A FACTOR ANALYSIS OF HIDROCHEMICAL COMPOSITION OF LLOBREGAT RIVER BASIN

N. Otero[1], R. Tolosana-Delgado[2] and A. Soler[1]
[1]Departament de Cristal·lografia, Mineralogia i Dipòsits Minerals. Facultat de Geologia
Universitat de Barcelona, Barcelona, Spain; notero@geo.ub.es
[2]Departament d'Informàtica i Matemàtica Aplicada.
Universtat de Girona, Girona, Spain; raimon.tolosana@udg.es

Abstract

Hydrogeological research usually includes some statistical studies devised to elucidate mean background state, characterise relationships among different hydrochemical parameters, and show the influence of human activities. These goals are achieved either by means of a statistical approach or by mixing models between end-members. Compositional data analysis has proved to be effective with the first approach, but there is no commonly accepted solution to the end-member problem in a compositional framework.

We present here a possible solution based on factor analysis of compositions illustrated with a case study. We find two factors on the compositional bi-plot fitting two non-centered orthogonal axes to the most representative variables. Each one of these axes defines a subcomposition, grouping those variables that lay nearest to it. With each subcomposition a log-contrast is computed and rewritten as an equilibrium equation. These two factors can be interpreted as the isometric log-ratio coordinates (ilr) of three hidden components, that can be plotted in a ternary diagram. These hidden components might be interpreted as end-members.

We have analysed 14 molarities in 31 sampling stations all along the Llobregat River and its tributaries, with a monthly measure during two years. We have obtained a bi-plot with a 57% of explained total variance, from which we have extracted two factors: factor G, reflecting geological background enhanced by potash mining; and factor A, essentially controlled by urban and/or farming wastewater. Graphical representation of these two factors allows us to identify three extreme samples, corresponding to pristine waters, potash mining influence and urban sewage influence. To confirm this, we have available analysis of diffused and widespread point sources identified in the area: springs, potash mining lixiviates, sewage, and fertilisers. Each one of these sources shows a clear link with one of the extreme samples, except fertilisers due to the heterogeneity of their composition.

This approach is a useful tool to distinguish end-members, and characterise them, an issue generally difficult to solve. It is worth note that the end-member composition cannot be fully estimated but only characterised through log-ratio relationships among components. Moreover, the influence of each end-member in a given sample must be evaluated in relative terms of the other samples. These limitations are intrinsic to the relative nature of compositional data.

1 Introduction

The statistical analysis of compositional data (Aitchison, 1986) has offered solutions to many problems, specially those related to distances: cluster analysis (Martin-Fernandez, 2001), discriminant analysis (Barceló-Vidal, 1996), regression (Buccianti et al, 1999; Daunis-i-Estadella et al., 2002), even principal component analysis (Aitchison 1984, 2001, Aitchison and Greenacre, 2002), have been adapted to the compositional distance, with different degrees of clarity and success. However, there are important problems, especially in geosciences, yet unsatisfactorily solved: two of these are the end-member problem and factor analysis.

The so-called end-member problem attempts to find an unknown, though considered small, number of sources and their *pure* composition, in order to be able to reproduce every sample as a convex linear combination of these target sources (Renner, 1995). There are several *Euclidean* attempts to solve this

problem: analyzing data sets in an R-mode factor analysis framework (Renner, Glasby and Walter, 1997), putting forward some procedures to expand too-small a priori known convex hulls (Renner, 1995, Weltje, 1997), using simulated annealing to choose end-member composition from an *a priori* library (Penn, 2002), among others (details in Renner, Glasby and Walter, 1997; Weltje, 1997; and Penn, 2002). From another point of view, once known the number and composition of the present end-members, there are several alternative approaches to the linear mixture model, as well as some tests to compare them (Aitchison and Bacon-Shone, 1999).

Factor analysis is a modification of a principal component analysis where once the principal components are extracted they are rotated/recombined in order to obtain some "factors" that are easier to interpret than principal components themselves. Bi-plots and log-contrasts have been used to reduce the dimensionality of a data set (a classical PCA application) by selecting a most-variant subcomposition (Aitchison, 1984). However, the recent introduction of isometric log-ratio (*ilr*) transformations and coordinates (Egozcue et al., 2003) offers the definition of basis associated to partitions and, particularly, to subcompositions. Thus, factorial subcompositions become *ilr* axes.

The aim of this study is to develop a possible alternative solution to the end-member problem based on subcompositional factor analysis of compositional data. We illustrate our approach with a data set of the Llobregat River Basin (NE Spain) containing 14 molarities measured monthly in 31 sampling stations.

2 Methodology

First step: computing the bi-plot

Compositional bi-plots were initially developed based on a *clr* transformation (Aitchison, 1986), although today they are identified with a Singular Value Decomposition (SVD) of the compositional data matrix according to the Euclidean space structure of the Simplex (Aitchison and Greenacre, 2002). Let us review this last work to introduce the necessary notation and matrices further used in this communication.

The data set is expressed as a matrix $\underline{\boldsymbol{X}}$ with N rows (individuals) and D columns (components), where all elements in any row sum up to the unity. This data matrix is *clr*-transformed, and further centered to obtain another matrix $\underline{\boldsymbol{Z}}$ with the same dimension, but where all elements sum up to zero both by rows and columns; this matrix has a rank $r \leq \min(N,D)$. Then, $\underline{\boldsymbol{Z}}$ can be factorized as the product

$$\underline{\boldsymbol{Z}} = \underline{\boldsymbol{U}} \cdot \underline{\boldsymbol{\Gamma}} \cdot \underline{\boldsymbol{V}}^t, \tag{1}$$

where $\underline{\boldsymbol{U}}$ and $\underline{\boldsymbol{V}}$ are the matrices of left and right singular vectors, and $\underline{\boldsymbol{\Gamma}}$ is the diagonal ($r \times r$) matrix of singular values. All these matrices have a rank r. $\underline{\boldsymbol{U}}$ has as many rows as individuals in $\underline{\boldsymbol{Z}}$, and $\underline{\boldsymbol{V}}$ has as many rows as variables in $\underline{\boldsymbol{Z}}$; both of them have r independent columns, called singular vectors. The covariance bi-plot (the kind of bi-plot most used in compositional applications) rewrites $\underline{\boldsymbol{U}}=\underline{\boldsymbol{F}}$ as a matrix containing the standard coordinates of each individual in the r-dimensional space, and $\underline{\boldsymbol{\Gamma}} \cdot \underline{\boldsymbol{V}}^t = \underline{\boldsymbol{G}}^t$ as a matrix with the principal coordinates of the variables in this space.

The Eckart-Young Theorem (Eckart and Young, 1936) states that, from all matrices of rank r^*, the best approximation to $\underline{\boldsymbol{Z}}$ can be computed using (1) with the first r^* singular values and their corresponding singular vectors. This approximate matrix is expressed as $\underline{\boldsymbol{Z}}_{r^*}$. As a consequence, the best approximation of rank r^* to the covariance matrix of the original data ($\underline{\boldsymbol{S}} = 1/N \cdot \underline{\boldsymbol{Z}}^t \cdot \underline{\boldsymbol{Z}}$) can be computed as $\underline{\boldsymbol{S}}_{r^*} = 1/N \cdot \underline{\boldsymbol{G}} \cdot \underline{\boldsymbol{G}}^t$. In Principal Component Analysis (PCA), there exists a statistic describing the general goodness of this approximation, the *total explained variance* (Γ_{r^*}), computed as

$$\Gamma_{r^*} = \frac{\mathrm{Tr}(\underline{\boldsymbol{S}}_{r^*})}{\mathrm{Tr}(\underline{\boldsymbol{S}})} = \frac{\sum_{i=1}^{r^*} \gamma_i^2}{\sum_{i=1}^{r} \gamma_i^2}, \tag{2}$$

where γ_i is the i-th singular value of the matrix $\underline{\boldsymbol{\Gamma}}$. Equation (2) gives a global measure of fit of the reduced system to the original system. When we need an assessment of the degree of approximation for each variable, then PCA offers the *communality* (ζ) statistic

