Martín-Fernández, J.A., Palarea-Albaladejo, J. y Barceló-Vidal, C., 2011. Técnicas composicionales para concentraciones geoquímicas por debajo del límite de detección. *Boletín Geológico y Minero,* 122 (4): 459-468
ISSN: 0366-0176

Técnicas composicionales para concentraciones geoquímicas por debajo del límite de detección

J. A. Martín-Fernández[(1)], J. Palarea-Albaladejo[(2)] y C. Barceló-Vidal[(1)]

(1) Dept. Informàtica i Matemàtica Aplicada, Campus Montilivi, Univ. de Girona, E-17071 Girona, Spain.
josepantoni.martin@udg.edu; carles.barcelo@udg.edu
(2) Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom.
javier@bioss.ac.uk

RESUMEN

A menudo, las técnicas log-cociente se aplican en estudios geoquímicos. Cuando algunos elementos químicos están presentes en las muestras a concentraciones prácticamente inapreciables, se registran como valores por debajo del límite de detección, simplemente como ceros o bien anotados con una etiqueta del tipo *menor que*. La presencia de estas observaciones impide la aplicación directa de las técnicas log-cociente de análisis de datos composicionales. En este trabajo se presentan los fundamentos teóricos de las propuestas más recientes para tratar este tipo de observaciones. Se dan pautas para su aplicación práctica y se ilustra su funcionamento mediante ejemplos con datos geoquímicos.

Palabras clave: diagrama ternario, imputación, log-cociente, porcentajes, símplex

Compositional techniques to deal with concentrations below the detection limit in geochemistry

ABSTRACT

In most geochemical analyses log-ratio techniques are required to analyse compositional data sets. When a chemical element is present at a low concentration it is usually identified as a value below the detection limit and added to the data set either as zero or simply by attaching a less-than label. In any case, the ocurrence of such concentrations prevents us from applying the log-ratio approach. We review here the theoretical bases of the most recent proposals for dealing with these types of observation, give some advice on their practical application and illustrate their performance through some examples using geochemical data.

Keywords: imputation, log-ratio, percentages, simplex, ternary diagram

ABRIDGED ENGLISH VERSION

Introduction

In most geochemical analyses log-ratio techniques are required to analyse compositional data sets. In such studies some chemical elements are present at low concentrations. When these concentrations are so low as to be undetectable they are usually identified as values below detection limit (VBDL) and added to the data set either as zero or simply by attaching a less-than label. In any case, the ocurrence of such concentrations prevents us from applying the log-ratio approach. Consequently these values require some sort of prior treatment. This problem is usually covered under the heading rounded zeros in the compositional data literature. Martín-Fernández et al. *(2011) discuss extensively types of zeros and their appropriate treatment.*

We review and compare here two different methods for rounded zeros: the multiplicative replacement and the alr-EM algorithm. The performance of both methods is illustrated by some examples in which the function EMcomp (Palarea-Albaladejo and Martín-Fernández, 2008) for Matlab® v. 2008 is used.

Methods

Using a non-parametric imputation approach, the multiplicative replacement *method (1) imputes zeros with values* δ_{ij} *equal to 65% of the threshold* ε_{ij} *and adjusts the whole composition by the so-called multiplicative modification. This method has reasonable properties from a compositional point of view (Martín-Fernández* et al.*, 2003; Martín-Fernández and Thió-Henestrosa, 2006). In particular, it is "natural" in the sense that it recovers the "true" composition when replacement values are identical to the non-observed values; and it is also coherent with the basic operations on the simplex. This coherence implies that the covariance structure of subcompositions with no zeros is maintained.*

The database Cenozoic Volcanic Rocks of Hungary (Ó.Kovács and Kovács, 2001) was set up in order to contribute to our understanding of the petrogenetic processes that occurred in the Carpatho-Pannonian region. Our dataset consists of 959 unaltered rock samples and nine major oxides, SiO_2, TiO_2, Al_2O_3, Fe_2O_3*total, MgO, CaO,* Na_2O, K_2O *and* P_2O_5, *from this database. The pattern and the location of VBDL in*

