Egozcue, J. J. y Pawlowsky-Glahn, V., 2011. Análisis composicional de datos en Ciencias Geoambientales. *Boletín Geológico y Minero,* 122 (4): 439-452
ISSN: 0366-0176

Análisis composicional de datos en Ciencias Geoambientales

J. J. Egozcue[1] y V. Pawlowsky-Glahn[2]

(1) Dep. Matemática Aplicada III, Universidad Politécnica de Cataluña, Jordi Girona 1-3, 08034-Barcelona, Spain.
juan.jose.egozcue@upc.edu
(2) Dep. Informática y Matemática Aplicada, Universidad de Girona, Campus Montilivi Edif. P4, 17071-Girona, Spain.
vera.pawlowsky@udg.edu

RESUMEN

Los datos composicionales (concentraciones) son frecuentes en geociencias. No tener en cuenta el carácter composicional puede llevar a conclusiones erróneas. La correlación espuria (K. Pearson, 1897) tiene consecuencias devastadoras. Partiendo de los trabajos de J. Aitchison en los años 80, surge una metodología libre de estos inconvenientes. La geometría del símplex permite representar composiciones en coordenadas ortogonales, pudiendo aplicarse métodos estadísticos usuales. Facilita la computación y el análisis. El uso de (log) cocientes evita la interpretación individual de concentraciones desatendiendo su carácter relativo. Se ilustra con un conjunto de análisis hidro-químicos.

Palabras clave: biplot, correlación espuria, dendrograma, geometría de Aitchison, símplex

Compositional data analysis in geo-environmental sciences

ABSTRACT

Compositional data (concentrations) are common in geosciences. Neglecting its character may lead to erroneous conclusions. Spurious correlation (K. Pearson, 1897) has disastrous consequences. On the basis of the pioneering work by J. Aitchison in the 1980s, a methodology free of these drawbacks is now available. The geometry of the simplex allows the representation of compositions using orthogonal co-ordinates, to which usual statistical methods can be applied, thus facilitating computation and analysis. The use of (log) ratios precludes the interpretation of single concentrations disregarding their relative character. A hydro-chemical data set is used to illustrate the point.

Key words: Aitchison geometry, biplot, dendrogram, simplex, spurious correlation

ABRIDGED ENGLISH VERSION

Introduction

Most data in the geosciences are compositional in character. They describe quantitatively the parts of a whole. The units are usually mg/l, %, ppm, molarity or molality. Compositional data appear in all branches of science and technology. One of the first examples where problems in compositional data analysis were detected was described by Pearson (1897). The study of compositional data in geology goes back to the 1960s. The problem was called "the closure problem" (Chayes, 1960). J. Aitchison (1982, 1986) introduced a solution with the analysis of log-ratios. In 2001 several advances were made (Pawlowsky-Glahn and Egozcue, 2001, Billheimer et al., *2001, Aitchison* et al., *2002). At present, the analysis of compositional data can be reduced to three steps: transformation of data into log-ratio co-ordinates; statistical analysis of these co-ordinates as real variables; and interpretation of obtained models in co-ordinates, or re-expressing results as compositions.*

Our aim here is to review the basics of compositional analysis as applied today. Problems posed by conventional analysis are illustrated by water analysis data of deep aquifers (Moeller et al., *2008).*

Compositional data analysis without caution

Data published by Moeller at al. (2008) illustrate the behaviour of traditional statistical methods. Deep groundwater salinization is studied using major elements (mg/l). The data taken into account are: extraction depth, h; temperature, T; total dissolved matter, TDS; and concentrations in mg/l of Na, K, Mg, Ca, Cl, SO_4 and HCO_3. The relationships between the concentrations and between the concentrations and the external variables h, T and TDS are analysed. A standard exploratory analysis of raw data is presented in Table 1 and Figure 1. Table 2 shows how raw correlations can change on account of different sub-compositions, thus visualising the spurious correlation phenomenon (Aitchison, 1986). This also affects correlations between concentrations and external variables. Figure 2 shows two biplots obtained from raw concentrations for two different sub-compositions where possible similarities between the two cases are fortuitous.

Compositional alternative

J. Aitchison (1982, 1986) put forward some principles for compositional analysis (Barceló-Vidal et al.*, 2001; Aitchison and Egozcue, 2005, Martín-Fernández* et al.*, 2003, Egozcue, 2009). The immediate consequences of the definition of compositional data are that they appear as vectors of positive components, and that only ratios of parts provide information. They lead to the formulation of the following principles:* scale invariance; Sub-compositional coherence, *including* sub-compositional dominance; symmetrized relative scale*; and* permutation invariance.

Geometric framework

The development of the concepts proposed by Aitchison (1986) has lead to the so called Aitchison geometry of the simplex *(Egozcue and Pawlowsky-Glahn, 2001). It is a Euclidean geometry and requires specific definitions of the operations and metrics.* Perturbation *is defined in (2) and* powering *in (3). These defined operations meet the requirements of the operations of a vector space and can be interpreted intuitively. Scale-invariant log-ratios that remove units of the parts involved are called log-contrasts and they have the form of (5). Any appropriate representation of a composition must be based on log-contrasts.*

The centred log-ratio, *clr, and its inverse were defined by Aitchison in 1986 (6-7). The clr representation of compositions consists of* D *coefficients (log-contrasts), although the dimension of the simplex is* D-1*. Perturbation and powering of compositions are transformed into the addition and multiplication of real vectors of* D *components (8). The representation of clr can be used to define a metric structure in* S^D*. The Aitchison inner product, norm and distance are defined in (9-11). These definitions configure a Euclidean structure on the simplex. Thus the usual tools in these spaces (orthonormal bases, orthonormal co-ordinates, orthogonal projections etc.) are available in the simplex.*

An orthonormal basis in S^D *is a set of unitary mutually orthogonal compositions. For a fixed orthonormal basis, the vector of co-ordinates is called an* isometric log-ratio transformation *(ilr) (Egozcue* et al.*, 2003) of the composition (12). The inverse ilr transformation (13) allows us to recover the original composition. The ilr transformation has properties of isometry (14-16).*

Orthonormal co-ordinates can be defined according to the problem to be solved. One such technique is based on a sequential binary partition (SBP) of the composition (Egozcue and Pawlowsky-Glahn, 2005, 2006b). Each partition, of a total of D-1*, results in an ilr co-ordinate and is called* balance*, the structure of which facilitates interpretation. Table 5 describes the process of SBP as an example. Each of these partitions results in a balance (17), a log-ratio of the geometric means of parts in each separated group.*

Centres and variability

Following the approach set out in Pawlowsky-Glahn and Egozcue (2001), the variability of a random composition ***X*** *compared to a composition* ***z*** *is* $\mathrm{Var}(\mathbf{X},\mathbf{z}) = \mathrm{E}[\mathrm{d}_a^2(\mathbf{X},\mathbf{z})]$*. The composition* ***z*** *that minimizes* $\mathrm{Var}(\mathbf{X},\mathbf{z})$ *is called the* centre *of* ***X****, and the minimum variability obtained is* total variance*. The centre, total variance and its components can be estimated in ilr co-ordinates (Pawlowsky-Glahn and Egozcue, 2001, 2002). The variability of a sample is analysed using the matrix of variances of all simple log-ratios (Aitchison, 1986). The compositional* biplot *is a graphical representation of variability (Aitchison and Greenacre, 2002), consisting of a principal-component analysis of the matrix of clr transformed data. The* biplot *simultaneously displays a projection of the data and of the clr centred variables. Figure 3 shows the* biplot *for the example data set. The main elements for interpretation of the* biplot *are the links between the rays. The link between two rays is roughly proportional to the variance of the simple log-ratio between the two parts. In addition, approximately perpendicular links indicate a low correlation between the corresponding simple log-ratios. Figure 3 shows that a large part of the variance is associated with the log-ratios of Cl and Na with* HCO_3*, and that the variance of ln (Na/Cl) is quite small. Moreover, it suggests that log-ratios containing* HCO_3 *correlate poorly with those that do not contain* HCO_3*. This is consistent with the hypothesis that some of the water samples examined contain significant amounts of sodium chloride in solution whilst others contain higher concentrations of bicarbonate. See also the compositional dendrogram (Figure 4), which allows us to display simultaneously the chosen balances, their means and variances (Egozcue and Pawlowsky-Glahn, 2006; Thió-Henestrosa* et al.*, 2008). Inspection of the dendrogram in Figure 4 confirms the impressions obtained from the compositional biplot.*