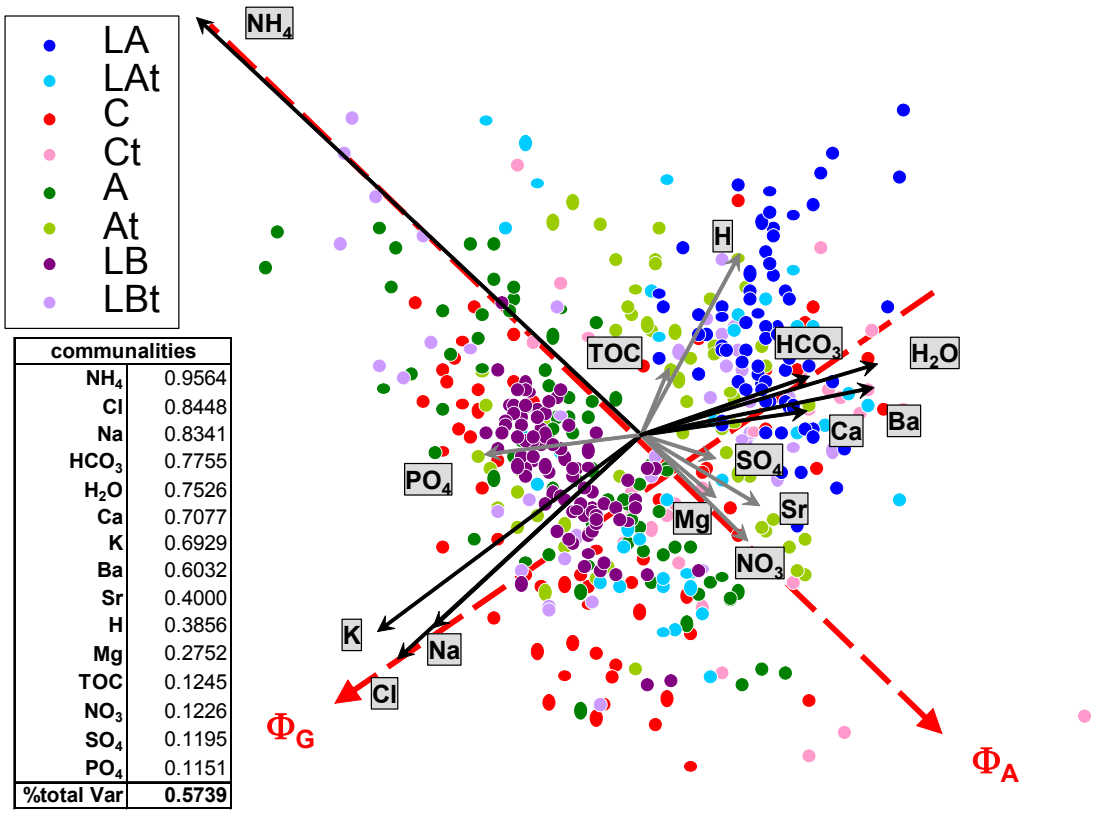
$$\zeta_i = \frac{s_{r^*,ii}}{s_{ii}} = \frac{\sum_{j=1}^{r^*} g_{ij}^2}{\sum_{n=1}^{N} z_{ni}^2}, \qquad (3)$$

being s_{ii} the i-th diagonal element of the covariance matrix (thus, the covariance of the i-th variable), $s_{r^*,ii}$ the i-th diagonal element of the approximate covariance matrix, g_{ij} and z_{ni} elements of the matrices $\underline{\boldsymbol{G}}$ and $\underline{\boldsymbol{Z}}$ respectively. Both the total explained variance and the communality of a variable are positive and their maximal value (meaning perfect fitting) is the unity.

If r^*=2 then the SVD can be easily represented in a *bi-plot* (Gabriel, 1971): individuals are represented as dots, using their principal coordinates of the matrix $\underline{\boldsymbol{F}}$, while variables are plotted as arrows (*rays*) with the tail in the origin and the head at the coordinate points in matrix $\underline{\boldsymbol{G}}$ (the principal components). Figure 1 contains the bi-plot of the Llobregat River data set: there is also the communality of each variable (those with ζ_i<0.5 have been shaded), as well as the total explained variance, a low 57%.

Interpretation of the bi-plot is based on the *links*: a link is the segment between two variable ray heads:

- the length of the *ij*-link between variables $z_{\cdot i}$ and $z_{\cdot j}$ is proportional to Var[ln($z_{\cdot i}$ / $z_{\cdot j}$)],
- the cosine of the angle formed by the *ij*-link and *kl*-link is proportional to Corr[ln($z_{\cdot i}$ / $z_{\cdot j}$), ln($z_{\cdot k}$ / $z_{\cdot l}$)],
- if a single link passes through several ray heads, then the angles between their *ij*-links are close to zero, their correlation coefficient is almost 1, and the represented variables are mutually proportional



communalities	
NH_4	0.9564
Cl	0.8448
Na	0.8341
HCO_3	0.7755
H_2O	0.7526
Ca	0.7077
K	0.6929
Ba	0.6032
Sr	0.4000
H	0.3856
Mg	0.2752
TOC	0.1245
NO_3	0.1226
SO_4	0.1195
PO_4	0.1151
%total Var	0.5739

Figure 1 Bi-plot of the Llobregat River data set, according to different sub-basins; communalities (ζ_i) of each variable are included in a table. The codings are: *LA*-high Llobregat course, *C*-Cardener, *A*-Anoia, *LB*-low Llobregat course; a *t* indicates the tributaries of the corresponding sector.

Second step: fitting two orthogonal axes

The goal of this step is the definition of two independent subcompositions that describe most of the variability, using the facts that:

- a single link among several variables defines a subcomposition with a single degree of freedom,
- the center of the bi-plot is an arbitrary point, and the only interesting feature are the links (of those variables with high communality), thus it is not necessary that the links pass through the center,
- bi-plots are representations of a SVD in the geometry of the simplex (Aitchison, 2001), thus the principal axes defining the bi-plot are naturally orthogonal and isometric with respect to the Aitchison's metric of the simplex, thanks to this fact, "independence of the subcompositions" will be translated to "choosing a pair of orthogonal links".

This fit can be done visually in the bi-plot. The development of an automatic procedure on the coordinates in the $\underline{\boldsymbol{G}}$ matrix is left for further work. Each one of the variables that lay acceptably near a link will be included in the corresponding subcomposition. Figure 1 shows these two axes, labeled as $\boldsymbol{\Phi}_{\mathbf{G}}$ (from *geological*) and $\boldsymbol{\Phi}_{\mathbf{A}}$ (from *anthropogenic*). These labels have been chosen due to a prior knowledge on the natural processes and pollution sources in the area.

Third step: computing the factor log-contrasts

Once fitted the two axes and grouped the original components in two subcompositions, a known log-contrast analysis (Aitchison, 1984) should be applied to each one of them. However, if the bi-plot explains a low proportion of total variance or one of the subcompositions is mainly formed with low-communality variables, then this procedure could lead to inconsistent (non-orthogonal) factors.

As an alternative, since the variables are already expressed in a maximal variant (*ilr*) plane in the $\underline{\boldsymbol{G}}$ matrix, we suggest to use it, applying a projection and renormalization procedure, as follows:

1. define the head and the tail of the main link
2. compute the new rays from the tail of the main link to each one of the involved variable heads
3. compute the scalar projection of each new ray in the main link; this value is the non-centered weight of each variable in the factor
4. compute the arithmetic mean of the non-centered weights and subtract it from them, to compute the centered weights
5. round reasonably the centered weights to integer values that sum up to zero; the Appendix contains the weights of the variables linked to each of the two factors as well as these intermediate steps of computation.
6. compute the norm of the vector of weights, and compute the factor values $\boldsymbol{\Phi}$ for each individual n; these factors are like

$$\Phi_{(AG),n} = \frac{1}{|\boldsymbol{B}|}\sum_{i=1}^{D}\beta_{(AG),i}\cdot z_{ni}\ , \quad 0=\sum_{i=1}^{D}\beta_{(AG),i}\ ,$$

where $\underline{\boldsymbol{B}}=(\beta_i)$ is the vector of integer weights obtained in the last step; in our example, this can be done with expressions:

$$\Phi_G = \frac{1}{\sqrt{166}}\log\frac{\left[Na^+\right]^6\cdot\left[Cl^-\right]^6\cdot\left[K^+\right]^6}{\left[Mg^{2+}\right]^2\cdot\left[SO_4^{2-}\right]^2\cdot\left[Ca^{2+}\right]^3\cdot\left[HCO_3^-\right]^3\cdot\left[H_2O\right]^4\cdot\left[Ba^{2+}\right]^4}\ , \tag{4}$$

$$\Phi_A = \frac{1}{\sqrt{32}}\log\frac{\left[Mg^{2+}\right]\cdot\left[SO_4^{2-}\right]\cdot\left[Sr^{2+}\right]\cdot\left[NO_3^-\right]^2}{\left[NH_4^+\right]^5}\ , \tag{5}$$

7. ensure orthogonality of computed weight vectors; if the two subcompositions do not have any common component, then they will be automatically orthogonal; otherwise, it is easy to compute the angle α between them:

$$\cos\alpha = \frac{\langle \boldsymbol{B}_A | \boldsymbol{B}_G \rangle}{|\boldsymbol{B}_A| \cdot |\boldsymbol{B}_G|},$$

which in our case gives a value of cos α = 0.055 thus α =86.85º. This means that factors Φ_A and Φ_G are not actually coefficients in an *ilr* basis, since they are not perfectly orthogonal. However, we accept this discrepancy in order to obtain a better interpretability of the factors. A future development of this technique could be the systematization of the fitting process to directly avoid these deviations from orthogonality, for instance, accounting for the uncertainty in the positions of the head of each ray.