the dataset is summarised in Table 1, where it can be seen that three components, SiO_2, Al_2O_3 and K_2O, have no rounded zeros. Among the rest, 101 compositions have at least one zero value and the zeros are mainly concentrated in components P_2O_5, TiO_2 and MgO. Because the number of VBDL in the data matrix is quite low (1.8%) it seems reasonable to consider a simple-substitution strategy (Martín- Fernández et al. *2003). It is logical to expect that any replacement strategy will restore the 'true' sample when the VBDL are replaced by the 'true' values. Sample* **x**** in Table 2 is an artificial sample obtained from sample* **x** *by making the TiO_2, MgO and P_2O_5 parts equal to zero, and then closed to obtain once more a sum equal to 100. In our example we use the values 0.14, 0.13 and 0.03 to replace the zeros in TiO_2, MgO and P_2O_5 respectively. Table 2 shows the results obtained from applying the different treatments of zero (Martín-Fernández* et al., *2003) to* **x****: the one proposed by Aitchison (1986), simple replacement, and multiplicative replacement (1). Note that multiplicative replacement is the only treatment that restores the true composition of* **x** *exactly.*

For the previous example in volcanic rocks, Ó. Kovács and Kovács (2001) consider a common threshold $\varepsilon = 0.01\%$ for all the components. According to Martín-Fernández et al. *(2003), we take it that δ equals 65% of the threshold, $\delta = 0.0065\%$, and apply the multiplicative replacement (1). Once the replacement is made some multivariate methods are usually applied. Two groups of samples exist in the database: alkaline basalts and the calc-alkaline series. Figure 1 shows a clr-biplot of the data obtained from the multiplicative replacement (1). This type of representation consists of a biplot diagram of the clr-transformed data, that is to say, the clr-biplot is a diagram in the clr-transformed space. Note that the first two axes of the diagram capture 76.61% of the total variability. The question that naturally arises is, "How robust are the results of a multivariate analysis in relation to the values imputed by the replacement method?" Therefore, a sensitivity analysis of the results in relation to these values must be carried out. Aitchison (1986) suggests that it is sufficient to perform an analysis of sensitivity of the values δ_{ij} in the rank $\varepsilon_{ij}/10 \leq \delta \leq \varepsilon_{ij}$,* where ε_{ij} is the threshold associated to the zero problem analysed.

For the database of Cenozoic volcanic rocks we are interested in linear discriminant analysis (LDA) using log-ratio methods. The good separation of the alkaline basalts and the calc-alkaline series spotted in the clr-biplot (Figure 1) is numerically confirmed by the 2-group LDA applied to the clr-transformed dataset: the misclassification rate is 3.96%. A sensitivity analysis must now be conducted. Figure 2 shows the variation pattern of the LDA misclassification rate when the value δ=0.000065 simultaneously varies for all parts between 0.00001 and 0.0001. For values around the imputed value the LDA rate is reasonably stable; when δ tends to zero, on the other hand, the misclassification rate increases because the separation between the two groups becomes more blurred.

The alr-EM algorithm [Equations (2) and (3)] consists of a modification of the classical EM algorithm, which is applied in combination with the additive log-ratio transformation (alr). This method takes into account that the imputed value must be lower than the given detection limit and is independent of the divisor selected for the alr transformation. Additionally, the zero values are imputed conditionally on the information included in the observed data. Reasonable estimations are obtained and, since this method is also coherent with the basic operations on the simplex, minimum distortion is produced. Figure 3 shows the clr-biplot of the data obtained from the alr-EM method applied to the database of Cenozoic volcanic rocks. In this case the first two axes of the diagram capture 77.21% of the total variability. Note that here the difference between the two methods (Figures 1 and 3) is negligible due to the small number of rounded zeros in the data matrix (1.8%). To illustrate the performance of both methods with a larger number of VDBL we select a subsample of the database. Firstly, we select the 101 compositions having at least one zero value. Secondly, from the remaimng 858 samples without rounded zeros we pick 35 at random. Therefore, the final subsample is made up of 136 samples and the corresponding data matrix has 12.5% entries equal to zero. Figure 4 shows the clr-biplots obtained after applying both VBDL replacement methods introduced above. The variability turns out to be 69.25% and 74.33%, respectively.

