i^* are the ilr coordinates of $\mathbf{x}$ with respect to the element of the basis $\mathbf{e}_i$, $i = 1, 2, \ldots, D-1$.

In summary, the ilr coordinates grouped in $\mathbf{x}^*$ can be computed as

$$\mathbf{x}^* = V^T log(\mathbf{x}), V^T V = I_{D-1}, \quad (4)$$

where V is a $(D, D-1)$-matrix whose rows are $\operatorname{clr}(\mathbf{e}_i)$ and, therefore, adding to 0.

The previous Section 3.3, shows the usefulness of representing CoDa in coordinates, particularly, orthogonal principal coordinates. However, principal coordinates are data-driven and not designed for easy solving of research questions. The ilr coordinates (Egozcue et al. (2003), also known as olr after Martín-Fernández (2019)) can be designed by the user according to his/her needs. The procedure to this end is called *Sequential Binary Partition* (SBP) (Egozcue and Pawlowsky-Glahn, 2005, 2006). It consists of a partition of the composition into two groups of parts, and then each group of the first partition is split into two groups. This procedure is iterated until all groups contain only one part, which is attained after $D-1$ partitions. The selection of the groups in each partition is arbitrary. The k-th partition is associated with an ilr

coordinate whose expression is

$$b_k = B(\mathbf{x}_{+k}/\mathbf{x}_{-k}) = \sqrt{\frac{n_{+k}n_{-k}}{n_{+k}+n_{-k}}}\ \ln\frac{g_m(\mathbf{x}_{+k})}{g_m(\mathbf{x}_{-k})}, k = 1, 2, \ldots, D-1, \quad (5)$$

where $\mathbf{x}_{+k}$, $\mathbf{x}_{-k}$ are the groups of parts marked with +1 and -1 respectively in the k-th partition, i.e. those assigned to the first and second subgroup in this step (see Table 6), and n_{+k}, n_{-k} are the number of parts in those groups; $g_m(\cdot)$ denotes the geometric mean of the argument. All b_k in Eq. (5) are balances (Egozcue and Pawlowsky-Glahn, 2005), that is a normalized log-ratio of geometric means of two non-overlapping groups of parts. The value of the square root is the normalizing constant so that it corresponds to a unitary vector of the basis of the simplex, thus making the magnitudes of the balances comparable.

The SBP shown in Table 6 could have been obtained from the subjective criterion of the analyst. The first partition step separates F and K from the rest of the ions since it is assumed that the main source of these elements is volcanic rocks. The third partition step, labeled pb_1, separates NO_3 the only element associated with anthropic activities. The next step, labeled pb_4, separates SO_4 from the remaining elements but there is no clear reason for that except to facilitate interpretation of the next steps. These steps try to group Cl-Na and HCO_3 with Ca-Mg. The selection of signs in a partition step is irrelevant, it is only associated with a change of orientation of an axis.

However, the SBP in Table 6 was obtained using the *principal balances* (PB) technique (Martín-Fernández et al., 2018). It tries to approximate the principal coordinates (CoDa-PCA) by balances, like in Eq. (5). The first PB (labeled in Table 6 as pb_1) has the largest variance within the set of PBs. Next PBs, being orthogonal to the first one, have maximum variance, and so on up to the $(D-1)$-th PB. Labels in Table 6 indicate the decreasing order of variance of those PBs.

A way of visualizing the SBP used to obtain coordinates is the *CoDa-dendrogram*. It represents the SBP as a dendrogram showing the partitions as in Fig. 4 (look at the tree in black).

Taking advantage of the dendrogram, the sample mean and variance of the balance coordinates can be represented on it. Each horizontal bar is assumed to have an equal length, in Fig. 4 they are scaled to the segment $(-6, 6)$. Vertical lines are anchored in them at the value of the sample mean of the corresponding balance-coordinate. In black, the balance corresponds to the overall sample, and, in different colors, the balances are those of the different sub-basins thus comparing the mean value of the balance coordinate in sub-basins. Note that the mid-point of the bar is the zero of the coordinate balance, and the more right anchoring, the more positive is the mean value of the balance. The length of vertical bars corresponds to the sample variance of the corresponding coordinate balance in the overall sample (black) and colors for the sub-basins. Since

Table 6
Each column represents a partition into two groups of parts, one group is labeled as +1 and the other one with -1; parts that do not participate in the partition are labeled 0. The rows n_+ and n_- are the number of parts labeled with 1 and -1 respectively in each partition. In the row labeling the sequence of partitions *pb* is *principal balance* and the number is the order in decreasing variance.

	pb_3	pb_5	pb_1	pb_4	pb_2	pb_7	pb_6	pb_8
HCO_3	+1	0	−1	+1	+1	0	−1	+1
F	−1	+1	0	0	0	0	0	0
Cl	+1	0	−1	+1	−1	+1	0	0
NO_3	+1	0	+1	0	0	0	0	0
SO_4	+1	0	−1	−1	0	0	0	0
Ca	+1	0	−1	+1	+1	0	−1	−1
Mg	+1	0	−1	+1	+1	0	+1	0
Na	+1	0	−1	+1	−1	−1	0	0
K	−1	−1	0	0	0	0	0	0
n_+	7	1	1	5	3	1	1	1
n_-	2	1	6	1	2	1	2	1