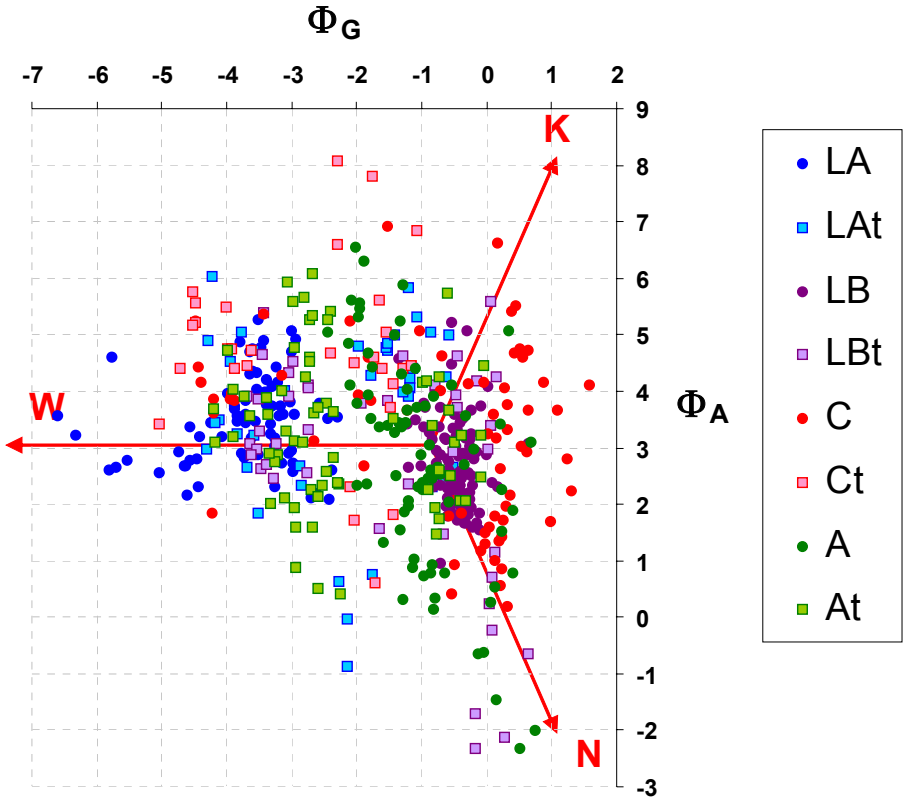


Figure 2 Scatterplot of factors $\boldsymbol{\Phi_G}$ vs $\boldsymbol{\Phi_A}$, with indication of principal *clr*-directions.

Fourth step: bivariate analysis of the factors

The computed values of the two factors can be plotted in a scatterplot (Figure 2). This plot is in fact (almost) the rotated bi-plot, and it is the representation of the composition expressed in (almost) isometric log-ratio coordinates defined by factors $\boldsymbol{\Phi_G}$ and $\boldsymbol{\Phi_A}$. Interesting issues of this plot are explained in section 3. However, since this representation is a projection in a nearly *ilr*-plane, we may try to re-express these factors as coordinates of a new composition formed with three *hidden* components. These hidden factorial components shall coincide with a set of clr-directions projected onto the same *ilr*-plane, thus mutually at 120º: The same scatterplot of Figure 2 includes them, visually fitted to the data dispersion.

Fifth step: computation of the factorial components

Once decided the three directions of the hidden factorial components, we can compute them applying usual expressions of calculation of components from *ilr* coordinates. This gives in our case an expression like

$$(K, N, W) = \mathcal{C}\left(\exp\left(\tfrac{1}{\sqrt{6}}\,\mathbf{\Phi}_{\mathbf{G}} + \tfrac{1}{\sqrt{2}}\,\mathbf{\Phi}_{\mathbf{A}}\right), \exp\left(\tfrac{1}{\sqrt{6}}\,\mathbf{\Phi}_{\mathbf{G}} - \tfrac{1}{\sqrt{2}}\,\mathbf{\Phi}_{\mathbf{A}}\right), \exp\left(-\tfrac{2}{\sqrt{6}}\,\mathbf{\Phi}_{\mathbf{G}}\right)\right), \quad (6)$$

where $\mathcal{C}$ is the closure operation, and (*K, N, W*) the new hidden composition. Afterwards, they can be represented in a ternary diagram, as it is shown in Figure 3

Finally, if some end-members actually *exist* in the problem under analysis, they might lay near the *clr*-axis defined in the last step: this is a conjecture based on the observation of the presented plots and the knowledge of the composition of the known sources of water and pollution in this basin (see the details on this topic in section 3). Furthermore, if the procedure here explained does not lead to a clear triad of suitable *clr*-directions, it seems reasonable to statistically test whether the mixture model is acceptable. In our opinion, the development of this test should be based on geometry-free procedures, preferably with a probabilistic approach, like the general models of upper order statistics or of extreme values (Embrechts, Klüppelberg and Mikosch, 1997).

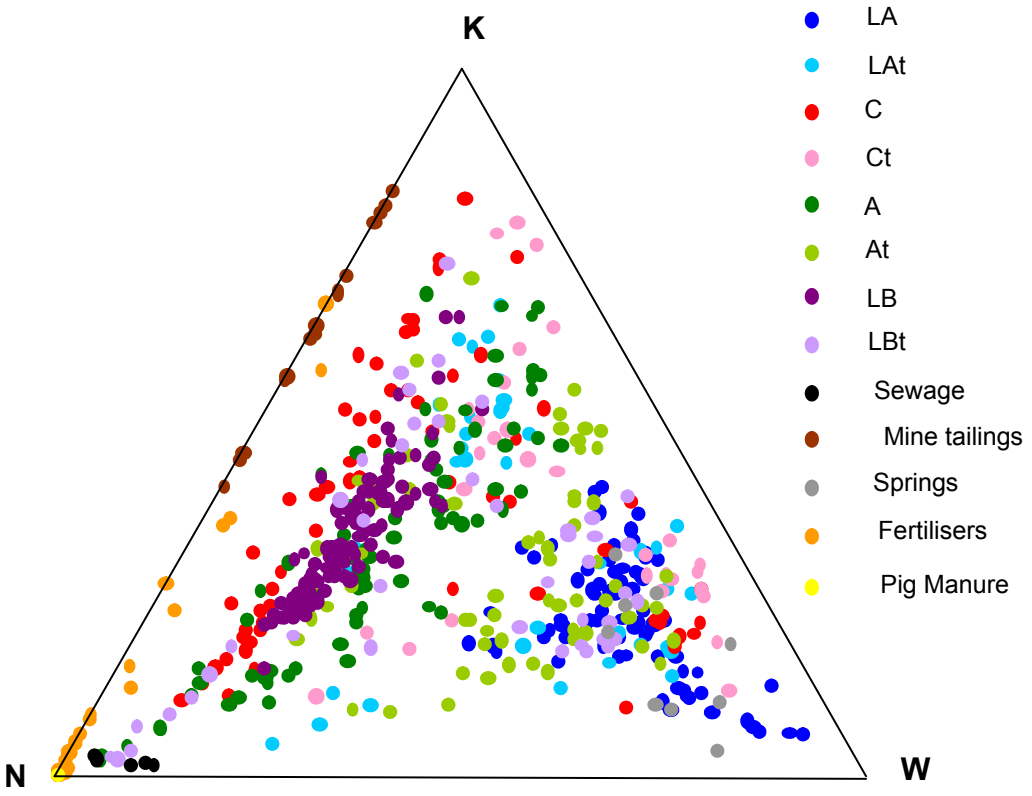


Figure 3 Ternary diagram of the computed hidden composition: W (pristine waters) K (potash sources) and N (ammonium sources).