In Figure 4 the differences between the two replacement methods are greater than in Figures 1 and 3. For example, in Figure 4B the variable clr(P_2O_5) still has more variability than the variable clr(TiO_2). Nervertheless, for the *multiplicative replacement* the behaviour is the opposite (Figure 4A). Note that component P_2O_5 has the largest number of rounded zeros (Table 1). Because *multiplicative replacement* imputes zeros by the same value (65% of the threshold) the variability decreases. Also in Figure 4A, some samples located far away from the centre and corresponding to alkaline rocks show a linear pattern. These samples contain VBDL in the same components and replacement by a constant value causes linearity in the clr-transformed space. In addition, the total variability of the data set obtained after *multiplicative replacement* is 14.51, clearly smaller than the total variability (27.17) of the data set obtained from the alr-EM method.

Conclusions

When the number of rounded zeros is small the modified EM algorithm does not significantly improve the performance of the non-parametric multiplicative replacement. Nevertheless, case studies reveal its potential in the presence of a higher number of zero values. We have introduced here the theoretical background of both methods and given some practical advice. Lastly, we have illustrated their performance by means of geochemical examples.

Introducción

Cuando un investigador trabaja con la concentración de un elemento químico está tratando con la abundancia relativa del elemento respecto a un total. De manera que aquellos estudios geoquímicos en que los datos a analizar estadísticamente representan concentraciones de elementos, son estudios en los que las técnicas composicionales son apropiadas y útiles. Los últimos avances en el campo de los datos composicionales nos muestran que su análisis debe basarse en transformaciones log-cociente (ej. *clr*, del inglés *centred log-ratio*) o en una representación en coordenadas respecto a una base ortonormal (Egozcue *et al.*, 2003). En ambos casos, se pasa a trabajar con un nuevo conjunto de datos transformados resultado del cálculo de cocientes entre componentes y de la aplicación de la función logarítmica. Todas estas operaciones exigen que los valores de la matriz de datos original sean estrictamente positivos.

La mayoría de los estudios geoquímicos incluyen elementos químicos cuyos niveles de concentración son muy bajos. En aquellas muestras en las que la concentración de un elemento no llega al límite de detec-

ción del aparato de medida, el analista suele indicarlo registrando un valor cero o con una simple etiqueta tipo *menor que*. A estos valores se les conoce por las siglas VBDL del inglés *Value Below Detection Limit*. En estas situaciones no es viable aplicar transformaciones log-cociente o calcular coordenadas respecto a una base. Por otro lado, la aplicación de la mayoría de técnicas de análisis requiere que se disponga de muestras completas, es decir, muestras donde se han medido todas las variables. Una alternativa consistiría en no utilizar las muestras que contengan VBDL. Sin embargo, esa opción malogra los recursos invertidos en el proceso de muestreo e implica la consiguiente pérdida de información. En consecuencia, surge la necesidad de realizar un tratamiento previo de los vectores de observaciones que contienen VBDL. Ya en la monografía de Aitchison (1986) se propuso un técnica de imputación que años más tarde fue mejorada por Martín-Fernández *et al.* (2003). Los recientes avances en este tipo de tratamiento se han inspirado en la adaptación de las metodologías para datos perdidos (*missing data*).

En la siguiente sección presentamos diferentes situaciones en las que los conjuntos de datos composicionales pueden contener VBDL y, en general, ceros. A continuación, centrándonos en el caso de los VBDL, se introducen con detalle el método de reemplazamiento multiplicativo y el método EM-modificado. El funcionamiento de ambos métodos se ha ilustrado con ejemplos realizados con el software numérico Matlab® v. 2008, utilizando la función *EMcomp* introducida en Palarea-Albaladejo y Martín-Fernández (2008). Finalmente, en las conclusiones, se desgranan las ventajas e inconvenientes de ambos métodos.

Tipos de ceros en conjuntos de datos composicionales

Según su naturaleza, distinguiremos tres tipos de ceros: *esenciales*, *de conteo*, y *por redondeo*. Es importante identificar el tipo de cero con el que se corresponden los valores nulos presentes en un conjunto de datos, ya que el tratamiento a aplicar depende de la tipología de estos ceros. El lector interesado encontrará en Martín-Fernández *et al.* (2011) una discusión extensa sobre los tipos de ceros en conjuntos de datos composicionales y las propuestas más recientes para su tratamiento.