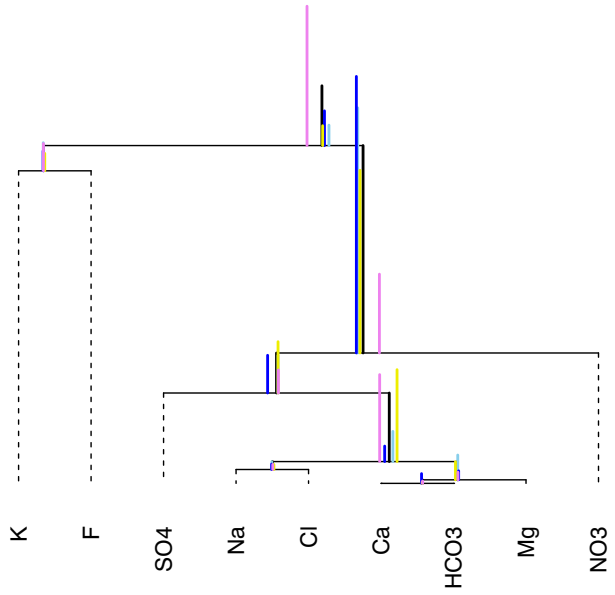

Fig. 4. CoDa-dendrogram for the Tevere basin (black tree) corresponding to the SBP coded in Table 6. HT sub-basin in sky-blue; MT in blue; NE in yellow and LT in violet. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\mathrm{TotVar}(X) = \sum_{j=1}^{D-1} \mathrm{Var}\left(\mathrm{ilr}_j(X)\right), \quad (6)$$

is a decomposition of the total variance into orthogonal components, the lengths of vertical bars in the CoDa-dendrogram represent such a decomposition of the total variance for the overall basin (black) and each of the sub-basins (colour).

Fig. 4 immediately reveals that the balances $B(\mathrm{Cl}/\mathrm{Na})$ and $B(\mathrm{HCO_3}/\mathrm{Ca})$ have low variance (short vertical bars) suggesting an association between involved elements (see also Table 4). Balance $B(\mathrm{NO_3}/\mathrm{HCO_3, Cl, SO_4, Ca, Mg, Na})$ (labeled pb_1 in Table 6) is identified as having the largest sample variance. Again, the coherent behavior of Cl, Na and Ca, Mg is related to a geochemical affinity in natural processes and the source. On the other hand, the balance $B(\mathrm{NO_3}/\mathrm{HCO_3, Cl, SO_4, Ca, Mg, Na})$ appears to link different processes, natural and anthropic, often working on different scales, thus explaining the higher variability (mixing of processes and sources).

An interesting feature of the CoDa-dendrogram is the visualization of differences between centers of sub-basins taking into account the respective variabilities. For instance, the mentioned $B(\mathrm{NO_3}/\mathrm{HCO_3, Cl, SO_4, Ca, Mg, Na})$ differentiates the center of LT in front of other sub-basins whose values are hardly different. This confirms the larger relative abundance of NO_3 in sub-basin LT. Balance $B(\mathrm{F, K/all\ other\ parts})$, labeled pb_3, presents a similar situation, a larger relative abundance of F and K in LT compared to other sub-basins. The only balance that seems to distinguish (in mean) the four sub-basins is $B(\mathrm{Cl, Na/HCO_3, Ca, Mg})$ (pb_2). However, this observation should be tinged or modulated by observing the variances in sub-basins. These variances in LT and NE are large and distinguishing them from other sub-basins can be non-significant. Contrarily, the variance in sub-basins HT and MT are smaller and the mean balance can be considered different. An analysis of variance reveals that the sample means of this balance in HT and MT cannot be considered equal (p-value 2.05e-05). See also boxplots comparing the four sub-basins in the Supplementary materials Fig. S.3.

4. Are the mean compositions of sub-basins equal?

The comparison of the four sub-basins has been stated as a research question. This comparison can be studied from different perspectives. The first way can be the comparison of the compositional centers, reported in Table 2, and of the Aitchison distances between them, which are listed in Table 3. Although these Tables suggest some differences between centers, a statistical contrast is necessary. The problem can be translated into a standard one in real multivariate statistics making use of the principle of working in coordinates (Mateu-Figueras et al., 2011), that is, translate the centers into ilr-coordinates, for instance, those used in the CoDa-dendrogram (Fig. 4) and, then, proceed as in standard multivariate statistics. The first step is to perform a MANOVA (Sierra et al., 2017). It confirms that there are significant differences between the centers (Pillai statistics p-value $< 10^{-7}$), but this only confirms what was clear from the exploratory analysis. The corresponding post-hoc tests only check which ilr-coordinates are different for each sub-basin. This was shown graphically in Fig. 4. The main inconvenience is that this depends on the selection of the ilr-coordinates. What is needed is to know which are the characteristics of compositions that differentiate sub-basins. These characteristics should be simple expressions that admit simple interpretations. These functions should be scale invariant and hence log-contrasts are candidates. Preferably, they should be sparse, that is, involving only a few elements. The simplest expressions fulfilling these conditions are the pairwise log-ratios or sparse balances. In this way, the goal is to identify balances involving few elements, able to discriminate between two sub-basins. This can be achieved using procedures like *selbal* (Rivera-Pinto et al., 2018).