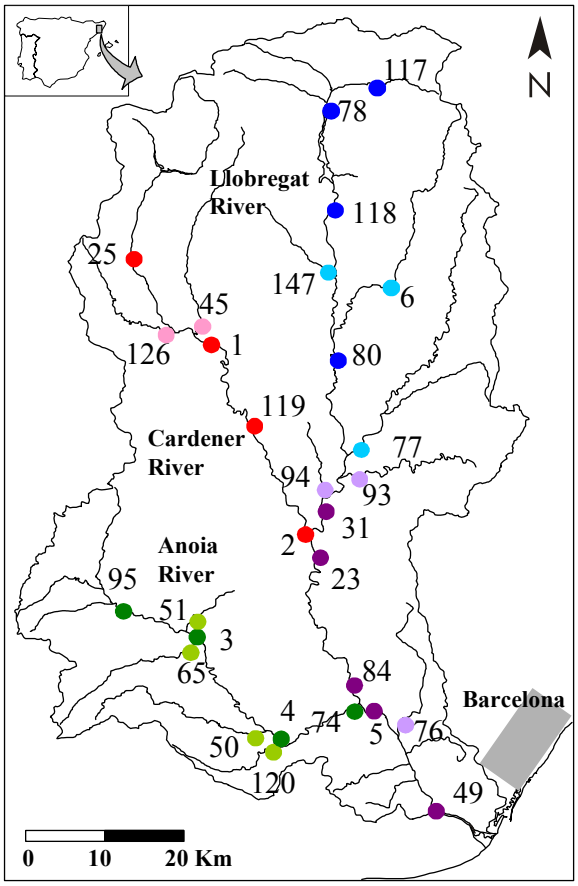
3 Application

This methodology has been tested with a compositional data set obtained from the Llobregat River Basin. Measurements were taken at 31 stations, monthly between June 97 and January 99, plus another in April 99. Out of the more than 30 suitable variables in the data file (Soler et al., 2002), we studied 14 molarities: H+, Na+, K+, Ca2+, Mg2+, Sr2+, Ba2+, NH4, Cl-, HCO3-, NO3-, SO42-, PO43- and total organic carbon –TOC–.

The Llobregat River is located in NE Spain. It drains an area of 4948.2 Km^2, and is 156.6 Km long, with two main tributaries, the Cardener and Anoia Rivers (Figure 4). The headwaters of the Llobregat and Cardener Rivers are in a rather unpolluted area of the Eastern Pyrenees. Mid-waters, the rivers flow

through a densely populated and industrialized area, where potash-mining activity occurs and there are large salt mine tailings stored with no water proofing. Moreover, in this area the main land use is agriculture and stockbreeding. The lower course flows through one of the most densely populated areas of the Mediterranean region and waters receive large inputs from industry and/or urban origin while intensive agriculture activity is again present near the mouth, at the Llobregat delta. Anoia River is quite different. Its headwaters are in an agricultural area, down-water it flows through an industrialized zone (paper mills, tannery and textile industries), and near the confluence with the Llobregat River the main land use is agriculture again, mainly vineyards, and there is a decrease in industry and urban contribution (Soler et al., 2002).

Results of previous work on this area show that the chemistry of most stream waters is mainly controlled by the weathering of the Tertiary chemical sediments within the drainage basin (Soler et al., 2002). Figure 5 shows a simplified lithologic map of the Llobregat basin, where the geochemical signature of natural bedrock is derived from weathering of a) limestone and marls, with major ions HCO3−, Ca2+, Mg2+ and minor ions Sr2+ and Ba2+, b) gypsum, with major ions SO42−, Ca2+ and minor ions Sr2+ and Ba2+, and c) halite and silvite, with major ions Na+, Cl−, K+ and minor ions Mg2+, SO4 and Ca2+.



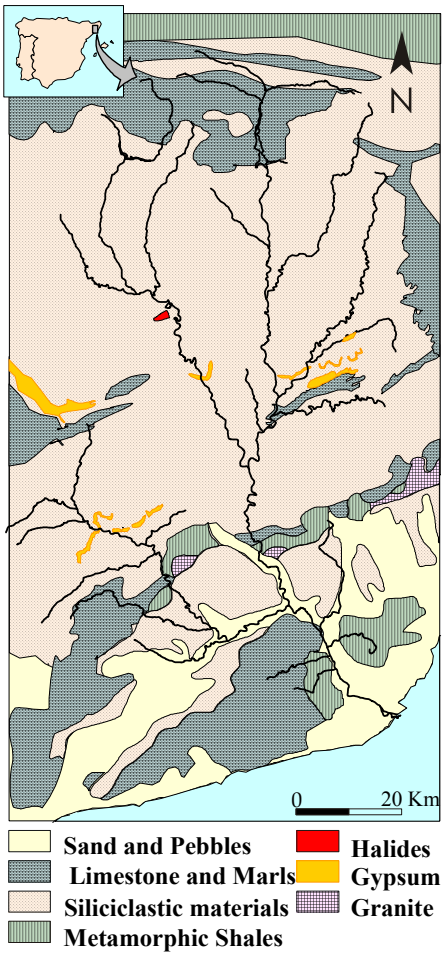


Figure 4 (left) Llobregat basin showing sample location.

Figure 5 (right) Geological map of the Llobregat River Basin

The major sources of anthropogenic pollution in the basin are mainly identified Soler et al. (2002): a) potash mine tailing, with major ions Na+, Cl−, K+ and minor ions Mg2+, SO4 and Ca2+; b) fertilizers with NO3−, PO43−, K+ and SO4 as major ions -; c) stockbreeding NH4+, NO3−, and maybe TOC; and d) urban and/or industrial sewage, with major ions as NH4+, PO4 and TOC. In addition, coupling chemical data with isotopic compositions from strontium (Antich et al., 2000, 2001) and sulfur (Otero and Soler, 2002), the contribution of these sources to water pollution has been quantified to some extent.

Given the rather high variability of this basin, both in the richness of its geological background and the human activities developed on it, different groups are displayed in the plots. A first division is made between the three major rivers: Anoia, Cardener and Llobregat, the latter further divided into higher and lower course. A priori, each group is known to have a distinct geochemical print: Anoia River is richer in sulfates, Cardener River has the salt outcrops and major potash mining activities in its basin, and high Llobregat River flows across a carbonate and siliciclastic landscape. It goes without saying that the lower course of the Llobregat River mixes all of these prints, and is characterized by high values of all the major ions.

Results of a previous study highlighted two factors defined as equilibrium equations between some of the available components (Tolosana-Delgado et al., submitted), thus offering a reduction of dimensionality following the log-contrast approach (Aitchison, 1984). A detailed analysis of these two factors suggested the existence of end-members, supported by the available analysis of known inputs in the area.

Univariate characterisation of factor logcontrasts (Φ_G and Φ_A)

Results of Factor Φ_G throughout the sampling period are shown in Figure 6 and Figure 7. Factor Φ_G is surely reflecting the geological background, enhanced by mining. Low values indicate that the sample has low anthropogenic influence. High values of that factor are achieved by high concentrations in Na, Cl and K, the major ions of halide outcrops and potash mine lixiviates. However, in the areas without halide outcrops or mining activity, high values of that factor must have an alternative anthropogenic source, as urban and industrial sewage (Na, Cl) or fertilisers (K).

- Lowest values of Φ_G, from -6 to -4, correspond to pristine waters, as it can be observed at locations 117, 25, 45 and location 6, to some extent.

- Highest values of Φ_G, up to +1, correspond either to potash mining influence, which is very clear at locations 119 and 2, and down-waters from location 31; or to high urban/industrial input, as in locations 3 and 76. Location 119, after the mining area, has the highest value. Location 3 has similar values although it is located outside the mining area; this can be explained by the contribution of the local industrial activity, specially the tanneries, that use Na-Cl brines from the mining areas. To support this, at location 3, the lowest values of this factor are achieved in aug-97 and aug-98; notice that August is holiday time and the industrial activity decreases. This fact is reflected down-waters, in locations 4 and 74, where the minimum values of Φ_G are reached in the same month.

- There is a sudden increase of Φ_G passing through the mining areas at the middle section of the Cardener and the Llobregat Rivers. Another slight increase is observed in the Llobregat River after the confluence with the Cardener River.

- At several sampling locations, an overall increase in the Φ_G value is observed throughout the sampling period, e.g. locations 3, 4, 50, 31, 65, 74, 77, and 93; this is possibly indicating an increasing influence of anthropogenic sources. At location 119 there is an overall diminution of Φ_G from jun-97 to aug-99, suggesting an improvement in the control of the potash mining lixiviates.