Prediction of a composition with an external variable

The data of Moeller et al. *(2008) suggest the question, "Does the Cl and Na content come from the dissolution of sodium chloride?" A compositional sample cannot tell anything about the origin of dissolved salts or processes that generated the observed concentrations (Aitchison, 1986; Aitchison and Egozcue, 2005). External variables need to be considered to reach conclusions that are not compositional. The advantage of ilr co-ordinates can be analysed by using standard statistical techniques. In the present case a regression model that predicts the composition from the TDS has been fitted. Compositional regression was introduced in Aitchison and Shen (1980), but having the representation in balances of Table 5, simple regressions of each balance on logit(TDS)=ln(TDS)/(10^6-TDS) can be established (Table 6).*

The regression results (Table 6 and Fig. 5) show that balance b_2 *has a significant regression with logit (TDS), whilst* b_3 *shows reduced variability (Fig. 4), indicating that the Cl/Na ratio is in stoichiometric equilibrium as regards the sodium chloride. This can be interpreted as showing that an increase in the dissolved mass is accompanied by an increase in mass of sodium chloride in solution but that the Cl/Na ratio is maintained.*

Conclusion

Geo-environmental analysis often deals with chemical compositions. The direct application of statistical methods designed for real multivariate data may be misleading or even meaningless. Methods based on the analysis of log-ratios are able to counter these difficulties in accordance with the principles of invariance of scale and sub-compositional coherence. The recognition of the simplex as a sample space of compositional data, and of its Euclidean structure (Aitchison geometry), allows the representation of compositions in real co-ordinates, to which usual methods of multivariate statistics for real data can be applied. This representation requires the analyst to abandon interpretation in terms of a single component but reasoning permanently on the ratios of parts: an effort that generally leads to greater methodological rigor.

Introducción

Gran parte de los datos que aparecen en geología, minería y ciencias ambientales tienen carácter composicional. Es decir, describen cuantitativamente las partes que forman un todo. Aparecen en forma de concentraciones, proporciones, frecuencias absolutas o relativas, con unidades como mg/l, %, partes por millón (ppm) de masa o de volumen, molaridad, molalidad, etc. Frecuentemente, el total no tiene interés especial. Por ejemplo, el tamaño de la muestra de una roca o el volumen de agua de un acuífero recogido para su análisis. Los datos composicionales aparecen en todas las ramas de la ciencia y la técnica y no sólo en las geociencias. Sin embargo, en geociencias, la frecuencia con que aparecen es enorme y ha provocado que sea en este ámbito en el que históricamente se les haya prestado mayor atención. Atención que ha sido motivada por los problemas que surgen al aplicar el análisis estadístico tradicional a datos de tipo composicional. Sin embargo, uno de los primeros ejemplos en que se detectaron los problemas mencionados procede de la morfología biológica y de uno de los fundadores de la estadística moderna: K. Pearson (1897). En geología el auge del estudio de los datos composicionales se remonta a los años 50 y 60 del siglo XX. En ese momento el problema se denomina de *datos clausurados (closed data)*. Entre los geólogos que estudiaron y advirtieron sobre los problemas del análisis estadístico destaca F. Chayes (1960). Pero no se llegó a una vía factible de análisis hasta los años 80, en que J. Aitchison (1982, 1986) introdujo el análisis de log-cocientes (*log-ratios*) junto con sus principios básicos, las técnicas y los modelos correspondientes. A pesar de las ventajas que ofrecían las técnicas de log-cocientes y las correspondientes transformaciones de los datos, éstas no tuvieron el éxito que cabía esperar y una buena parte de los científicos continuaron aplicando el análisis estadístico tradicional sin precaución alguna. A partir del año 2000 se producen diversos avances en los aspectos formales del análisis (Pawlowsky-Glahn y Egozcue 2001, Billheimer *et al.* 2001, Aitchison *et al.* 2002) que permiten una mayor sistematización de los métodos ya propuestos en los años 80. En la actualidad el análisis de datos composicionales puede reducirse a tres pasos: la transformación de los datos a coordenadas de tipo log-cociente; el análisis estadístico (tradicional) de dichas coordenadas como variables reales; y la interpretación de los modelos obtenidos en las propias coordenadas o volviendo a expresar los resultados en términos de composiciones. Estas técnicas *composicionales* aún no son de uso generalizado, pero se advierte un creciente interés en diversas ramas científicas, incluyendo las geo-ciencias.

El objetivo de la siguiente exposición es revisar los elementos básicos del análisis de datos composicionales tal como se aplican en la actualidad. Los problemas que se presentan en un análisis convencional se ilustran mediante unos datos de análisis de aguas de acuíferos profundos (Moeller *et al.,* 2008). A continuación se revisan los principios en los que se basa el análisis de datos composicionales. Las secciones siguientes se dedican a las bases de la geometría del símplex, considerado como espacio muestral de los datos composicionales y las consecuencias estadísticas que se derivan. Utilizando el ejemplo mencionado, se describen sucintamente las herramientas exploratorias que son propias del análisis composicional. Y se acaba con un ejemplo de regresión en que la variable respuesta es la composición geoquímica de las aguas analizadas.

Análisis de datos composicionales sin precaución

Para ilustrar el comportamiento anómalo de los métodos estadísticos tradicionales aplicados a datos composicionales se usan parte de los datos publicados por Moeller at al. (2008). Se consideran los elementos mayores (mg/l) del análisis de aguas subterráneas profundas para estudiar su salinización, supuestamente provocada por la presencia de paleosalmueras profundas en la cuenca sedimentaria del norte de Alemania. Los datos fueron extraídos de un conjunto de 28 pozos con profundidades que oscilan entre 50m y 2250m. Se ha suprimido un pozo porque faltan algunos datos. De los diversos datos aportados en Moeller *et al.* (2008), se consideran la profundidad de la extracción h, la temperatura T, el total de materia disuelta (TDS) y las concentraciones en mg/l de Na, K, Mg, Ca, Cl, SO_4, HCO_3. Aquí se propone como objetivo analizar las relaciones entre las concentraciones de los distintos elementos de la composición química observada, y entre ellas y las variables externas h, T, y TDS. Normalmente, un análisis exploratorio básico contiene la estima de las medias mediante un promedio aritmético para cada variable y el de la matriz de correlaciones entre todas las variables. Estos resultados se detallan en la Tabla 1. A simple vista destaca la correlación entre la concentración de Cl con la de Na (0.996) y también con la de Mg (0.808) que parecen comprensibles dado que se trata de un estudio sobre salinización. El analista puede concluir que, en la muestra, cuando se registra un aumento de la concentración de Cl, las de Na y Mg aumentan casi proporcionalmente. También destacan las altas correlaciones de TDS con algunas concentraciones como Na

	h	TDS	T	Na	K	Mg	Ca	Cl	SO_4	HCO_3
media	549.36	80830.75	21.11	28841.32	279.49	474.53	1202.54	48157.75	1424.86	352.32
mediana				9450.00	49.00	198.00	325.00	16950.00	725.00	372.00
m. geom.				7993.19	68.06	175.24	415.13	11529.21	364.80	321.20

	h	TDS	T	Na	K	Mg	Ca	Cl	SO_4	HCO_3
h	1.000									
TDS	0.605	1.000								
T	0.665	0.401	1.000							
Na	0.576	0.998	0.366	1.000						
K	0.079	0.621	0.097	0.609	1.000					
Mg	0.515	0.809	0.378	0.787	0.852	1.000				
Ca	0.863	0.652	0.717	0.613	0.201	0.539	1.000			
Cl	0.624	0.999	0.419	0.996	0.611	0.808	0.670	1.000		
SO_4	0.067	0.714	-0.009	0.726	0.747	0.703	0.149	0.693	1.000	
HCO_3	-0.583	-0.672	-0.602	-0.662	-0.253	-0.563	-0.564	-0.677	-0.402	1.000

Tabla 1: Medias muestrales y matriz de correlación de las variables consideradas. Las unidades originales de la profundidad h son m, la temperatura T en grados centígrados, TDS en mg/l; las demás concentraciones se dan en mg/l.
Table 1: Sample means and correlation matrix of the variables considered. The original units of depth, h, are in m, temperature, T, is in centigrade and TDS is in mg/l; other concentrations are in mg/l.