Table 7 shows the balance that best discriminates the first sub-basin from the second one (balance larger than or equal to the threshold). The accuracy of the classifier is quantified by ROC-AUC (Receiver Operating Curve-Area Under Curve) (Fawcett, 2006). The ROC-AUC takes values between 0 and 1. A large value, close to 1, indicates a nearly perfect separation of the two groups.

The main advantage of the discrimination by a balance is that it allows an easy interpretation of differences between sub-basins. In fact, the species involved in the balances highlight the lithological differences of the four sub-basins. In particular, the HT sub-basin seems to be well discriminated from MT and NE by the relative increase in Na compared to the other species considered in the balances ($B(\mathrm{Na, Ca/HCO_3})$, $B(\mathrm{Cl/Na})$). This could be linked to the greater contribution in Na from the upper part of the catchment due to silicate weathering reactions. LT is distinguished by elements that typically derive from water-rock interaction processes with potassic and ultrapotassic volcanic complexes, whereas those related to the weathering of carbonate rocks are characteristic of NE.

5. Is there any relation between height and ion composition? – Regression

This question can be answered from different points of view but regression is the more intuitive one. Even so, one can think of predicting the height of the sample from the composition of ions, or vice-versa taking the composition as a response and trying to predict it from the height. For simplicity, we have only considered the height, but other covariates could be included in the analysis, for instance, pH and/or conductivity.

Table 7
Discriminating balance between two sub-basins. The threshold is the value of the balance that best separates the two sub-basins. ROC-AUC is a measure of accuracy in the prediction of the first sub-basin by the value of the balance over the threshold.

Sub-basins		Balance	Threshold	ROC-AUC
HT	MT	$B(\mathrm{Na, Ca/HCO_3})$	−1.388	0.855
HT	NE	$B(\mathrm{Cl/Na})$	0.262	0.835
HT	LT	$B(\mathrm{F, Cl/HCO_3})$	−3.819	0.882
MT	NE	$B(\mathrm{HCO_3, Cl/Na, K})$	2.377	0.923
MT	LT	$B(\mathrm{SO_4/Cl})$	0.678	0.859
NE	LT	$B(\mathrm{HCO_3/F, Na, K, Mg})$	4.021	0.873

5.1. Height as a response in regression

Taking height, h, as a response from the composition can be tackled by a simple multiple regression. In this case, the principal balance-coordinates computed following the SBP in Table 6 can be taken as covariates. There are other alternatives, for instance, the principal coordinates obtained in the CoDa-PCA whose loadings are shown in Table 5. In any case, attention is restricted to linear models like

$$h = \beta_0 + \beta_1 \left(\sum_{i=1}^{D} \alpha_i log X_i \right) = \beta_0 + \beta_1 \phi(\mathbf{x}), \sum_{i=1}^{D} \alpha_i = 0 \ , \tag{7}$$

where the condition on the α_is assures that the predictor is scale invariant and the D parts X_i represent the observed concentration of ions dissolved in water (Aitchison and Bacon-Shone, 1999). The predictor is a log-contrast of the composition, a scale-invariant linear combination of the logs of the parts. When the regression is based on coordinates, such as ilr or principal coordinates, the α_i coefficients are obtained by combining the coefficients of the coordinates $\mathbf{x}^*$ (Eq. 4) times the estimated regression coefficients.

Table 8 shows the results for different approaches. The row *general* is obtained by regressing h on all principal balances coded in Table 6. The $R^2 = 0.59$ is not a high one, thus indicating that the model shows a trend more than a prediction of h. A regression using the CoDa-principal components gives exactly the same result. This is due to the invariance of regression models under rotation of orthogonal axes. However, this general model lacks easy interpretability as all ions are involved with different coefficients. Significance tests allow the removal of some principal balances (pb_3, pb_5, pb_7, pb_8), but this does not improve interpretability. An important simplification is obtained by forcing sparsity, making some of the coefficients α_i to be equal to zero. There are procedures that force this simplification, see for instance Shi et al. (2016) using the Lasso techniques (Tibshirani, 1996). In particular, using the method called elastic-net proposed in Susin et al. (2020), the result in the row *elastic net* is obtained. Note that, at a very low cost in R^2, two zeros are introduced in the regression model (for *log*Mg and *log*K). This simplifies the model but still seven ions, with different weights, participate in the predictor. A further simplification consists of making the $\phi(\mathbf{x})$ in Eq. (7) to be a balance (Eq. 5), that is, the coefficients α_i attain only three different values: 0, a negative value for those elements in the denominator of the balance and another positive for the elements in the numerator. Based on expert knowledge, the following balance including only anions was selected