Factor Φ_A has a clearly anthropogenic influence as NH_4^+ or NO_3^- only have significant anthropogenic sources. However, the interpretation of this factor is more difficult due to its high variability and the presence of anthropogenic compounds in both numerator and denominator. Pristine waters have intermediate values, low values can be interpreted as a high urban contribution, influenced by the NH_4^+ contents, and high values are not clearly associated to a pollution source, although fertilizers are good candidates due to their high NO_3^- and SO_4^{2-} concentration.

- Pristine streams have intermediate values, from +2 to +4, as it can be observed at locations 117 and 25, in the headwaters.

- Negative values, from 0 to –2, correspond to high urban influence, as it can be observed at location 76, and at several samples of locations 147, 94 and 3.

- Highest values, from +5 to +7, are in locations 95, 51, 45 and 126, placed in agricultural areas. One possible origin of these values is a fertilizer influence.

- There is a sudden decrease of Φ_A from location 119 to location 2 at the Cardener River, after the city of Manresa, which can be explained by a high urban contribution. Another sudden decrease is observed between locations 95 and 3 in the Anoia River, when passing the city of Igualada.

- The tendencies throughout the sampling period are not so clear as those involving factor Φ_G, but some trends are observed: at location 2, in the Cardener Basin a diminution of Φ_A is observed throughout the sampling period, suggesting an increasing contribution of the urban input. Location 95 in the Anoia Basin shows the same trend, suggesting an increase in the urban influence or/and a diminution in the rural (fertilizer) influence. A similar tendency is observed at locations 74 and 65. At locations 80 and 31 an increase throughout the sampling period is observed.

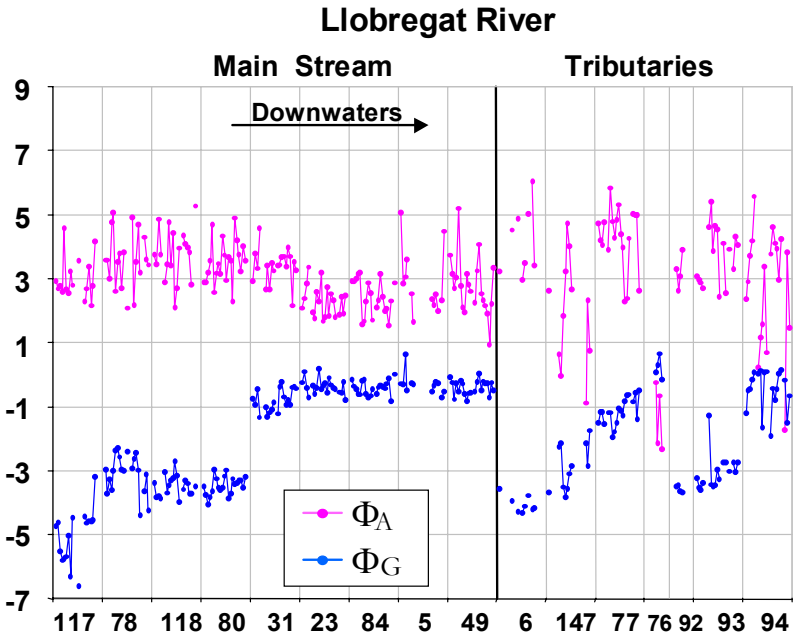


Figure 6 Down-waters evolution of factors Φ_G and Φ_A in the Llobregat River main stream, and some tributaries. At each location samples are represented throughout the sampling period (jun-97-aug-99).

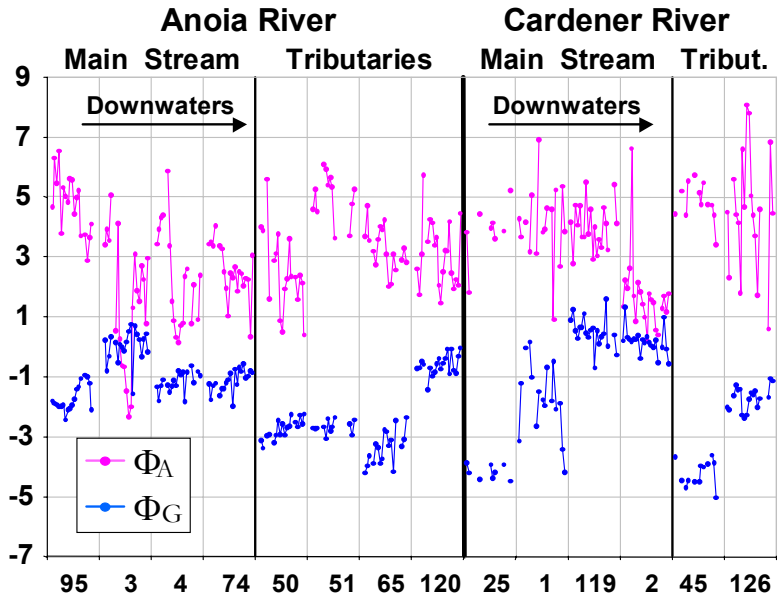


Figure 7 Down-waters evolution of factors Φ_G and Φ_A in the Cardener and Anoia Rivers and their tributaries. At each location samples are represented throughout the sampling period (jun-97-aug-99).

Bivariate analysis of factor logcontrast (Φ_G vs Φ_A)

Figure 8 shows a plot of Φ_G vs Φ_A. Considering only the main course of each River, Llobregat River High waters (117, 78, 118 and 80), together with location 25 from the headwaters of the Cardener River, are plotted in an area clearly distinct from the other samples, characterized by negative values of Φ_G (lower than –3) and intermediate values of Φ_A (from +2 to +4). Anoia River samples, except location 3, fall in an area with Φ_G between -2 and -1 and Φ_A-values from 0 to +6, Cardener River locations 119 and 2, and location 3, cluster together in an area with the highest values of Φ_G, up to +1, and Φ_A ranges from 0 to +5. Llobregat River Low waters are all clustered in an area between Cardener and Anoia Rivers with Φ_G from -1 to 0 and Φ_A from +1 to +4. Detailed down-water evolution shows the following issues:

- Llobregat headwaters are characterized by negative values of Φ_G, down to -6, and intermediate values of Φ_A, from +2 to +4, these values indicate a possible end-member: pristine waters. Downwaters, at location 78 there is an increase in Φ_G while Φ_A remains fairly constant. Location 118 shows a slight diminution of Φ_G; one possible explanation is the presence of the main sweet water reservoir in the basin, located between locations 78 and 118, which can induce an homogenizing effect. Location 80 plots in the same area indicating that there is no significant influence of the tributaries 6 and 147, probably due to their small flow. From location 80 to location 31 there is a drastic increase of Φ_G as the River crosses the potash mining area, although the influence of the tributaries discharging between these two locations can not be discarded, especially locations 77 and 94. Downstream, at location 23 the samples show an increase in Φ_G coupled with a decrease in Φ_A showing the influence of the input of the Cardener River, controlled by urban wastewater (see location 2). Down-waters of location 23 all samples have similar values of Φ_G, and the distinctive change is an increase in the variability of Φ_A at location 5, after the confluence with the Anoia River, influenced by the variability of that factor at location 74.