Las últimas contribuciones relevantes en el estudio de los ceros *esenciales* (ej. Bacon-Shone, 2008) coinciden en definir este tipo de cero como un valor nulo *verdadero*. Es decir, que no es debido a la incapacidad de detección de un valor muy pequeño como consecuencia del diseño experimental del muestreo o de la precisión del aparato de medida. A los ceros *esenciales* también se les conoce como ceros *estructurales* o *absolutos*. Dadas sus características, se deduce que no pueden recibir el mismo tratamiento que los ceros por redondeo o VBDL. En la actualidad no existe una metodología general para tratar los ceros esenciales. En la literatura más reciente se proponen principalmente dos aproximaciones al problema: una, basada en un modelo binomial condicional logístico-normal (Aitchison y Kay, 2003), aplicable a componentes de tipo continuo; y otra, basada en la distribución log-Poisson normal (Bacon-Shone, 2008), propuesta para componentes de tipo conteo.

Como hemos dicho, cuando un investigador decide utilizar cualquiera de estas dos alternativas está asumiendo que los ceros representan concentraciones nulas reales. Una situación distinta es la de los ceros de tipo *conteo*. En este contexto, se entiende que la composición se calcula a partir de un vector de recuento, esto es, a partir de una realización de una variable aleatoria categórica. Las componentes de este vector proporcionan el número de veces (frecuencia absoluta) que las diferentes categorías se observan sobre los ítems que componen una unidad muestral. En aquellos casos en que el investigador no está interesado en el número total de ítems (suma total de los valores del vector de recuento) ni en las frecuencias de las diferentes categorías, procederá a dividir éstas por el total de ítems y trabajará con el vector composicional de frecuencias relativas asociado, siendo útil en este caso la aplicación de las técnicas composicionales de análisis. Puede suceder que algunas de las categorías de la variable tengan asociadas probabilidades muy bajas y que, debido al diseño del muestreo o simplemente a un tamaño muestral pequeño, el conteo de alguna de esas categorías sea igual a cero. Este cero en una categoría no representa un valor *verdadero*, sino que es resultado del muestreo. Por este motivo, en estos casos, se asume implícitamente que de haber trabajado con una unidad muestral de mayor tamaño se podría haber registrado una frecuencia muy pequeña pero positiva en la categoría en cuestión. Para este tipo de ceros, Pierotti *et al.* (2009) aplican un tratamiento basado en un planteamiento *Bayesiano-multiplicativo*, donde se combinan técnicas bayesianas de imputación (Walley, 1996) con una modificación multiplicativa de los valores no nulos (Martín-Fernández *et al.*, 2003).

En los estudios donde las variables son de tipo continuo (ej. porcentajes de peso, tiempo o longitudes) los ceros más usuales son los ceros por *redondeo*, en los que nos centraremos en este trabajo.

Bajo esta denominación se incluyen tanto los VBDL como los ceros debidos a la utilización de un número insuficiente de cifras decimales significativas a la hora de plasmar numéricamente las frecuencias relativas de las distintas categorías. En este último caso, el valor nulo de la matriz de datos no es realmente un cero, sino una cantidad tan pequeña que por efecto del redondeo se ha transformado en un cero. Es decir, el cero presente en la componente *j*-ésima de una composición representa un valor pequeño inferior a un cierto error de redondeo ε ($\varepsilon = 10^{-d}$, donde *d* es el número de cifras significativas), que suele ser el mismo para todas las componentes de la matriz de datos. Por el contrario, los VBDL presentes en los datos –indicados por un cero o por una etiqueta– representan valores muy pequeños que ha sido imposible medir con exactitud. En este caso, cuando la observación *i*-ésima contiene un VBDL en la componente *j*-ésima sabemos que en realidad debería contener corresponde a un valor de la concentración inferior a ε_{ij}, donde ε_{ij} es el umbral o límite de detección. El hecho de incluir los dos subíndices en el umbral ε_{ij} responde a considerar el caso más general. De esta manera, se contempla la posibilidad de que el umbral sea diferente para cada componente *j* –debido, por ejemplo, a que el instrumento de medida sea diferente o tenga distinta sensibilidad según cual sea la componente– y, también, que una misma componente tenga distintos umbrales para distintas observaciones *i* –porque, por ejemplo, sean muestras analizadas en diferentes laboratorios.