$$\phi(\mathbf{x}) = B(\mathrm{HCO_3/Cl\ NO_3\ SO_4}) = \sqrt{\frac{3}{4}}\ log\frac{\mathrm{HCO_3}}{(\mathrm{Cl\ NO_3\ SO_4})^{1/3}} \ ,$$

and the result is in the row *user*. There is a new reduction of R^2 to 0.54 but this is a price compensated by a clear increase of interpretability. The data, classified by sub-basins, and the regression line are shown in Fig. 5, where it is clear that the regression model only points out a trend with the height.

There are also some procedures to select a balance automatically. Here the *selbal* method (Rivera-Pinto et al., 2018) was used. The balance selected is in this case $B(\mathrm{HCO_3/Cl})$, which increases interpretability to a maximum but R^2 is now reduced to 0.34.

5.2. Composition as a response in regression

Predicting an ion composition of $D = 9$ parts on the height of the sample is a hopeless task unless there are strong associations between

Table 8
Estimated α_i $i = 1, 2, \ldots, D$ coefficients in Eq. (7). Column R^2 is the determination coefficient in the regression. Rows correspond to different degrees of simplification of the predictor. See the text for an explanation.

Log-contrast	HCO_3	F	Cl	NO_3	SO_4	Ca	Mg	Na	K	R^2
General	63.715	−33.499	−17.679	−43.529	−58.830	127.271	6.936	−56.393	12.009	0.59
Elastic net	93.996	−18.919	−44.746	−43.364	−50.563	97.624	0	−34.028	0	0.59
User	54.168	0	−18.056	−18.056	−18.056	0	0	0	0	0.54
Selbal balance	123.88	0	−123.88	0	0	0	0	0	0	0.34

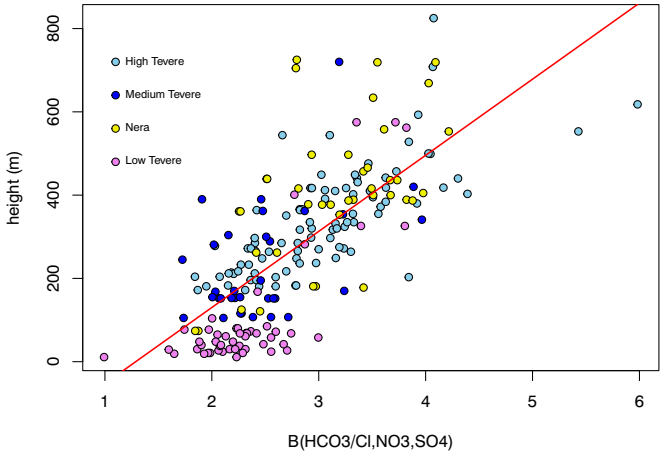

Fig. 5. Scatterplot of height of sampling points with the balance $B(\mathrm{HCO_3/Cl, NO_3\ SO_4})$. Points are colored according to the sub-basin in which they were collected. The regression line (red) is also shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the elements, that in this case were not detected (see Table 4). The more flexible and easy-to-understand procedure consists of expressing the composition in orthogonal coordinates and then regressing these coordinates on the covariates, in this case only h (Egozcue et al., 2012). The orthogonality of the coordinates allows us to reduce the problem to fit $D-1$ simple regressions which can be fitted independently. Although this is an advantage, this independent fitting makes the model dependent on the particular ilr-coordinates chosen.

In our case, we can represent the observed composition of ions by the principal balance coordinates defined by the SBP coded in Table 6. These coordinates are balances, and they are regressed only on height h, although other covariables could be added. The least squares regression models are

$$pb_i = \beta_{i0} + \beta_{i1} h + \varepsilon_i, i = 1, 2, \ldots, D-1, \tag{8}$$

where ε_i are residuals whose sum of squares is to be minimized. As expected, most of these regressions are not significant and all of them have small R^2. The best fit corresponds to $pb_3 = B(\mathrm{F, K/all\ other})$ ($R^2 = 0.3266$, p-value $< 10^{-5}$, $\widehat{\beta}_{31} = 0.0027$, $\widehat{\beta}_{30} = 0.8137$). Although, the regressions in Eq. (8) do not contribute much to the knowledge about the relation between h and the observed composition, the regression of pb_3 still provides some interpretation. R^2 is the fraction of variance of pb_3 explained by h. Also, the fact that $\beta_{31} < 0$ is rejected to be null, suggests that $pb_3 = B(\mathrm{F, K/all\ other})$ is influenced positively by height h, that is, there is some weak evidence that relative to other elements, the geometric mean of F and K increases with h. That can be interpreted as that the volcanic contributions vaguely increase with h.