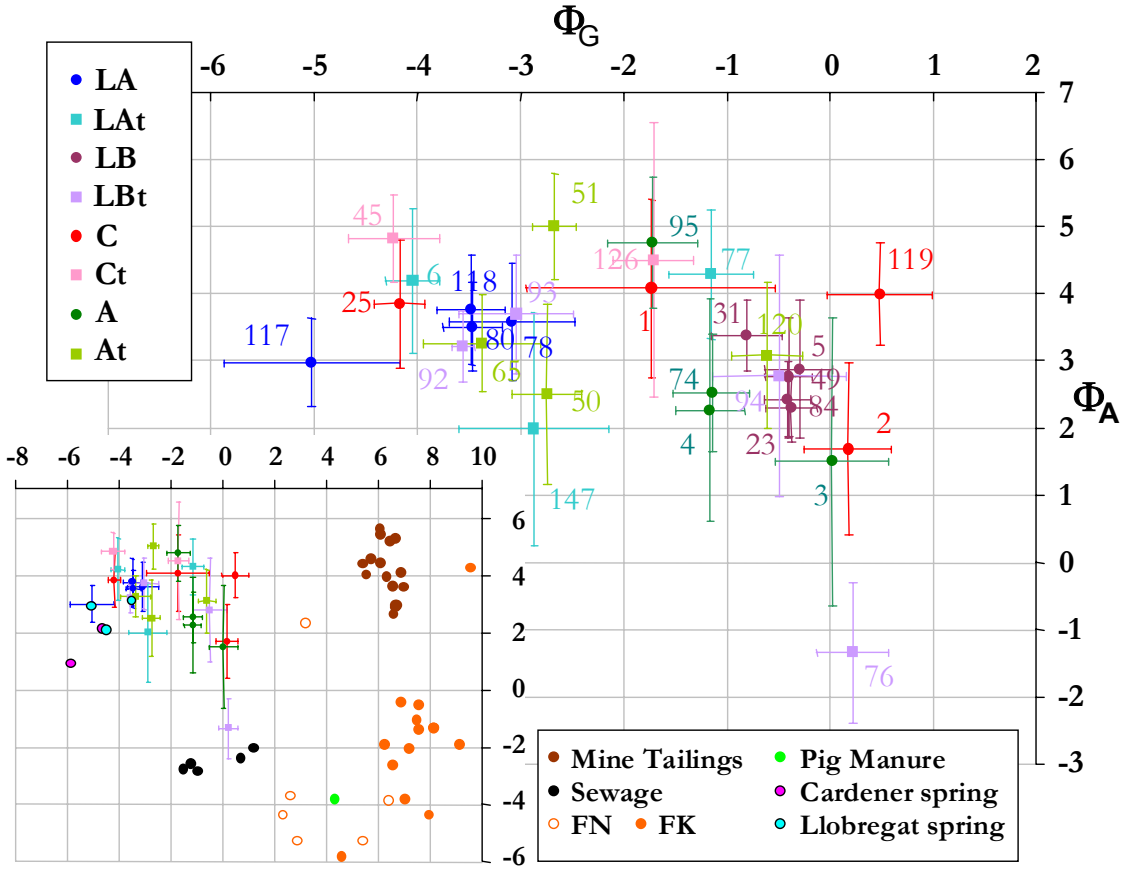


Figure 8 Factor Φ_G vs factor Φ_A showing mean values of each sampling station and the standard deviation. The main inputs in the area are also represented in the small figure (*FN* and *FK* are respectively Nitrogen- and Potassium-rich fertilizers).

- Regarding the tributaries of the Llobregat River, they usually show higher variability than the main stream. Location 6 has values similar to those from pristine waters; location 147 has values of Φ_G slightly higher than location 6 and Φ_A shows lower values and large variability, besides as Φ_G increases, Φ_A decreases (see Figure 6), indicating an influence of urban wastewater. Location 77 shows the highest Φ_G values of all the sub-basin; moreover Φ_A has values up to +6, suggesting fertilizers as the probable source of pollution. Location 76 presents high Φ_G-values, up to +0.5, coupled with the lowest values of Φ_A for the entire basin, down to –2.5, marking a second possible end-member: the urban/industrial influence. Locations 92 and 93 have values of Φ_G and Φ_A similar to relatively pristine waters in the main course, indicating low anthropogenic influence. Location 94 is characterized by high Φ_G-values and intermediate Φ_A-values, with a large variability in both factors, some samples are similar to location 76, indicating an urban influence; and other samples are close to location 119, suggesting a contribution of the potash mining activity.

- Cardener River headwaters have similar values to those of Llobregat headwaters, with negative values of Φ_G, down to -4, and intermediate values of Φ_A, from +3 to +5. Downwaters, at location 1, there is an overall increase in Φ_G and samples show the highest variability of that factor for the entire basin, (due to a bad sampling location, or an influence by location 126). After the mining areas and the evaporitic outcrops, at location 119, there are no changes in Φ_A and there is a drastic increase in Φ_G up to +1, the highest values of the entire basin, indicating a third possible end-member: the potash mining lixiviates. Down-waters of location 119, at location 2, there is a slight diminution in Φ_G and a clear decrease in Φ_A indicating a reduction in the mining influence and an increasing contribution of urban wastewater. This evolution has been already observed by Tolosana-Delgado et al. (submitted).

- The two tributaries collected at the Cardener Basin plot in two clusters well differentiated. Location 45 has values similar to pristine waters but with Φ_A-values slightly higher, possibly influenced by fertilizers, since mean values of NO_3 in location 45 are tenfold those from location 25 and it has lower NH_4 contents. Samples of location 126 show intermediate values of Φ_G (around -1) and a high variability of Φ_A (from +0.5 to +8), reaching the highest values of that factor for the entire basin. This high variability could be explained by a mixture of two populations, one with a urban contribution (low Φ_A-values) and the other with a "rural" influence (high Φ_A-values).

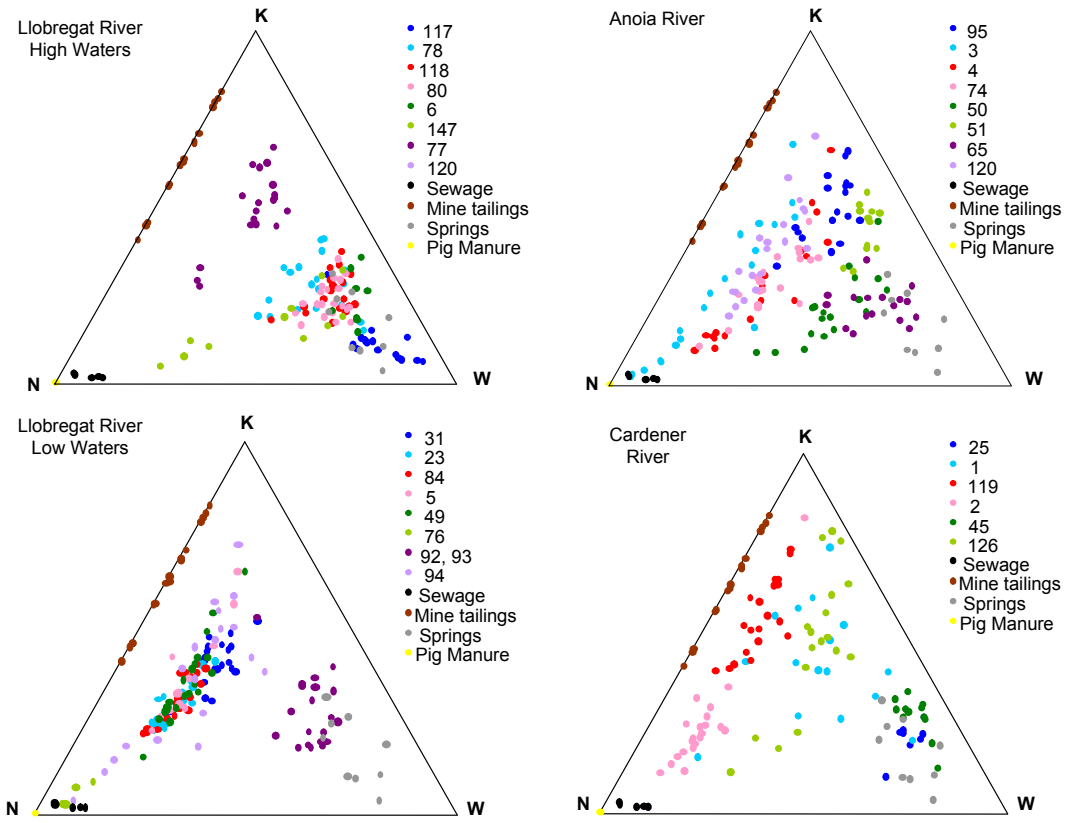


Figure 9 Ternary diagrams (K, N, W) of each sub-basin.

- Anoia River headwaters at location 95, fall far from pristine waters of the Llobregat and the Cardener Rivers, with Φ_G ranging from –1.5 to 0 and Φ_A from +2.5 to +6.5. This values can not be explained by the different geological background of the sub-basin, as, if there are no halide outcrops, an anthropogenic influence must be invoked to achieve the Φ_G-values reported; fertilizers are the most suitable input to explain this values. Downwaters, at location 3, samples are characterized by higher values of Φ_G, from 0 to +1.5, and a large variability of Φ_A, from –2 to +5, this sample has a clear urban contribution and, as explained in section 3.1, the high values of Φ_G can be achieved with the contribution of the local industrial activity. Downstream of location 3, the locations 4 and 74 plot both within the same area, with lower values of Φ_G and minor variability of Φ_A, indicating a dilution with less polluted waters.