En el escenario de ceros por *redondeo*, el valor *verdadero* es desconocido pero poseemos información de su valor máximo. Nos encontramos en una situación en la que un valor es *perdido* a causa del propio valor (demasiado pequeño). Esta particularidad ha llevado a que los últimos progresos en su tratamiento se hayan inspirado en algunas de las técnicas generales para el tratamiento de valores censurados y, en su caso más general, los valores perdidos (o faltantes) de tipo NMAR (*Not Missing At Random*) recogidas, p.e., en Little y Rubin (2002). En un sentido muy amplio, las técnicas para datos perdidos pueden clasificarse en técnicas paramétricas y no-paramétricas. Las primeras (ej., el algoritmo EM) se basan en distribuciones multivariantes de probabilidad –habitualmente relacionadas con la distribución normal–, de las que se estiman sus parámetros. Las técnicas no paramétricas se clasifican según cual sea la *técnica de imputación* utilizada. En el contexto de los ceros por *redondeo*, el término *imputación* significa la substitución del cero (o la etiqueta) por un valor "pequeño" sin necesidad de asumir ningún modelo de probabilidad para el conjunto de datos.

Reemplazamiento por imputación multiplicativa

Consideremos que la D-composición $\mathbf{x}_i=(x_{i1}, x_{i2}, ..., x_{iD})$ contiene Z ceros por redondeo y deseamos reemplazar $\mathbf{x}_i$ por una nueva composición $\mathbf{r}_i=(r_{i1}, r_{i2},..., r_{iD})$ sin ceros. En Martín-Fernández *et al.* (2003) se propone la fórmula

$$r_{ij} = \begin{cases} \delta_{ij} & \text{si } x_{ij} = 0, \\ \left(1 - \dfrac{\sum_{k|x_{ik}=0} \delta_{ik}}{c}\right) x_{ij} & \text{si } x_{ij} > 0; \end{cases} \qquad (1)$$

donde δ_{ij} es un valor pequeño, menor que el umbral ε_{ij} dado, y c es la constante –habitualmente 1, 100 (%) o 10^6 (*ppm*)– de la restricción de suma-constante que caracteriza los vectores composicionales. Nótese que en (1) se modifican también las componentes de $\mathbf{x}_i$ diferentes de cero con el fin de conservar el cumplimiento de esta restricción. Esta modificación, que es de tipo multiplicativo, y la imputación directa del valor δ_{ij} dotan a esta fórmula de reemplazamiento de unas propiedades (Martín-Fernández *et al.*, 2003; Martín-Fernández y Thió-Henestrosa, 2006) mejores que las del reemplazamiento propuesto originalmente en Aitchison (1986, página 269). En la literatura se conoce la expresión (1) como reemplazamiento *multiplicativo*. En estos mismos trabajos (Martín-Fernández *et al.*, 2003; Martín-Fernández and Thió-Henestrosa, 2006) los autores, desde un punto de vista teórico, también comparan las propiedades del reemplazamiento multiplicativo con las propiedades del reemplazamiento *simple*. Este tipo de reemplazamiento consiste en substituir los ceros por los valores δ_{ij} y, a continuación, simplemente realizar la clausura de todo el vector. Cuando se aplica este procedimiento a una composición con ceros, el valor δ_{ij} imputado inicialmente queda modificado como consecuencia de la clausura que se realiza a posteriori. Esta particularidad implica una pérdida de naturalidad en el tratamiento del valor cero puesto que el valor que finalmente reemplaza al cero no es el valor δ_{ij} decidido por el investigador. Es más, el valor finalmente imputado dependerá de la cantidad de ceros que tenga la composición original.

Con el propósito de ilustrar de manera práctica el comportamiento del reemplazamiento multiplicativo, lo aplicamos a continuación a un conjunto de datos geoquímicos de rocas volcánicas del Cenozoico de Hungría (Ó.Kovács y Kovács, 2001). El conjunto de datos consta de 959 muestras de rocas volcánicas cuya 9-composición en óxidos mayores [SiO_2; TiO_2; Al_2O_3;

Fe_2O_3total; MgO; CaO; Na_2O; K_2O; P_2O_5] es el objeto del estudio.

En la Tabla 1 se muestra el patrón y la cantidad de ceros según su localización en las diferentes variables. Sólo tres componentes no contienen ceros –SiO_2, Al_2O_3 y K_2O–, y 101 composiciones tienen al menos un cero, los cuales se localizan principalmente en los óxidos P_2O_5, TiO_2 y MgO. El mínimo valor observado en estas tres componentes, y por extensión en el conjunto de datos, es 0.01%.