(0.998), Mg (0.809), Cl (0.999), o la única correlación negativa con HCO_3 (-0.672). Las correlaciones positivas podrían interpretarse como que los incrementos de masa disuelta van acompañados de un incremento proporcional de las concentraciones de Na, Mg y Cl. La correlación negativa con el ion bicarbonato también parece razonable si se interpreta que los aportes de masa disuelta proceden de diversos cloruros. Otras correlaciones resultan difíciles de interpretar, como la del bicarbonato y el Ca, que resulta ser -0.564.

La Tabla 2 muestra que, dependiendo de la normalización de las concentraciones de los iones, se obtienen resultados muy distintos. Incluso las correlaciones más elevadas y cercanas a 1 de la Tabla 1 pasan a ser correlaciones negativas. Y esto es así no solo para correlaciones entre concentraciones, sino también para correlaciones entre una concentración y variables externas como TDS o la profundidad. La conclusión es que las correlaciones que involucran variables composicionales son espurias y no tiene sentido interpretarlas (Aitchison, 1986).

Las medias de las concentraciones presentadas en la Tabla 1 también presentan problemas. Frecuentemente, las distribuciones muestrales de cada concentración suelen ser muy asimétricas y, como consecuencia, las medias y medianas pueden diferir sustancialmente. La Figura 1 presenta los diagramas de caja (*box-plot)* de las concentraciones; en ellos se

	Cl	Cl	Cl	HCO_3	K	TDS	TDS	h
	Na	Mg	Ca	Ca	Na	Na	Cl	Ca
O	0.996	0.808	0.670	-0.564	0.609	0.998	0.999	0.863
A	0.840	-0.564	-0.695	0.714	-0.323	0.424	0.376	-0.148
B	-0.359	-0.892	-0.852	***	0.517	-0.017	0.312	-0.194

Tabla 2. Correlaciones muestrales obtenidas entre algunos pares de variables en tres situaciones diferentes: O concentraciones originales en mg/l; A concentraciones en masa sobre el total de la subcomposición; B concentraciones en masa sobre el total de la subcomposición suprimiendo el ion bicarbonato.
Table 2. Sample correlations between some pairs of variables in three different situations: O: original concentrations in mg/l; A: concentrations in mass of the total sub-composition; B: concentrations in mass of the total sub-composition omitting the bicarbonate ion.

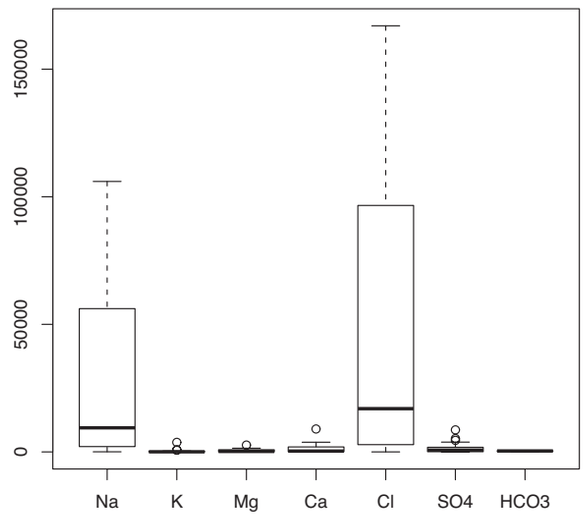

Figura 1. Diagramas de caja de las concentraciones de los elementos analizados.
Figure 1. Box-plot of the concentration of elements analysed.

aprecian las asimetrías mencionadas y que la diferencia de escala hace ilegible el resultado de algunas componentes. La cuestión es que la definición de media implica la aceptación de la métrica ordinaria en los números reales. Tomando como referencia algunos de los datos de concentración de Na, al calcular la media se acepta que la distancia entre 50 mg/l y 150 mg/l de Na es la misma que entre $106 \cdot 10^3$ y $106 \cdot 10^3 + 100$, cuando en el primer caso una de las concentraciones es triple de la otra y en el segundo caso ambas concentraciones pueden considerarse prácticamente idénticas. Si aceptamos que la escala de los datos es relativa, una primera aproximación a una media sería la media geométrica (esto equivale a tomar logaritmos, hacer la media aritmética de ellos, y entonces tomar exponencial del resultado). Si se comparan los valores de la media, la media geométrica y la mediana (Tabla 1) se observa una mejor concordancia entre las dos últimas. La conclusión es que para el cálculo de valores centrales es conveniente tener en cuenta la escala de los datos, que en el caso de concentraciones puede, en primera aproximación, considerarse relativa.

Para insistir en lo inadecuado del análisis de las concentraciones sin tomar precauciones de acuerdo a su carácter composicional, se presenta el análisis de componentes principales (PCA) de la composición hidroquímica de los datos en los casos O (datos originales en mg/l) y A (datos en tanto por uno en masa sobre el total de la subcomposición). Ya que en ambos casos la información utilizada es la misma, sería deseable que los resultados fueran, al menos, parecidos. Sin embargo, puesto que el PCA está íntimamente relacionado con la matriz de covarianzas de los datos, las posibles similitudes de ambos análisis pueden considerarse fortuitas. La Figura 2 representa los *biplots* para el caso O (85% variabilidad represen-

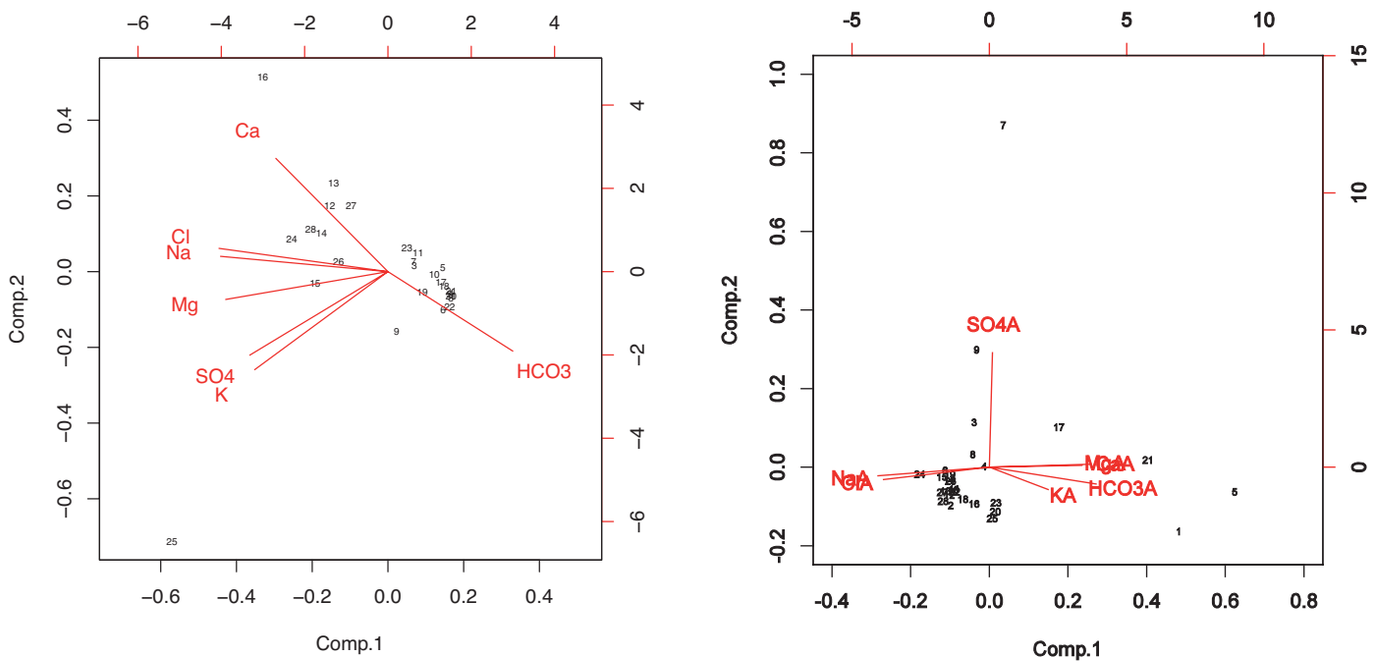

Figura 2. *Biplot* de las concentraciones utilizando la matriz de correlaciones. Izquierda: situación O, concentraciones en mg/l. Derecha: situación A, concentraciones en tanto por uno de TDS.
Figure 2. Biplot of concentrations using the correlation matrix. Left: situation O, concentrations in mg/l. Right: situation A, concentrations in parts per one of TDS.

tada) y A (75% variabilidad representada). En ambos casos se ha utilizado la matriz de correlación de las concentraciones estandarizadas; si se utiliza la matriz de correlación de las concentraciones sin estandarizar, las diferencias de varianza entre las variables oculta la estructura del *biplot*. En cualquier caso, ambos *biplots* son difícilmente comparables, de acuerdo con la diferente estructura de correlación (espuria) de los dos casos O y A.