However, any function of the composition can be regressed on covariates like h or the pH. For instance, any balance, appearing in an SBP can be taken as a response to a regression. The only condition is that results are interpretable and therefore sensible for the research.

6. How do samples cluster according to their ion composition?

Unsupervised cluster analysis (Kozak and Scaman, 2008) can give several different results depending on the clustering method and the preprocessing of the data set. Consequently, the question in the title has multiple answers. First of all, not all the information in the data set provides the same information for clustering, in our case it depends on the subcomposition chosen for clustering samples. The whole available composition of ions is here taken as a reference but, if the subcomposition $(\mathrm{F, K})$ is alternatively considered, one can expect substantially different results.

In order to answer the posed question, it seems appropriate to provide some preliminary information, for instance, the compositional centers of the data in each sub-basin (Table 2). This suggests the use of a k-means method (Emre Celebi et al., 2013) to obtain a clustering according to the ion composition. To carry out the cluster analysis using compositional data some previous steps are required since most available software requires real data and a distance or similarity for real data. Therefore, the first step should be expressing the compositional data as real data. This can be achieved in several ways, for instance, represent the data in ilr coordinates like those obtained as principal balances (contrast matrix coded in Table 6), or as clr, which allows the computation of the Aitchison distance as a Euclidean distance. In the case of k-means cluster analysis, the initial centers of the four clusters should also be provided in the chosen representation of the data, that is in ilr coordinates or in clr representation. Then, one can proceed to any cluster analysis that uses Euclidean distances.

Fig. 6 shows different clustering results of the data set. In the right panel, the sampling points are colored as the sub-basins in Fig. 1 (sky-blue for HT, blue for MT, yellow for NE, and violet for LT). Additionally, samples taken at elevations higher than 500 m are marked with a red square. In the middle panel, the same points are colored as the clusters obtained using k-means. The colors are assigned as the initial centroids proposed. The right panel shows the hierarchical cluster obtained for 4 groups, colored as the previous similar groups.

Comparing the left and middle panels in Fig. 6, the first observation is that the sub-basin MT (medium Tevere) is blurred and merged with the other groups. That is, the geographic characterization of MT does not correspond to the ion composition. This is in accordance with the fact that MT is highly geologically heterogeneous compared to the other sub-basins, which instead are marked by well-defined lithological differences. A second observation is that violet points, originally assigned to LT (lower Tevere) and associated with a larger anthropogenic influence, extend to other sub-basins that also have such influence. Finally, the changes of group corresponding to samples over 500 m show an erratic behavior, probably due to pristine waters mainly influenced by the rock composition of the bed of the current or to the effects of local springs.

More difficult is the comparison of the left panel and the right one in Fig. 6. Although the shape of sub-basin cluster LT is still identifiable, other groups are mixed, thus supporting the idea that cluster analysis can result differently depending on methods and procedures.

7. There are no zeros in this data set. What can be done if zeros are present?

In fact, the data set of ions dissolved in the Tevere basin waters does

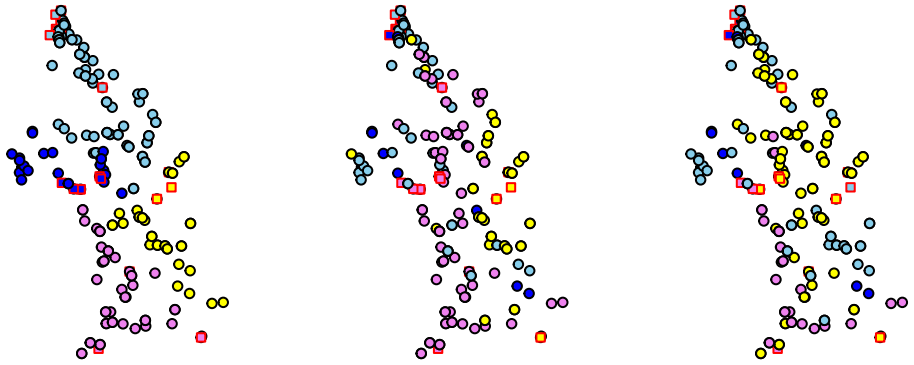

Fig. 6. Left: sample points classified by sub-basins (see Fig. 1). Middle: sample points clustered using K-means; initial centroids equal to centers in the four sub-basins (see Table 2). Right: sample points classified by hierarchical cluster (four groups), using the Aitchison distance and Ward's method. Points with red square markers have heights higher than 500 m. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