(1986), el reemplazamiento simple, y el reemplazamiento multiplicativo. Con el propósito de ilustrar mejor los diferentes comportamientos de los respectivos tratamientos, se escogió una composición inicial **x** sin ningún cero en sus componentes. La composición $\mathbf{x}^*$ es una composición *artificial* que se ha obtenido a partir de la composición **x** (Tabla 2) forzando que las componentes TiO_2, MgO, y P_2O_5 tomen el valor cero y clausurándola a continuación para forzar que la suma sea igual a 100.

Número de ceros en la composición	Patrón de ceros									Cantidad de composiciones
	SiO_2	TiO_2	Al_2O_3	Fe_2O_3_tot	MgO	CaO	NaO_2	K_2O	P_2O_5	
0										858
1		0								8
1					0					12
1									0	39
2		0			0					3
2		0							0	25
2					0				0	3
2						0			0	1
3		0		0					0	1
3		0			0				0	5
3		0					0		0	3
3					0	0			0	1
Número de ceros	0	45	0	1	24	2	3	0	78	
Mínimo valor observado(%)	42.3	0.01	8.21	0.26	0.01	0.07	0.10	0.62	0.01	

Tabla 1. Patrón de ceros en el conjunto de datos de rocas volcánicas. "0" representa que la componente incluye ceros.
Table 1. Pattern of zeros in the volcanic-rock data set. '0' means that the component includes a zero value.

Nótese que los 153 ceros que contiene en total la matriz 959*x*9 de datos representa un porcentaje muy bajo (1.8%) del total (8631) de elementos de la matriz. Según se muestra en Martín-Fernández *et al.* (2003), este hecho hace recomendable considerar la aplicación del reemplazamiento multiplicativo puesto que en este caso se producirá una distorsión leve en la estructura de covarianzas log-cociente, es decir, en los valores que recoge la matriz de varianzas-covarianzas de los datos transformados.

En la Tabla 2 se recogen los resultados que se obtienen al aplicar los diferentes tratamientos de ceros a una misma composición: el propuesto por Aitchison

Para ilustrar que el reemplazamiento multiplicativo tiene un comportamiento más *natural* que los otros dos métodos, reemplazamos los ceros en $\mathbf{x}^*$ utilitzando un valor de imputación δ_{ij} igual al valor original que toma la componente en la composición **x**. De un tratamiento de ceros coherente se espera que al substituir el cero por su valor real toda la composición recupere sus valores verdaderos. En nuestro ejemplo, utilizamos los valores 0.14, 0.13, y 0.03, para reemplazar, respectivamente, el cero presente en $\mathbf{x}^*$ en los óxidos TiO_2, MgO, y P_2O_5. Como se puede comprobar en los resultados de la Tabla 2, únicamente el reemplazamiento multiplicativo recupera de nuevo toda la

Composición	SiO_2	TiO_2	Al_2O_3	Fe_2O_3_tot	MgO	CaO	Na_2O	K_2O	P_2O_5
x	75.0775	0.1400	14.3957	1.6495	0.1300	0.9397	2.9091	4.7286	0.0300
x^*	75.3033	0.0000	14.4390	1.6545	0.0000	0.9425	2.9179	4.7428	0.0000
Aitchison	75.2885	0.0415	14.4242	1.6397	0.0385	0.9277	2.9031	4.7280	0.0089
Simple	75.0782	0.1395	14.3958	1.6495	0.1296	0.9397	2.9092	4.7286	0.0299
Multiplicativo	75.0775	0.1400	14.3957	1.6495	0.1300	0.9397	2.9091	4.7286	0.0300

Tabla 2. Reemplazamientos de Aitchison, simple y multiplicativo aplicados a una misma composición **x***. La composición "artificial" **x***, con ceros en los elementos TiO_2, MgO y P_2O_5, se obtiene a partir de la composición **x.**
Table 2. Replacements according to Aitchison; simple and multiplicative replacements applied to composition ***x****. The "artificial" composition* ***x****, with zeros in the elements TiO_2, MgO and P_2O_5, is obtained from composition* ***x.***

composición original, mostrando una naturalidad que los otros dos reemplazamientos no poseen.