La alternativa composicional

La experiencia en análisis como el presentado en la sección anterior ha sido acumulada durante un centenar de años. Basándose en ella J. Aitchison (1982, 1986) formuló unos principios a los que debe responder la caracterización y análisis de datos composicionales. Se han dado varias reformulaciones de esos principios (Barceló-Vidal *et al.* 2001; Aitchison y Egozcue 2005, Martín-Fernández *et al.* 2003, Egozcue 2009) que los han detallado de acuerdo con los desarrollos de la teoría. En primer lugar es necesaria una definición de datos composicionales que, esencialmente, coincide con la dada en las primeras líneas de la introducción: *los datos composicionales son datos que describen cuantitativamente las partes de un todo y aportan solo información relativa entre sus componentes*. De esta definición se extraen algunas consecuencias inmediatas. Los datos composicionales aparecen en forma de vectores de dos o más componentes positivas, aunque es frecuente que se suprima una de las componentes. Solo las razones o cocientes entre las componentes aportan información. Esto excluye los datos que contienen componentes intrínsecamente nulas, que no pueden aportar información relativa. Estas consideraciones llevan a la formulación de los siguientes principios.

***Invariancia por escala**: los vectores de componentes positivas proporcionales representan la misma composición. Dicho de otra forma, al multiplicar una composición por una constante, por ejemplo al pasar de tantos por uno a tantos por cien, obtenemos la misma composición y la información contenida en ella es totalmente equivalente.*

Según el principio de invariancia por escala, todos los vectores de *D* componentes positivas que son proporcionales son equivalentes y representan la misma composición. Es por tanto lógico elegir un representante de la clase de equivalencia para facilitar el trabajo e interpretación. La forma tradicional de elegir el representante es normalizar el vector para que sus componentes sumen una constante prefijada κ, que puede ser 1, 100, 1000, 10^6, o cualquier otra que pueda ser conveniente. Para simbolizar esta operación de elección de representante se utiliza el operador *clausura*. Si $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ es un vector de *D* componentes positivas se define su clausura por

$$C\mathbf{x} = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right) \qquad (1)$$

Las componentes del vector clausurado se denominan *partes*, referidas al total κ. El conjunto de vectores de *D* componentes positivas cuyas componentes suman la constante κ constituyen el símplex de *D* partes, que se denotará por S^D. Las composiciones equivalentes a **x** quedan entonces representadas por $C\mathbf{x}$.

composición O		**d. euclídea = 0.010**					
Na	**K**	**Mg**	**Ca**	**Cl**	**SO4**	**HCO3**	**otros**
1.580E-04	7.400E-06	1.400E-05	4.000E-05	6.300E-05	3.000E-06	5.030E-04	9.997E-01
3.300E-03	1.880E-05	5.000E-05	6.200E-05	5.100E-03	1.900E-05	4.940E-04	9.915E-01
subcomposición A		**d. euclídea = 0.777**					
Na	**K**	**Mg**	**Ca**	**Cl**	**SO4**	**HCO3**	
2.004E-01	9.386E-03	1.776E-02	5.074E-02	7.991E-02	3.805E-03	6.380E-01	
3.649E-01	2.079E-03	5.529E-03	6.856E-03	5.639E-01	2.101E-03	5.462E-02	

Tabla 3. Dos primeras composiciones del ejemplo expresadas en tanto por uno de masa. Situación original (O) considerando el resto hasta 1. Situación (A): subcomposición eliminando la componente otros. Se dan las distancias euclídeas entre las dos composiciones en ambas situaciones. Se viola la dominancia subcomposicional.

Table 3. Two initial compositions of the example in parts per one of mass. Original situation (O) including the remainder to 1. Situation (A): sub-composition after eliminating the component "others". The Euclidean distances between both compositions in both situations are given. The principle of sub-compositional dominance is violated.

Coherencia subcomposicional*: cuando se examina un subconjunto de las partes de una composición, una subcomposición, se requiere que los resultados del análisis no sean contradictorios con los obtenidos de la composición original. La coherencia puede resumirse en dos criterios: (A) el principio de invariancia por escala se aplica a cada una de las subcomposiciones posibles; (B) (dominancia subcomposicional) si se utiliza una distancia o divergencia para comparar composiciones, esa distancia o divergencia tiene que ser menor cuando se comparan las respectivas subcomposiciones.*

Este principio es más sutil que el de invariancia por escala, y es de importancia capital. En el ejemplo de la sección anterior, este principio aseguraría que el análisis de la composición completa de 7 componentes en mg/l más la octava, representando a *otros* elementos incluyendo el agua, (situación O), debe ser coherente con el análisis de la subcomposición que abarca exclusivamente las primeras 7 componentes y que se expresa en tanto por uno de masa sobre la masa total de esas 7 componentes (situación A). Esto debe ocurrir porque la información que da una subcomposición son las razones entre las partes que la constituyen, que son las mismas que en la composición original. Los ejemplos de matrices de correlaciones entre las diferentes concentraciones en las situaciones O, A, B, son ejemplos de violación del principio de coherencia subcomposicional (Tabla 2).

La dominancia subcomposicional exige que la forma de medir distancias en la composición y las subcomposiciones siga las reglas de una proyección: las distancias se reducen cuando se observa una proyección. Es lógico preguntarse si la distancia euclídea ordinaria entre vectores reales puede ser aplicada para medir distancias entre composiciones. De hecho se violan tanto el principio de invariancia por escala como la dominancia subcomposicional. Si dos vectores de componentes positivas se multiplican por una contante positiva *c*, entonces la distancia euclídea entre ellas queda multiplicada por *c* violando así el principio de invariancia por escala. Para ver que esa distancia no es subcomposicionalmente dominante podemos considerar los dos primeros datos del ejemplo de la sección anterior dados en la Tabla 3 expresados en tanto por uno (masa). En la situación original (O) se tiene en cuenta la componente *otros*. En la situación A se suprime la componente *otros* y se ha clausurado a tanto por uno. Se observa que la distancia euclídea en el caso O, de mayor dimensión, resulta menor que en el caso A, de menor dimensión.

Escala relativa simetrizada*: cada una de las partes de una composición tiene escala relativa.*

La importancia de la diferencia entre dos concentraciones depende del valor de dichas concentraciones. Como se ha mencionado antes, a propósito de las distancias, una concentración de 150mg/l de Na debe considerarse triple de otra de 50 mg/l, mientras que concentraciones $106 \cdot 10^3$ y $106 \cdot 10^3$ +100 mg/l apenas se diferencian entre si. Una forma tradicional de tratar escalas relativas es medir las diferencias mediante los respectivos logaritmos; por ejemplo las dos primeras concentraciones darían una diferencia ln(150)-ln(50)=ln(3). Sin embargo, esto debe ser válido para cada una de las partes de una composición. Imaginemos una composición de dos partes, por ejemplo, Na y todos los demás elementos incluyendo el agua. Parece lógico exigir que las diferencias en concentraciones apreciadas para el Na debieran ser valoradas de la misma forma por un analista que trabajara con la concentración complementaria. Es decir, supongamos que el litro de disolución tuviera masa 10^6 mg; el analista que observa el Na compara los valores en tanto por uno $50 \cdot 10^{-6}$ y $150 \cdot 10^{-6}$; en cambio el que trabaja con la componente *otros* observa los valores (en tanto por uno) $(10^6\text{-}50) \cdot 10^{-6}$ y $(10^6\text{-}150) \cdot 10^{-6}$. Las comparaciones que realicen ambos analistas debiera conducir a la misma conclusión, pero como se muestra en la Tabla 4 no es este el caso, ni siquiera tomando logaritmos. Este tipo de simetría entre las comparaciones de una parte y de su complementaria exige transformaciones de escala simetrizadas. La transformación *logit* o *logística* es el prototipo de ellas; consiste en el logaritmo del cociente de una parte con su com-

concentr. mg/l		absol.	log	logit
50	150	100	1.0986	1.0987
106000	106100	100	0.0009	0.0011
999950	999850	-100	-0.0001	-1.0987

Tabla 4. Comparación de dos concentraciones en escala absoluta, relativa (logarítmica) y relativa simétrica (logit). Las diferencias absolutas son simétricas pero no relativas. La escala logarítmica refleja el carácter relativo pero no es simétrica respecto al complementario. La escala logit refleja tanto el carácter relativo como la simetría.