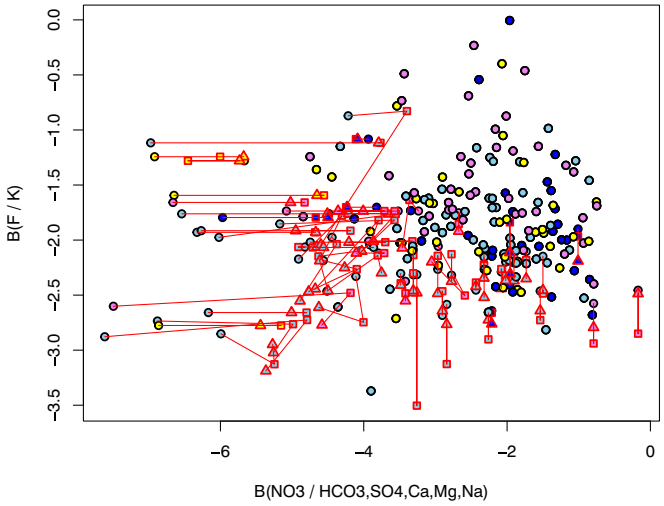

Fig. 7. Comparison of zero substitutions. Original data set, black circles. Random Log-Normal substitution, red squares. Log-ratio EM substitution, red triangles. Red lines join the original data point, Random Log-Normal, and log-ratio EM substitutions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

not contain any zero. This is not the most frequent situation, especially if trace elements are included in the reported sample. Zeros in compositional samples are so frequent that some researchers focus their research in order to include zeros as regular compositional data (e.g Butler and Glasbey, 2008; Tsagris, 2021). However, the log-ratio approach to CoDa does not accept zeros as compositional parts and considers them as irregular data. The reason is the definition of CoDa as positive data with relative scale, better than the old definition of vectors adding to a constant. Certainly, zero is not relative to anything.

In geochemistry most occurrences of zeros are due to values under the detection limit of measurement devices, thus they can be considered as censored data when the detection limit is known. In order to demonstrate the case of data under detection limit and their possible treatment, artificial zeros are created in the reference data set. This was done assuming that F has detection limit $dl(\mathrm{F}) = 0.0999$ and $dl(\mathrm{NO_3}) = 0.57$ (mg/L). This generates 42 artificial zeros in F and 37 artificial zeros in $\mathrm{NO_3}$. See the artificial zero pattern in Supplementary material, Fig. S.5.

The ideal situation to afford the treatment of zeros is to know the sampling distribution which allows to establish the likelihood of unknown parameters given the data (including the zeros). Since this is not the case, a substitution of zero data should be used. Depending on the hypothesis, the substitution is carried out using different methods. Two of them were selected (Palarea-Albaladejo and Martín-Fernández, 2013; Palarea-Albaladejo and Martín-Fernández, 2015). The first one assumes that the sampling probability distribution of the concentration of the element containing a zero is a log-normal distribution (LN). Then a value under *dl* is simulated according to LN. The second method assumes that the compositional sample is normal in the simplex (Mateu-Figueras et al., 2013; Pawlowsky-Glahn et al., 2015). The normal model of the ilr coordinates allows prediction of the values under *dl* using the EM-method (Dempster et al., 1977); the resulting algorithm is named log-ratio EM or lr-EM (Palarea-Albaladejo and Martín-Fernández, 2015).

Both procedures, LN and lr-EM, were applied to substitute the artificial zeros. Fig. 7 shows a projection on the plane of the two principal balances $pb_1 = B(\mathrm{NO_3/HCO_3, SO_4, Ca, Mg, Na})$ and $pb_5 = B(\mathrm{Mg/HCO_3, Ca})$ of the original data (circled points colored by sub-basins). Also, substituted zeros are shown as red squares for the LN procedure and red triangles for lr-EM. A red line joins the original data that was artificially made zero with the LN substitution and the lr-EM substitution (see also another projection in the Supplementary material Fig. S.6).

The sampling points that were transformed into artificial zeros in the left-hand side of Fig. 7 are shifted right after the substitution of zeros, both using LN or lr-EM procedures. These points correspond to zeros created in $\mathrm{NO_3}$. Also, most artificial zero points placed at the lower part of Fig. 7 (corresponding to zeros in F) are shifted down after replacing zeros. The most extreme shifts are due to the fact that the assumptions for the substitution are not fulfilled. For instance, the LN substitution is based on the log-normal distribution of the elements. However, these hypotheses are rejected using the original data (Supplementary materials Fig. S.7). However, note that, in a standard situation, these hypotheses are hardly checked as, obviously, the true values are not known.

8. Software

Computations and figures were elaborated using R (R Development Core Team, 2004). Particularly, the following list of tasks use R-packages:

- Zero pattern and substitutions: R-package *zCompositions* (functions: zPatterns, multLN, lrEM) (Palarea-Albaladejo and Martín-Fernández, 2015).
- CoDa-dendrogram, ilr coordinates, variation array and Mahalanobis distances: R-package *compositions* (functions: CoDaDendrogram, MahalanobisDist) (Boogaart et al., 2009; Boogaart and Tolosana-Delgado, 2008).
- Principal balances: R-package *coda.base* by Comas-Cufí, (function: pb.basis), URL: https://CRAN.R-project.org/package=coda.base .