- Regarding the Anoia Tributaries, location 65 plot in the same area as the lowly polluted locations of the Llobregat high waters. Location 51 is characterized by Φ_G values close to -2 and it has the highest mean value of Φ_A (+5) in the entire basin, possibly influenced by fertilizers. Location 50 has similar values of Φ_G and lower values of Φ_A (+2.5) indicating a possible urban contribution. Location 120 has a high values of Φ_G, from -1 to +1, and maintains intermediate values of Φ_A; as for location 77, this behavior can not be explained by an urban contribution and another anthropogenic source must be invoked. This stream drains an agricultural area of vineyard, where K fertilizers are commonly used; this could explain the high values of Φ_G.

We have available analysis of the main inputs in the Llobregat River: springs (pristine waters), sewage (urban with variable industrial contribution), mine tailings, fertilizers and pig manure. We have calculated Φ_G and Φ_A of these inputs and plotted their values at Figure 8 together with the Llobregat River samples. Springs, mine effluents and sewage (and/or pig manure) seem to confirm the three end-members previously indicated by locations 117, 119 and 76 respectively. Fertilizers are not so clearly associated to an end-member due to the heterogeneity of their chemical composition. Besides, fertilizer factors are calculated with data from the bulk sample, and their values may not coincide with water polluted by fertilizers, e.g. NH_4 values are not representative, as volatilization and nitrification processes in the soil will affect the NO_3/NH_4 ratio. A similar problem would arise with the pig manure end member regarding NH_4, except if there is a direct spillage of pig manure on the stream.

Joint analysis of factorial components (K, N and W)

A ternary diagram of the factorial components, calculated as explained in section 2, fifth step, is shown at Figure 3. Due to the high variability of the different sub-basins, which can difficult their interpretation, detailed ternary diagrams of each sub-basin are shown in Figure 9, moreover the main inputs in the area are plotted in the same diagram.

Llobregat River high waters fall in an area close to the W end-member, indicating a major influence of pristine waters, except samples of location 77, that tend to the K end-member, and some samples of location 147 that tend to the N end-member. Cardener River high waters, together with location 45 lie close to the pristine springs, and the down-waters evolution towards K end-member at locations 1 and 119; and towards N end-member at location 2, are again observed. Samples of location 126 show a clear mixing trend between N and K end-members. Notice that the observed trend is a compositional line (Aitchison et al., 2002). Anoia River main stream lies away from pristine springs, location 95 tends to the K end-member, while downwaters there is a clear mixing trend between N and K, being location 3 the most influenced by anthropogenic inputs. Anoia tributaries plot closer to the W end-member than the main stream, except location 120. Regarding the Llobregat River low waters samples fall in a mixing trend between N and K end-members, excepting locations 92 and 93, closer to the W end-member.

With the ternary diagram, although absolute "mixing calculations" can not be made, we can calculate how far from pristine samples are those affected by anthropogenic inputs, computed as (non-closed) perturbations leading from that pristine sample to the sample at hand. We take as pristine waters locations 117 for the Cardener and Llobregat River, and for the Anoia River, due to the different geological background and the fact that its headwaters are indeed affected by anthropogenic inputs we choose as reference waters those from location 65. Results of this calculation are shown in Table 1. Although we cannot determine the absolute contribution of the different anthropogenic sources, in our case, as we have available analyses of the main pollutants, we can compare the results obtained for the samples with the results obtained with the inputs, giving a better approximation of the degree of "affection", e.g. location 76 and sewage have similar values, and values from location 119 are close to those from mine tailings.

Llobregat	ΔK	ΔN	ΔW
main: 117*	*1.000*	*1.000*	*1.000*
78	2.469	1.663	0.677
118	2.301	1.386	0.748
80	2.139	1.518	0.749
31	**3.693**	2.853	0.303
23	2.646	**4.140**	0.242
84	2.778	3.984	0.248
5	3.323	3.554	0.241
49	3.162	3.633	0.251
trib: 147	1.501	2.865	0.619
6	2.154	0.981	0.836
77	**4.524**	1.933	0.335
93	2.584	1.616	0.668
94	3.149	3.580	0.262
76	0.424	**7.002**	0.090

Cardener	ΔK	ΔN	ΔW
main: 25	1.857	1.055	0.867
1	3.898	1.914	0.429
119	**4.981**	2.580	0.162
2	2.152	**4.972**	0.176
trib: 45	2.402	0.726	0.842
126	**4.352**	1.618	0.412

Anoia			
main: 95	2.264	0.853	0.548
3	0.960	**2.994**	0.252
4	1.163	2.225	0.465
74	1.294	2.079	0.468
trib: 51	1.967	0.632	0.752
65*	*1.000*	*1.000*	*1.000*
(65)	2.035	1.700	0.734
50	0.927	1.514	0.832
120	1.695	1.892	0.379

K sources			
M	5.767	2.831	0.006
Mt	5.791	2.803	0.007

N sources			
P	0.040	7.866	0.003
SW	0.203	7.209	0.088

Table 1 Relative influence of K-Cl-Na pollution (K) and NH_4^+ pollution (N) in each location, compared with some known pollution sources: M=potash mining leaching, Mt=spring near potash mine, P= pig manure, SW=sewage. All location and source influences are expressed as non-closed perturbations from initial rather unpolluted states, marked by an asterisk (*): location 65 for Anoia waters, and location 117 for the other (including sources and location 65 itself enclosed in parentheses).

4 Discussion

The Llobregat River Basin, though relatively small, presents a high variability in its geological characteristics and human influence. However, we have been able to describe approximately half (57%) of this variability using only two factors: $\mathbf{\Phi_G}$ explains the geological background of the inflowing waters, and $\mathbf{\Phi_A}$ the dominant anthropogenic influence. As general features, it is interesting to notice: a) all river sources have low values of $\mathbf{\Phi_G}$, increasing down-waters; b) low values of $\mathbf{\Phi_A}$ clearly indicate sewage influence, while the presence of N-rich fertilizers should be noticed by an increment of this factor; c) a high influence of potash mining activities should increase specially $\mathbf{\Phi_G}$ and, in a minor degree, $\mathbf{\Phi_A}$; d) industrial sewage might be associated to increasing $\mathbf{\Phi_G}$ and decreasing $\mathbf{\Phi_A}$ values simultaneously.

Moreover, these factors can be converted to a hidden composition of the stated influences: W, K and N components represent the relative "influence" of pristine waters, potash pollution sources and ammonium pollution sources, respectively. To analyze this hidden composition, it is useful to compute the perturbation (Table 1) that leads from a supposedly pristine initial state to the actual state. Then, it becomes clear that in a general sense, the influence of both pollution sources increase down-waters, with special increments when passing the salt mines (locations 31 and 119, respectively mines in Sallent and Cardona-Súria), the cities (locations 126, 2+23 and 3, Solsona, Manresa and Igualada respectively) and some industrial areas (locations 3, specially tanneries, and 76, a complex mixing of industrial and urban wastewater, almost indistinguishable of a sewage water).

Finally, these factors are also useful to detect some time trends in these samples. Some locations in Anoia River show a minimum in $\mathbf{\Phi_G}$ at August, which shows again the human enhancement in a supposedly geological factor. Looking at time evolution, some locations show a better control of pollution sources (specially in the mining areas), although more of them show an inverse tendency: a clear increment of human influence, specially the inflow of urban sewage waters.

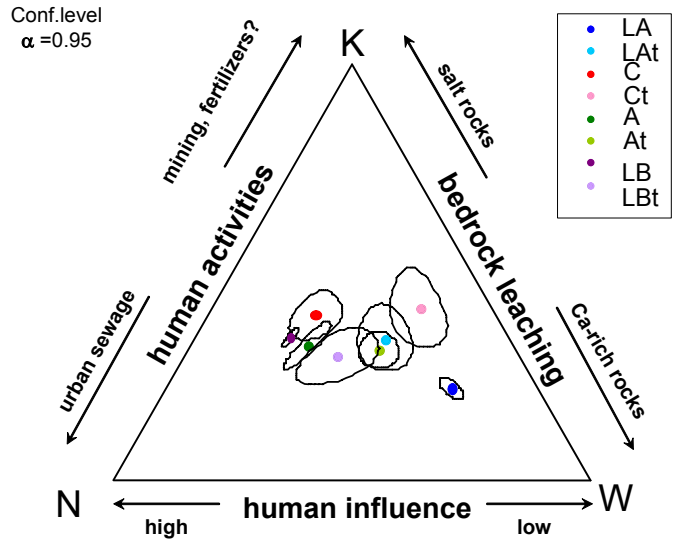


Figure 10 Summary ternary diagram, with influences and confidence regions of the geometric centers.