Table 4. Comparison of two concentrations in absolute scale, relative (logarithmic) and relative symmetric (logit). The absolute differences are symmetric but not relative. The logarithmic scale reflects the relative character, but is not symmetric with respect to the complementary. The logit scale reflects the relative character as well as the symmetry.

plementario. Por ejemplo la concentración de 50 ppm se transforma en ln(50/(10^6-50)). Debe notarse que la transformación logit es un log-cociente y, por tanto, invariante por cambio de escala.

Invariancia por permutación: *las conclusiones de un análisis composicional no deben depender de la ordenación de las partes*.

Esto parece obvio en muchos casos. En las composiciones geoquímicas es frecuente establecer un orden alfabético de las partes, prescindiendo de sus características. Sin embargo, en algunos casos las partes pueden considerarse ordenadas. Un ejemplo típico es la granulometría de un sedimento: las partículas se clasifican, después de un tamizado, por categorías de tamaños. Si se aplica un análisis composicional, la información debida a la ordenación de las clases no participa en él.

Marco geométrico

Para satisfacer los principios descritos en la sección anterior se requiere una geometría del símplex de *D* partes, S^D. El desarrollo de los conceptos propuestos por Aitchison (1986) ha llevado a la llamada *geometría de Aitchison* del símplex (Pawlowsky-Glahn y Egozcue, 2001) que, siendo una geometría de tipo euclídeo, requiere definiciones específicas de las operaciones y la métrica cuya apariencia es peculiar.

Consideramos las composiciones **x**, **y** en S^D, cuyas componentes se denotan por x_i, y_i, respectivamente. La *perturbación* de **x** con **y** se define como la composición

$$\mathbf{x} \oplus \mathbf{y} = C(x_1 y_1, x_2 y_2, ..., x_D y_D) \qquad (2)$$

y la *potenciación* de **x** con el número real α es la composición

$$\alpha \otimes \mathbf{x} = C(x_1^\alpha, x_2^\alpha, ..., x_D^\alpha) \qquad (3)$$

Se puede comprobar que si $\mathbf{n}=C(1,1,\ldots,1)$ entonces $\mathbf{x} \oplus \mathbf{n}=\mathbf{x}$, es decir, que una composición con todas las partes iguales es el elemento neutro de la perturbación. Estas operaciones definidas en S^D cumplen los requisitos de las operaciones de un espacio vectorial. Pero el principal mérito de la perturbación es que, además de atender a los principios del análisis composicional, tiene interpretación en el campo en que se trabaja. En el ejemplo de la composición hidroquímica de la sección segunda, es fácil concebir un tratamiento de las aguas, representadas por **x**, cuyo resultado sean las concentraciones iniciales multiplicadas cada una por una constante positiva. Esas constantes que caracterizan al tratamiento, agrupadas en un vector **y**, permiten calcular el resultado del tratamiento, que es $\mathbf{x} \oplus \mathbf{y}$. Debe notarse que no importa que las composiciones **x** (concentraciones iniciales) o **y** (coeficientes de transferencia) se presenten en forma clausurada porque $C\mathbf{x} \oplus C\mathbf{y} = \mathbf{x} \oplus \mathbf{y}$. La potenciación puede interpretarse como una aplicación reiterada del tratamiento o filtro. Imaginemos un tratamiento del agua mediante un filtro en disposición lineal cuyos coeficientes por unidad de longitud se representan por **y**. Entonces un filtro de α unidades de longitud tendría como coeficientes de transferencia $\alpha \otimes \mathbf{y}$ y el resultado del tratamiento sería $\mathbf{z} = \mathbf{x} \oplus (\alpha \otimes \mathbf{y})$. Si, conociendo el resultado **z,** nos preguntamos por la composición del agua de entrada, obtenemos $\mathbf{x} = \mathbf{z} \oplus (-\alpha \otimes \mathbf{y})$, que da sentido a la potenciación con un coeficiente negativo. Una forma natural de escribir la misma relación es utilizar la perturbación negativa definiendo el signo θ, con lo que la última expresión queda $\mathbf{x} = \mathbf{z}\ \theta\ (\alpha \otimes \mathbf{y})$.

El cambio de unidades de algunas o todas de las partes también puede considerase como una perturbación. Si **x** son las concentraciones iniciales en mg/l e **y** son los respectivos pesos moleculares de las partes, $\mathbf{x} \oplus \mathbf{y}$ será proporcional a la composición molal.

La invariancia por escala de las composiciones lleva de forma natural a utilizar cocientes entre las partes y, supuesta la escala relativa de los cocientes, a considerar sus logaritmos. Un primer intento de representar composiciones en forma de log-cocientes es la *transformación aditivo-logística* alr (Aitchison 1986). Si **x** es una composición de S^D entonces se define

$$\mathrm{alr}(\mathbf{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D}\right) \qquad (4)$$

donde el logaritmo ln se aplica a cada una de las componentes, de forma que la componente i-ésima es el log-cociente $\mathrm{alr}_i(\mathbf{x})=\ln(x_i/x_D)$. Se observa que el cociente elimina las constantes de clausura o unidades que puedan multiplicar a las partes y el logaritmo se hace cargo de la escala relativa. Esta transformación adolece de la falta de simetría al elegir una de las partes, en este caso la última, como denominador común, violando el principio de invariancia por permutación de las partes. La virtud de los log-cocientes simples, como los que intervienen en alr(**x**), es la eliminación de las unidades. Esta propiedad puede extenderse a log-cocientes más complejos, llamados log-contrates, como

$$\ln\left(\prod_{i=1}^{D} x_i^{\alpha_i}\right) = \sum_{i=1}^{D} \alpha_i \ln(x_i), \quad \sum_{i=1}^{D} \alpha_i = 0 \tag{5}$$

donde la condición de suma nula de los coeficientes reales α_i es necesaria para mantener la invariancia por escala. Es decir, una representación apropiada de las composiciones debe basarse en expresiones de tipo log-contraste.

Para superar la asimetría de la representación alr se definió la *representación log-cociente centrada* clr (Aitchison 1986),

$$\mathbf{v} = \mathrm{clr}(\mathbf{x}) = \ln\left(\frac{x_1}{g_{\mathrm{m}}(x)}, \frac{x_2}{g_{\mathrm{m}}(x)}, ..., \frac{x_D}{g_{\mathrm{m}}(x)}\right), \qquad g_m(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \tag{6}$$

donde las D componentes $\mathrm{clr}_i(\mathbf{x})=\ln(x_i/g_m(\mathbf{x}))$ son log-contrastes, es decir, los coeficientes suman cero (ver (5)). Conociendo clr(**x**) puede reconstruirse **x** con la transformación inversa

$$\mathbf{x} = \mathrm{clr}^{-1}(\mathbf{v}) = C\exp(\mathbf{v}), \tag{7}$$

donde la función exponencial se aplica a cada una de las componentes de **v** =clr(**x**). La representación clr de composiciones adolece de un problema: la dimensión del símplex como espacio vectorial es *D-1*, mientras que el número de componentes de clr(**x**) es *D*. Puede comprobarse que la perturbación y potenciación de composiciones equivale a la suma y multiplicación de vectores reales de D componentes:

$$\mathrm{clr}((\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{y})) = \alpha \cdot \mathrm{clr}(\mathbf{x}) + \beta \cdot \mathrm{clr}(\mathbf{y}), \tag{8}$$

para cualquier par de coeficientes reales α y β.