- Sparse regression: R-packages *coda4microbiome* and *selbal* (functions: coda_glmnet, selbal.cv) (Susin et al., 2020; Rivera-Pinto et al., 2018).
- CoDa-principal components and biplot: it was computed using the function svd (singular value decomposition) of the R-package *base*. However, the computation and plots can be carried out using the cited package *compositions* or *robCompositions* (Templ et al., 2011).

9. Conclusions

Many systems on the Earth shift abruptly from one given state to another when forced across a tipping point. Mass extinctions in ecosystems represent a typical example, as well as the change in the hydrological cycle due to global warming that, with cascade effects, modify forest cover, land use, and climate. Predicting and possibly avoiding regime shifts depends on our capacity to analyze data, taking into account the relationships among different components. From this point of view, the analysis of single variables is not particularly informative in itself, if the complexity and non-linearity of the processes have to be taken into account. Moreover, if high-dimensional systems are investigated, and compositional data are part of the framework, the choice of the adequate sample space is fundamental to obtain relevant evaluations and predictions. In the realm of geochemistry, this holds particularly true, as the relationships among chemical constituents are typically analyzed by traditional binary or ternary diagrams, without consideration of the appropriate sample space. This work represents an updated guide for the proper use of consistent statistical methods applied to compositional data. The step-by-step application obtained for Tevere (Tiber) river water chemistry, following the concept of *answering to research questions*, has allowed highlighting interesting features of the behavior of chemical species, as well as of the sub-basins dominated by different lithology. With the development of Artificial Intelligence data are more and more considered as the "natural capital" and, consequently, methods of data analysis should be up to the challenge. In this context, innovation in CoDA would represent an important turning point.

CRediT authorship contribution statement

Juan José Egozcue: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Caterina Gozzi:** Investigation, Visualization, Writing – original draft, Writing – review & editing. **Antonella Buccianti:** Conceptualization, Resources, Validation, Writing – review & editing. **Vera Pawlowsky-Glahn:** Conceptualization, Methodology, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Co-Guest Editor of the Special Issue "VSI:CoDA 40 years after 1982" of Journal of Geochemical Exploration - C.G.

Data availability

Data will be made available on request.