5 Conclusions

Using a methodology based on bi-plots and log-contrasts, we have been able to extract two factors from a data set of hydrogeochemical measurements in the Llobregat River watershed (NE Spain). The main interest of using this procedure to extract factors, is the ability to integrate many components in a single expression, mixing major and trace components. Since these factors are orthogonal, it seems reasonable to suppose them to be the *ilr* basis of a hidden composition, where each component represents an influence or source: W-pristine waters, K-potash sources and N-ammonium sources. Finally, the comparison of this hidden composition with a set of known pollution sources suggests the hidden components to be possible end-members. However, the actual composition in the original simplex of each end-member cannot be assessed. Thus we cannot obtain the contribution of each end-member to a given sample, only relative influences with respect to a chosen initial state are available from a compositional point of view.

Acknowledgements

We want to thank specially professors Dr, Vera Pawlowsky, Dr. Àngels Canals and Dr. Juan José Egozcue for their keen comments and inspiring ideas in the elaboration and revision of this document. This research has been funded by the *Dirección General de Enseñanza Superior e Investigación Científica* (DGESIC) of the Ministry of Education and Culture through the project BFM2000-0540, and by the projects HID99-0498 and REN2002-04288-CO2-02 of the *Comisión Interministerial de Ciencia y Tecnología* (CICyT), all of them institutions of the Spanish Government, as well as by the projects SGR01-00073 and 2001XT 00057, of the *Direcció General de Recerca* of the *Departament d'Universitats, Recerca i Societat de la Informació* of the *Generalitat de Catalunya*.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44 (2), 139–177.

Aitchison, J., 1984. Reducing the dimensionality of compositional data sets. Mathematical Geology 16 (6), 617–636.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK), 416 p.

Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, V. (Ed.), Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology. Vol. I, II and addendum. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 3–35.

Aitchison, J. and J. Bacon-Shone, 1999. Convex linear combinations of compositions. Biometrika 86 (2), 351-364.

Aitchison, J., 2002. Simplicial inference. In: Viana, M. A. G., Richards, D. S. P. (Eds.), Algebraic Methods in Statistics and Probability. Vol. 287 of Contemporary Mathematics Series. American Mathematical Society, Providence, Rhode Island (USA), 1–22.

Aitchison, J., and M. Greenacre, 2002. Biplots for compositional data. Applied Statistics 51 (4), 375–392.

Antich, N., A. Canals, A. Soler, D. Darbyshire, and B. Spiro, 2000. The isotope composition of dissolved strontium as tracer of pollution in the Llobregat River, northeast Spain. In A. Dassarges (Ed.), Tracers and Modelling in Hydrogeology, Proceedings of the TraM'2000 Conference, IAHS. 207–212.

Antich, N., A. Canals, A. Soler, D. Darbyshire, and B. Spiro, 2001. Strontium isotopes as tracers of natural and anthropic sources in Cardener River, Llobregat River Basin (Barcelona, Spain) (Los isótopos de estroncio como trazadores de fuentes naturales y antróopicas en las aguas del río Cardener, cuenca del río Llobregat). In A. Medina and J. Carrera (Eds.) Las caras del agua subterránea, Vol I, 413–420.

Barceló-Vidal, C., 1996. Compositional data mixtures (Mixturas de datos composicionales), unpublished PhD Thesis, Universitat Politècnica de Catalunya

Buccianti, A., V. Pawlowsky-Glahn, C. Barceló-Vidal, and E. Jarauta-Bragulat, 1999. Visualization and modeling of natural trends in ternary diagrams: a geochemical case study. In: Lippard, S. J., Næss, A., Sinding-Larsen, R. (Eds.), Proceedings of IAMG'99 — The fifth annual conference of the International Association for Mathematical Geology. Vol. I and II. Tapir, Trondheim (N), 139–144.

Daunis-i-Estadella, J., J.J. Egozcue, and V. Pawlowsky-Glahn, 2002. Least squares regression in the Simplex. In: Bayer, U., Burger, H., Skala, W. (Ed.), Proceedings of IAMG'02 — The eighth annual conference of the International Association for Mathematical Geology. Vol. I, Schriften der Alfred-Wegener-Stiftung, ISSN 0946-8979, Berlin (D), 411–416.

Eckart, C., and G. Young, 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, 2003. Isometric logratio transformations for compositional data. Mathematical Geology 35 (3), 249–300.

Embrechts, P., C Klüppelberg, and T. Mikosch, 1997. Modelling extremal values, Springer Verlag, Berlin.

Gabriel, K. R., 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (3), 453–467.

Martín-Fernández, J.A., 2001. Measures of difference and authomatic non-parametric classification of compositional data (Medidas de diferencia y clasificación automática no-paramétrica de datos composicionales), unpublished PhD Thesis, Universitat Politècnica de Catalunya

Otero, N. and A. Soler ,2002, Sulphur isotopes as tracers of the influence of potash mining in groundwater salinization in the Llobregat River Basin (NE Spain). Water Research (36), 3989–4000.

Penn, B.S., 2002. Using simulated annealing to obtain optimal linear end-member mixtures of hyperspectral data, Computers and Geosciences 28 (7), 809-817.

Renner, R.M., 1995. The construction of extreme compositions. Mathematical Geology 27 (4), 485-497.

Renner, R.M., G.P. Glasby, and P. Walter, 1997. Endmember analysis of metalliferous sediments from Galapagos Rift and East Pacific Rise between 2ºN and 42ºS. Applied Geochemistry 12, 383-395.

Soler, A., A. Canals, S. Goldstein, N. Otero, N. Antich, and J. Spangenberg, 2002. Sulphur and strontium isotope composition of the Llobregat River (NE Spain): Tracers of natural and anthropogenic chemicals in stream waters. Water, Air and Soil Pollution 136, 207–224.

Tolosana-Delgado, R.; N. Otero, A. Soler, V. Pawlowsky-Glahn and A. Canals. Relative vs absolute analysis of compositions: a comparative analysis in surface waters of a Mediterranean river. Water Research, submitted.

Weltje, J.G., 1997. End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem, Mathematical Geology 29, 503-549.

APPENDIX: Computation of factor log-contrasts

components	$\underline{G}$ matrix coordinates								
... in factor Φ_G	**horizontal**	**vertical**		**projection**	**centered**	**integers**	**ilr**	**other**	**ilr**
Na	-10.646	-9.880		-23.329	-13.948	-14	-0.412	6	0.466
K	-13.288	-9.827		-25.539	-16.158	-16	-0.471	6	0.466
Cl	-12.259	-11.453		-25.532	-16.151	-16	-0.471	6	0.466
Mg	3.912	-3.207		-7.452	1.929	2	0.059	-2	-0.155
SO_4	3.761	-1.308		-6.570	2.811	3	0.088	-2	-0.155
HCO_3	8.648	3.150		-0.061	9.320	9	0.265	-3	-0.233
H_2O	11.968	3.876		3.137	12.518	12	0.353	-4	-0.310
Ca	8.316	1.261		-1.346	8.035	8	0.235	-3	-0.233
Ba	11.775	2.540		2.263	11.644	12	0.353	-4	-0.310
factor Φ_G			**Σx**	-84.431	0.000	0	0.000	0	0.000
tail	8.500	3.500	**Σx^2**			1154	1.000	166	1.000
head	-13.000	-10.000							
... in factor Φ_A									
SO_4	3.761	-1.308		34.629	5.195	5	0.154	1	0.177
Sr	6.066	-3.698		37.940	8.506	8	0.246	1	0.177
NH_4	-22.963	20.753		0.000	-29.434	-29	-0.891	-5	-0.884
NO_3	5.847	-4.929		38.596	9.162	9	0.276	2	0.354
Mg	3.912	-3.207		36.005	6.571	7	0.215	1	0.177
factor Φ_A			**Σx**	147.170	0.000	0	0.000	0	0.000
tail	-23.000	20.700	**Σx^2**			1060	1	32	1
head	5.800	-5.000							

Coordinates of involved variables in the bi-plot (elements of the matrix $\underline{G}$) and computation steps of log-contrasts generating Φ_A and Φ_G factors. Two different solutions are proposed, labeled "integers" and "other"; the corresponding "ilr" column standardizes them to unit norm to allow us to compare them. These *ilr* columns are also the coefficients of computation of the *ilr* basis (Egozcue et al, 2003) linked to factors Φ_G and Φ_A.