La representación clr puede usarse para definir una estructura métrica en S^D. El producto escalar, norma y distancia de Aitchison en S^D se definen por

$$<\mathbf{x},\mathbf{y}>_a = <\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})> \tag{9}$$

$$\|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\|, \quad \mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})), \tag{10}$$

donde $<\cdot,\cdot>$, $\|\cdot\|$, $\mathrm{d}(\cdot,\cdot)$, denotan el producto escalar, la norma y la distancia euclídea ordinarias. Por ejemplo,

$$\mathrm{d}_a(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{D} (\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\mathbf{y}))^2}\ . \tag{11}$$

Estas definiciones constituyen la métrica de Aitchison del símplex. El producto escalar, la norma y la distancia de Aitchison obedecen a los principios de análisis composicional expuestos en la sección anterior y se convierten así en instrumentos de análisis libres de incoherencias. Pero además dan una estructura euclídea al símplex. Esto sugiere utilizar los instrumentos habituales en esos espacios: bases ortonormales, representación por coordenadas (ortonormales), proyecciones ortogonales, medidas de ángulos, definición de elipses, etc. Para dar este paso es conveniente disponer de algún método para construir bases ortonormales y las correspondientes coordenadas.

Una base ortonormal en S^D es un conjunto de composiciones $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ tal que $<\mathbf{e}_i, \mathbf{e}_j>_a=0$ cuando $i\neq j$ y $\|\mathbf{e}_i\|_a=1$. Fijada la base, las coordenadas de una composición son

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = (<\mathbf{x},\mathbf{e}_1>_a, <\mathbf{x},\mathbf{e}_2>_a, ..., <\mathbf{x},\mathbf{e}_{D-1}>_a) \tag{12}$$

de las cuales puede recuperarse la composición,

$$\mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{x}^*) = \bigoplus_{j=1}^{D-1} x_j^* \otimes \mathbf{e}_j\ . \tag{13}$$

La construcción de coordenadas ortonormales se ha llamado *transformación log-cociente isométrica,* ilr (Egozcue *et al.* 2003), porque las coordenadas $x_j^*=\mathrm{ilr}_j(\mathbf{x})$ tienen forma de log-contrastes y tienen propiedades de isometría:

$$\mathrm{ilr}((\alpha \otimes \mathbf{x}) \oplus (\beta \otimes \mathbf{y})) = \alpha \cdot \mathrm{ilr}(\mathbf{x}) + \beta \cdot \mathrm{ilr}(\mathbf{y}), \tag{14}$$

$$\|\mathbf{x}\|_a = \|\mathrm{ilr}(\mathbf{x})\|, \ \mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}(\mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y})), \tag{16}$$

paralelas a las dadas en las ecuaciones (9) y (10) para clr. La diferencia con aquéllas expresiones es que ahora el producto escalar, la norma y distancia de vectores de coordenadas ilr corresponden al espacio real de dimensión *D*-1, que es la dimensión del símplex S^D.

Como en cualquier otro espacio euclídeo, también en el símplex existen infinitas bases ortonormales. Una manera simple de representar composiciones en coordenadas ilr es realizar la descomposición en valores singulares de la matriz de la transformación clr de una muestra, operación que se realiza para presentar los *biplots* como se describe en la sección quinta. Pero

bal.	Na	K	Mg	Ca	Cl	SO_4	HCO_3	*r*	*s*
1	+1	-1	-1	-1	+1	-1	+1	3	4
2	+1	0	0	0	+1	0	-1	2	1
3	+1	0	0	0	-1	0	0	1	1
4	0	-1	-1	-1	0	+1	0	1	3
5	0	+1	-1	-1	0	0	0	1	2
6	0	0	-1	+1	0	0	0	1	1

Tabla 5. Código de signos de la SBP utilizada en el análisis de los datos. En cada fila las partes en un grupo se marcan con +1 y las del grupo alternativo con -1. El 0 indica que la parte correspondiente no participa en la partición. El número de signos + es *r* y *s* el de signos - en cada paso de partición.

Table 5. Sign-code of the SBP used to analyse the data. In each row parts in one group are marked with +1 and those of the alternative group with a -1. The 0 indicates that the corresponding part is not involved in the partition step. The number of + signs is r and the number of – signs is s in each partition step.

las coordenadas resultantes pueden ser difíciles de interpretar, dificultad frecuente en log-contrastes como (5). Por tanto es conveniente construir log-contrastes de interpretación adecuada al problema de que se trata. En particular, se trata de construir coordenadas ortonormales definidas por el analista de acuerdo con el problema que trata de resolver. Una de estas técnicas se basa en una partición secuencial binaria (denotada SBP por sus siglas en inglés, *Sequential Binary Partition*) de las partes de la composición (Egozcue y Pawlowsky-Glahn 2005, 2006b). Cada partición, de un total de *D*-1, da lugar a una coordenada ilr, ahora llamada *balance*, cuya estructura facilita la interpretación. En la Tabla 5 se describe el proceso de SBP. En la primera partición se ha agrupado Na, Cl y HCO_3 (signo +) contra SO_4, Mg, Ca, K (signo -). En la segunda partición, se separan Na, Cl de HCO_3. Y así sucesivamente. Cada una de estas particiones da lugar a un balance de la forma

$$b_j = \sqrt{\frac{rs}{r+s}} \ln \frac{g_{\mathrm{m}}(x_+)}{g_{\mathrm{m}}(x_-)} \ , \qquad (17)$$

donde $g_m(x_+)$ y $g_m(x_-)$ son las medias geométricas de las partes indicadas con signo + y con signo - en la partición j-ésima; y *r*, *s* son el número de partes con los signos + y – respectivamente. El nombre de balance proviene del hecho que es un log-contraste entre las medias geométricas de los grupos de partes que se han separado, por lo que pueden ser interpretados con más facilidad que otras coordenadas ilr de expresión más compleja. En la SBP de la Tabla 5 parecía evidente que Cl y Na debían mantenerse juntos hasta que finalmente se alcanzase el balance $b_3=2^{-1/2}$ ln(Na/Cl). Pero pueden existir dudas acerca de asociar estos elementos con HCO_3 o con HCO_3 y Ca. Se ha optado por la SBP de la Tabla 5 después de realizar un análisis exploratorio de los datos (sección quinta). Existen otras opciones de SBP, todas ellas desembocando en balances ortonormales: por ejemplo, la separación inicial de aniones y cationes puede ser útil cuando no se sospecha del origen de las sales disueltas.

Centros y variabilidad

La estadística trata de sintetizar la información en una muestra. La media y varianza-covarianza son los descriptores más frecuentes en escenarios multivariantes. Cuando se trata con datos composicionales no puede prescindirse de la geometría de su espacio muestral S^D y, en particular, de la distancia de Aitchison en S^D. Siguiendo el planteamiento en Pawlowsky-Glahn y Egozcue (2001), se define la variabilidad de una composición aleatoria **X** en S^D respecto a una composición **z** como $\mathrm{Var}(\mathbf{X},\mathbf{z}) = \mathrm{E}[\mathrm{d}_a^2(\mathbf{X},\mathbf{z})]$. La composición **z** que minimiza $\mathrm{Var}(\mathbf{X},\mathbf{z})$ se llama *centro* de **X**, y la variabilidad mínima obtenida *varianza total*. El centro es una composición que puede expresarse $\mathrm{Cen}[\mathbf{X}] = \mathrm{ilr}^{-1}\,(\mathrm{E}[\mathrm{ilr}(\mathbf{X})] = C\,\exp(\mathrm{E}[\ln \mathbf{X}]$, donde la última expresión también se ha dado como definición de centro (Aitchison, 1997). El centro de una composición aleatoria debe considerarse como el valor esperado o media de una composición. La varianza total puede descomponerse, al menos, de tres formas:

$$\mathrm{totVar}[\mathbf{X}] = \frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D} \mathrm{Var}\left[\ln\frac{X_i}{X_j}\right] = \sum_{i=1}^{D}\mathrm{Var}[\mathrm{clr}_i(\mathbf{X})] = \sum_{j=1}^{D-1}\mathrm{Var}[\mathrm{ilr}_j(\mathbf{X})]. \qquad (18)$$

La primera se basa en el aporte de varianza por cada uno de los log-cocientes simples posibles y, por ello, relativamente fácil de interpretar individualmente. Globalmente, la descomposición en componentes ilr resulta más interpretable, especialmente cuando se trata de balances, debido a la ortogonalidad de las coordenadas.

La estimación del centro, de la varianza total y sus componentes, puede hacerse en coordenadas ilr, y las propiedades de estos estimadores corresponden a las de los estimadores de medias y varianzas reales (Pawlowsky-Glahn y Egozcue (2001, 2002)).