Acknowledgements

JJE and VPG were supported by the Ministerio de Ciencia e Innovación, Spain, under the projects "CODA-GENERA" (Ref. PID2021-123833OB-I00) and "CONBACO" (Ref. PID2021-125380OB-I00). AB and CG acknowledge the support of the National Biodiversity Future Center (NBFC) and National Centre for HPC, Big Data and Quantum Computing to the University of Florence, Department of Earth Sciences, funded by the Italian Ministry of University and Research, PNRR, Missione 4 Componente 2, "Dalla ricerca all'impresa", Investimento 1.4, Projects CN00000033 and CN00000013.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2024.107385.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B (Stat Methodol.) 44, 139–177.
Aitchison, J., 1983. Principal component analysis of compositional data. Biometrika 70, 57–65. https://doi.org/10.1093/biomet/70.1.57.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK) ((Reprinted in 2003 with additional material by The Blackburn Press), London (UK). 416 p).
Aitchison, J., 1992. On criteria for measures of compositional difference. Math. Geol. 24, 365–379.
Aitchison, J., Bacon-Shone, J., 1999. Convex linear combination of compositions. Biometrika 86, 351–364.
Aitchison, J., Greenacre, M., 2002. Biplots for compositional data. J. R. Stat. Soc.: Ser. C: Appl. Stat. 51, 375–392.
Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Math. Geol. 32, 271–275.
Barceló-Vidal, C., Martín-Fernández, J.A., 2016. The mathematics of compositional analysis. Austrian J. Stat. 45, 57–71. https://doi.org/10.17713/ajs.v45i4.142.
Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96, 1205–1214.
Boogaart, K.G.v.d., Tolosana-Delgado, R., 2008. "Compositions": a unified R package to analyze compositional data. Comput. Geosci. 34, 320–338.
Boogaart, K.G.v.d., Tolosana-Delgado, R., 2013. Analysing Compositional Data with R. Springer, Berlin, p. 258.
Boogaart, K.G.v.d., Tolosana, R., Bren, M., 2009. Compositions: Compositional Data Analysis. URL. http://www.stat.boogaart.de/compositions (r package version 1.02-1).
Butler, A., Glasbey, C., 2008. A latent Gaussian model for compositional data with zeros. J. R. Stat. Soc.: Ser. C: Appl. Stat. 57, 505–520.
Chayes, F., 1971. Ratio Correlation. University of Chicago Press, Chicago, IL (USA), p. 99.
Dempster, A.P., Laird, N.M., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828.
Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In: Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, pp. 145–159.
Egozcue, J.J., Pawlowsky-Glahn, V., 2019. Compositional data: the sample space and its structure. TEST 28, 599–638. https://doi.org/10.1007/s11749-019-00670-6.
Egozcue, J.J., Pawlowsky-Glahn, V., 2023. Subcompositional coherence and a novel proportionality index of parts. SORT 47, 229–244. https://doi.org/10.57645/20.8080.02.7.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.
Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K., Filzmoser, P., 2012. Simplicial regression. The Normal model. J. Appl. Prob. Stat. 6, 87–108.
Emre Celebi, M., Kingravi, H.A., Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. 40 (1), 200–210.
Fawcett, T., 2006. An introduction to roc analysis. Pattern Recogn. Lett. 27, 861–874.
Filzmoser, P., Hron, K., Templ, M., 2012. Discriminant analysis for compositional data and robust parameter estimation. Comput. Stat. 27, 585–604.
Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Analysis. With Worked Examples in R. Springer Nature, Switzerland. https://doi.org/10.1007/978-3-319-96422-5 (280pp).
Gabriel, K.R., 1971. The biplot – graphic display of matrices with application to principal component analysis. Biometrika 58, 453–467.
Gower, J.C., Hand, D.J., 1996. Biplots. Chapman and Hall Ltd., London (UK), p. 277.
Gozzi, C., Buccianti, A., 2022. Assessing indices tracking changes in river geochemistry and implications for monitoring. Nat. Resour. Res. 31 (2) https://doi.org/10.1007/s11053-022-10014-1.
Gozzi, C., Filzmoser, P., Buccianti, A., Vaselli, O., Nisi, B., 2019. Statistical methods for the geochemical characterisation of surface waters: the case study of the Tiber river basin (Central Italy). Comput. Geosci. 131, 80–88. https://doi.org/10.1016/j.cageo.2019.06.011.
Gozzi, C., Sauro Graziano, R., Buccianti, A., 2020. Part–whole relations: new insights about the dynamics of complex geochemical riverine systems. Minerals 10. https://doi.org/10.3390/min10060501.
Gozzi, C., Dakos, V., Buccianti, A., Vaselli, O., 2021. Are geochemical regime shifts identifiable in river waters? Exploring the compositional dynamics of the Tiber River (Italy). Sci. Total Environ. 785, 147268 https://doi.org/10.1016/j.scitotenv.2021.147268.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer Series in Statistics, 2nd ed. Springer-Verlag, New York.
Kozak, M., Scaman, C.H., 2008. Unsupervised classification methods in food sciences: discussion and outlook. J. Sci. Food Agric. 88, 1115–1127.
Martín-Fernández, J.A., 2019. Comments on: compositional data: the sample space and its structure, by egozcue and pawlowsky-glahn. TEST 28, 653–657. https://doi.org/10.1007/s11749-019-00670-4.
Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2018. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298. https://doi.org/10.1007/s11004-017-9712-z.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The principle of working on coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications, 378. John Wiley & Sons, pp. 31–42.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2013. The normal distribution in some constrained sample spaces. Stat. Operat. Res. Trans. 37, 29–56.
Palarea-Albaladejo, J., Martín-Fernández, J.A., 2013. Values below detection limit in compositional chemical data. Anal. Chim. Acta 764, 32–43.
Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. zCompositions — r package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. 143, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019.
Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stochastic Environ. Res. Risk Assess. 15, 384–398.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Statistics in practice,. John Wiley & Sons, Chichester UK, p. 272.
R Development Core Team, 2004. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria (URL: http://www.r-project.org., iSBN 3-900051-00-3).
Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L., 2018. Balances: a new perspective for microbiome analysis. mSystems 3 (URL: https://api.semanticscholar.org/CorpusID:51706015).
Shi, P., Zhang, A., Li, H., 2016. Regression analysis for microbiome compositional data. Ann. Appl. Stat. 10, 1019–1040. https://doi.org/10.1214/16-AOAS928.
Sierra, C., Ruiz-Barzola, O., Menendez, M., Demey, J.R., Vicente-Villardon, J.L., 2017. Geochemical interactions study in surface river sediments at an artisanal mining area by means of canonical (manova)-biplot. J. Geochem. Explor. 175, 72–82.
Susin, A., Wang, Y., Le Cao, K.A., Calle, M.L., 2020. Variable selection in microbiome compositional data analysis. NAR Genom. Bioinf. 2 https://doi.org/10.1093/nargab/lqaa029.
Taussi, M., Gozzi, C., Vaselli, O., Cabassi, J., Menichini, M., Doveri, M., Romei, M., Ferretti, A., Gambioli, A., Nisi, B., 2022. Contamination assessment and temporal evolution of nitrates in the shallow aquifer of the Metauro River Plain (Adriatic Sea, Italy) after remediation actions. Int. J. Environ. Res. Public Health 19, 12231.
Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data. In: Compositional Data Analysis: Theory and Applications, pp. 341–355.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat Methodol. 58, 267–288.
Tsagris, M., 2021. The k-NN algorithm for compositional data: a revised approach with and without zero values present. J. Data Sci. 12, 519–534. https://doi.org/10.6339/JDS.201407-12(3).0008.