El análisis de la variabilidad de una muestra se realiza utilizando la matriz de varianzas de todos los log-cocientes simples, llamada por Aitchison *variation matrix* o matriz de variaciones. Una representación gráfica aproximada se obtiene con el *biplot* composicional (Aitchison y Greenacre, 2002). Consiste en un análisis de componentes principales de la matriz de datos transformados clr, convenientemente centrados y usando la descomposición de la matriz de covarianzas. El *biplot* muestra simultáneamente una proyección de los datos y de las variables clr centradas. La Figura 3 lo muestra para los datos del ejemplo. El principal elemento interpretativo del *biplot* son los enlaces entre los rayos.

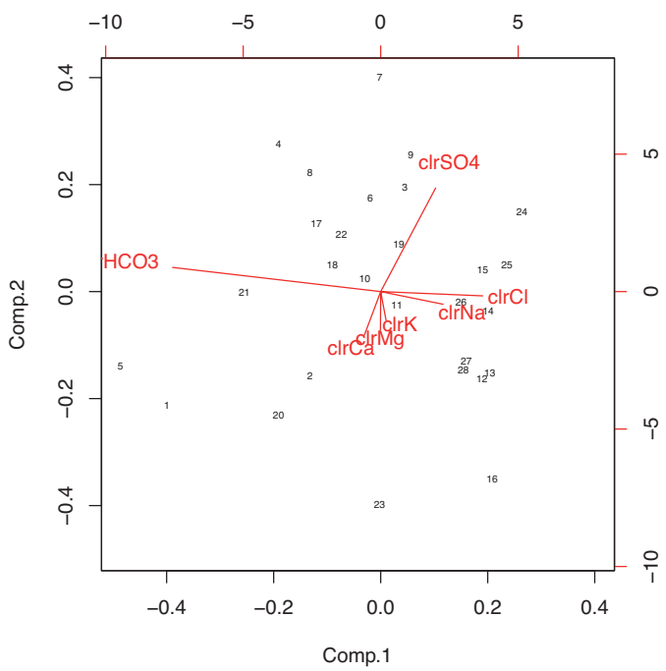

Figura 3. CoDa-biplot de los datos del ejemplo. Representa el 82.45% de la varianza de los datos.
Figure 3. CoDa-biplot of the data in the example, representing 82.45% of the variance in the data set.

Cada rayo representa una variable clr y su longitud la varianza asociada explicada en la proyección. El enlace entre dos rayos es aproximadamente proporcional a la varianza del log-cociente simple entre las dos partes. Además, enlaces aproximadamente perpendiculares indican una baja correlación entre los correspondientes log-cocientes simples. En la Figura 3 se aprecia que la mayor parte de varianza está asociada a los log-cocientes de Cl y Na sobre HCO_3, y que a su vez la varianza de ln(Na/Cl) debe ser pequeña. Por otra parte se sugiere que los log-cocientes que contienen HCO_3 están poco correlacionados con los que no contienen HCO_3. Esto concuerda, que no demuestra, con la hipótesis de que algunas de las muestras de aguas examinadas han disuelto cantidades apreciables de cloruro sódico, mientras otras contienen mayores concentraciones de bicarbonato (corresponden a las más superficiales). Estas consideraciones son las que han ayudado a tomar la decisión de elegir los balances con la SBP presentada en la Tabla 5. Cuando se ha definido una base de balances como la mencionada, el dendrograma composicional (Figura 4) permite visualizar simultáneamente la base de balances elegida, las medias de los balances, su varianza y la descomposición de la varianza total (Egozcue y Pawlowsky-Glahn, 2006; Thió-Henestrosa *et al.*, 2008). La inspección del dendrograma de la Figura 4 confirma las impresiones obtenidas del biplot composicional, al mismo tiempo que permite visualizar los valores medios de los balances representados por los puntos de enlace de las líneas horizontales en las verticales. La comparación de estas medias con los diagramas de caja permite una evaluación de la simetría de las distribuciones muestrales de los balances.

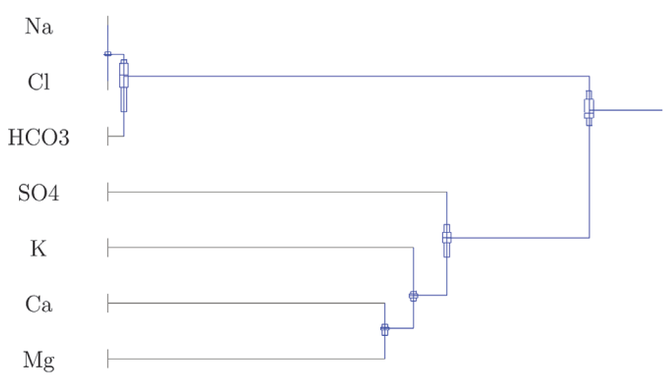

Figura 4. Dendrograma composicional de los datos del ejemplo. El dendrograma representa la SBP utilizada. Las longitudes de las líneas horizontales son proporcionales a la descomposición de la varianza por balances. Las líneas verticales tienen una escala común en el intervalo (-6,6) y muestran el diagrama de caja de la muestra en esa escala.
Figure 4. Compositional dendrogram of the data in the example. The dendrogram represents the SBP used. The lengths of the horizontal lines are proportional to the decomposition of the variance by balances. The vertical lines have a common scale in the interval (-6,6) and show the box-plot of the sample in that scale.

Predicción de una composición con una variable externa

Los datos seleccionados de Moeller *et al.* (2008) en el ejemplo de la sección 2 sugieren la pregunta: *el contenido en Cl y Na proviene de la disolución de cloruro sódico?* La muestra composicional solo describe las relaciones entre las diferentes partes y, por si misma, no puede informar acerca del origen de las sales disueltas o de los procesos que han llevado a las concentraciones observadas (Aitchison, 1986; Aitchison y Egozcue, 2005). Es preciso acudir a variables externas a la composición para llegar a conclusiones no composicionales. La ventaja de la representación en coordenadas de las composiciones es que éstas son variables reales y pueden tratarse, por tanto, mediante técnicas estadísticas estándar. En el caso que nos ocupa puede ajustarse un modelo de regresión que prediga la composición a partir del TDS (sólido disuelto total). La regresión con respuesta composicional fue propuesta en Aitchison y Shen (1980), pero disponiendo de la representación en balances de la Tabla 5 podemos establecer regresiones simples de cada uno de los balances sobre la transformación logit del TDS (ppm), ln(TDS)/(10^6-TDS) (Tabla 6). Esta transformación del TDS está de acuerdo con su carácter composicional (sólido, líquido).

	balance	logit(TDS)
b_1	{Na,Cl,HCO_3} vs. other	n.s.
b_2	{Na,Cl} vs. {HCO_3}	***
b_3	{Na} vs. {Cl}	**
b_4	{SO_4} vs. {K,Ca,Mg}	*
b_5	{K} vs. {Ca,Mg}	*
b_6	{Ca} vs. {Mg}	*

Tabla 6. Regresiones de los balances sobre logit de TDS. El balance b_2 tiene regresión claramente significativa (***, test F p-valor <0.001), y algo menos con el balance b_3 (**, test F p-valor <0.01); la regresión de los demás balances es menos significativa (*, <0.05), o no significativa (n.s.).
*Table 6. Regression of the balances on logit of TDS. Balance b_2 has a clearly significant regression (***, p-value of the F test <0.001), balance b_3 less so (**, p-value of the F test <0.01); for the other balances the regression is even less significant (*, <0.05) or not significant at all (n.s.).*

Los resultados de la regresión (Tabla 6, Fig. 5) muestran que el balance b_2 tiene una regresión significativa

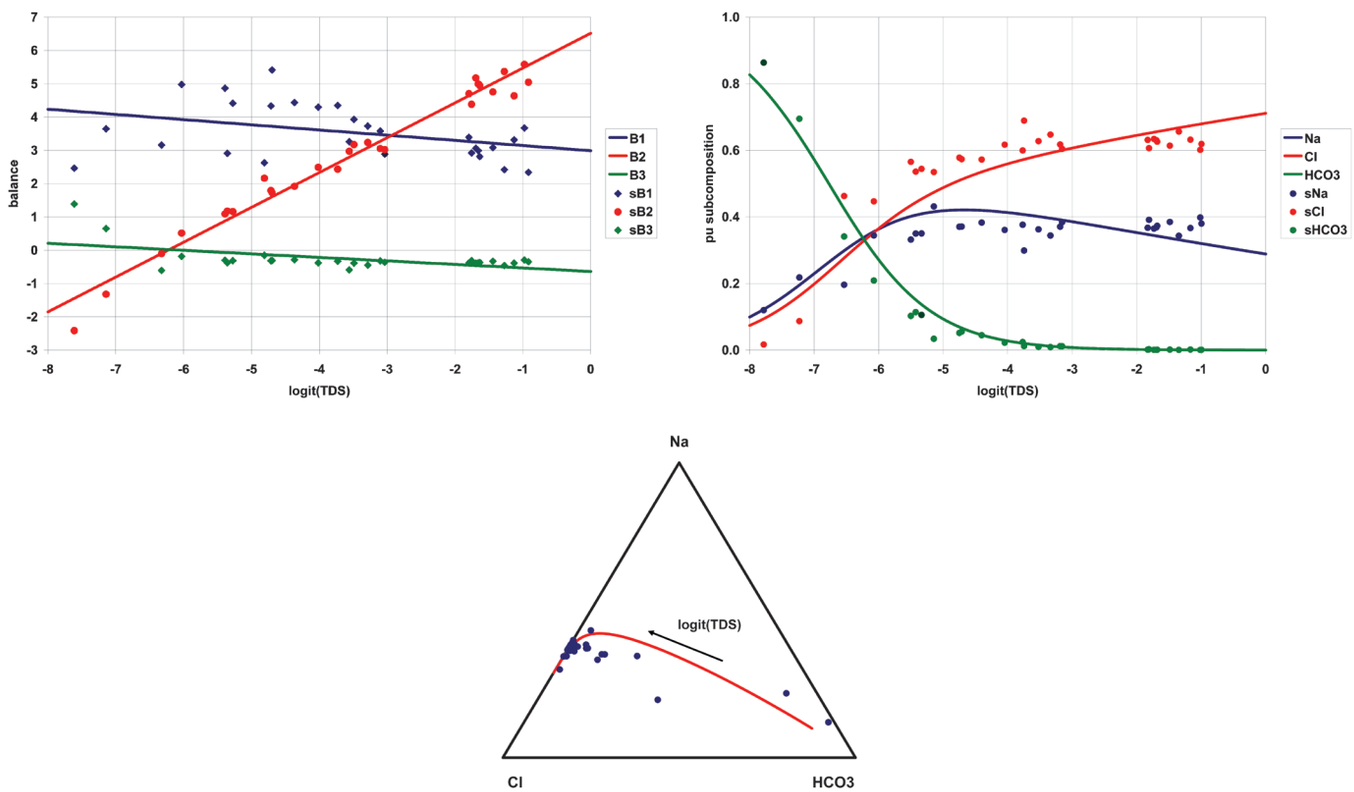

Figura 5. Izquierda: ajustes de b_1, b_2, b_3 a logit(TDS). El modelo de b_1 no es significativo. Derecha: los mismos modelos representados por partes en tanto por uno de la subcomposición Na, Cl, HCO_3. Inferior: trayectoria del modelo en el diagrama ternario de la subcomposición Na, Cl, HCO_3.
Figure 5. Left: adjustments of b_1, b_2, b_3 to logit(TDS). The model for b_1 is not significant. Right: the same models represented as parts per one of the sub-composition Na, Cl, HCO_3. Below: trajectory of the model in the ternary diagram of the sub-composition Na, Cl, HCO_3.

con logit(TDS); mientras que b_3 tiene una variabilidad reducida (Fig. 4), que indica que la relación Na sobre Cl sugiere el equilibrio estequiométrico del cloruro sódico. Esto puede interpretarse como que aumentos de masa disuelta vienen acompañados de un aumento de la masa de cloruro sódico disuelto.

Conclusión

Los análisis geoambientales tratan frecuentemente con composiciones químicas en sus diferentes versiones y unidades. Pero todos ellos comparten el carácter composicional. La aplicación imprudente de los métodos estadísticos diseñados para datos multivariantes reales puede llevar a conclusiones erróneas o sin sentido. Los métodos basados en el análisis de log-cocientes son capaces de afrontar las dificultades del análisis composicional de acuerdo con los principios básicos de invariancia por escala y coherencia subcomposicional. La concepción del símplex, como espacio muestral de datos composicionales, y su estructura euclídea (geometría de Aitchison) permiten representar las composiciones en coordenadas reales a las que se aplican los métodos de la estadística para datos reales multivariantes. Esta representación obliga al analista a abandonar la interpretación en términos de una sola componente y razonar permanentemente sobre los cocientes entre partes (o sus logaritmos): un esfuerzo que generalmente desembocará en un mayor rigor metodológico.

Agradecimientos

Esta investigación se ha desarrollado bajo los proyectos CoDA-RSS, Ref.: MTM2009-13272 y Ref.: *Ingenio Mathematica (i-MATH)* No. CSD2006-00032 (*Consolider - Ingenio 2010*) del Ministerio de Ciencia y Tecnología (España) y, también, bajo el proyecto Ref: 2009SGR424 de la *Agència de Gestió d'Ajuts Universitaris i de Recerca* de la *Generalitat de Catalunya* (España).

Referencias

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. 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). 416 p.

Aitchison, J. 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97 – The third annual conference of the International Association for Mathematical Geology,* Volume I, II and addendum, pp. 3-35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

Aitchison, J., Barceló-Vidal, C., Egozcue, J.J. and Pawlowsky-Glahn, V. 2002. A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In U. Bayer, H. Burger, and W. Skala (Eds.), *Proceedings of IAMG'02 - The eighth annual conference of the International Association for Mathematical Geology,* Volume I and II, pp. 387-392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Aitchison, J. and Egozcue, J.J. 2005. Compositional data analysis: where are we and where should we be heading? *Mathematical Geology,* 37(7), 829-850.

Aitchison, J. and Greenacre, M. 2002. Biplots for compositional data. *Journal of the Royal Statistical Society,* Series C (Applied Statistics), 51(4), 375-392.

Aitchison, J. and Shen, S.M. 1980. Logistic-normal distributions. Some properties and uses. *Biometrika,* 67(2), 261-272.

Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. 2001. Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 - The sixth annual conference of the International Association for Mathematical Geology,* pp. 20 p. CD-ROM.

Billheimer, D., Guttorp, P. and Fagan, W.F. 2001. Statistical interpretation of species composition. *Journal of the American Statistical Association,* 96(456), 1205-1214.

Chayes, F. 1960. On correlation between variables of constant sum. *Journal of Geophysical Research,* 65(12), 4185-4193.

Egozcue, J. and Pawlowsky-Glahn, V. 2006a. Exploring compositional data with the coda-dendrogram. In E. Pirard (Ed.), *Proceedings of IAMG'06 - The eleventh annual conference of the International Association for Mathematical Geology.*

Egozcue, J. and Pawlowsky-Glahn, V. 2006b. *Simplicial geometry for compositional data.* Special Publications Volume 264. Geological Society, London.

Egozcue, J.J. 2009. Reply to "On the Harker variation diagrams; ..." by J. A. Cortés. *Mathematical Geosciences*, 41(7), 829-834.

Egozcue, J.J. and Pawlowsky-Glahn, V. 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology,* 37(7), 795-828.

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

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35(3), 253-278.

Moeller, P., Weise, S.M., Tesmer, M., Dulsky, P., Pekdeger, A., Bayer, U. and Magri, F. 2008. Salinization of groundwater in the North German Basin: results from conjoint investi-

gation of major, trace element and multi-isotope distribution. *Int. J. Earth Sci.* 97, 1057-1073.

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

Pawlowsky-Glahn, V. and Egozcue, J.J. 2002. BLU estimators and compositional data. *Mathematical Geology* 34(3), 259-274.

Pearson, K. 1897. Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London,* LX, 489-502.

Thió-Henestrosa, S., Egozcue, J.J., Pawlowsky-Glahn, V., Kovàcs, L.O. and Kovàcs, G.P. 2008. Balance-dendrogram. A new routine of CoDaPack. *Computers and Geosciences,* 34, 1682-1696.

Recibido: diciembre 2010
Revisado: marzo 2011
Aceptado: julio 2011
Publicado: octubre 2011