El lector interesado encontrará más ejemplos en Martín-Fernández y Thió-Henestrosa (2006). En Martín-Fernández *et al.* (2003) se lleva a cabo un estudio comparativo desde un punto de vista teórico. En particular, los autores revisan con detalle las propiedades de los reemplazamientos en relación a las operaciones básicas en el símplex: subcomposición, perturbación, y transformación potencia; y en relación a los elementos básicos de la metodología log-cociente: distancia de Aitchison, media geométrica composicional, matriz de covarianzas y variabilidad total.

Un elemento clave del método de reemplazamiento multiplicativo es la decisión sobre qué valor δ_{ij} escoger en la fórmula (1). El valor utilizado debe ser un valor pequeño y no superior al umbral. En Martín-Fernández *et al.* (2003) se realizaron estudios de sensibilidad en función del valor δ_{ij} utilizado, y se mostró que los mejores resultados se obtienen utilizando δ_{ij} igual al 65% del valor del umbral ε_{ij}. Habitualmente, tras el tratamiento de los ceros, los investigadores realizan análisis multivariantes de los datos cuyos resultados se expresan a menudo mediante índices. Por ejemplo, en análisis discriminante se calcula la tasa de clasificación errónea; en regresión lineal múltiple, el coeficiente de determinación r^2; o en análisis de componentes principales, la proporción de variabilidad explicada por las diferentes componentes. Inmediatamente surge la pregunta de hasta qué punto el valor del índice obtenido depende de los valores δ_{ij} utilizados en el tratamiento previo de los ceros. Esta cuestión debe analizarse a través de un análisis de sensibilidad. En Aitchison (1986) se sugiere que es suficiente realizar un análisis de sensibilidad de los valores δ_{ij} en el rango $\varepsilon_{ij}/10 \leq \delta \leq \varepsilon_{ij}$, donde ε_{ij} es el umbral asociado al cero analizado.

Para el ejemplo anterior sobre rocas volcánicas (Ó.Kovács y Kovács, 2001), se considera un umbral ε=0.01% común a todas las componentes. De acuerdo con el criterio de Martín-Fernández *et al.* (2003), tomamos como valor δ el 65% del umbral, i.e. δ =0.0065%, y aplicamos el reemplazamiento multiplicativo. En el conjunto de datos existen dos grupos de rocas: basaltos alcalinos y rocas calc-alcalinas. Para representar el conjunto de datos utilizamos un diagrama *clr-biplot* (Figura 1). Este tipo de representación consiste en construir un diagrama biplot de los datos que se obtienen al aplicar la transformación clr a los datos originales, es decir, es una representación gráfica en el espacio *clr-transformado*. Nótese que los dos primeros ejes del diagrama logran capturar un 76.61% de la variabilidad total.

Como se puede apreciar en la Figura 1 los dos grupos de rocas aparecen localizados en diferente posi-

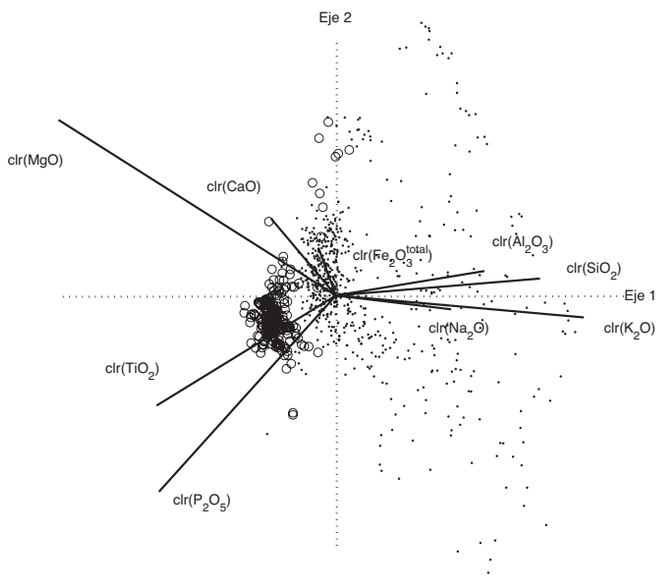

Figura 1. Diagrama clr-biplot de los datos de las rocas volcánicas tratados mediante reemplazamiento multiplicativo. Basaltos alcalinos (círculos) y rocas calc-alcalinas (puntos).
Figure 1. clr-biplot of volcanic-rock data after multiplicative replacement. Alkaline basalts (circles) and calc-alkaline rocks (dots).

ción, lo que sugiere la posibilidad de obtener una buena discriminación lineal. Una vez realizado el análisis discriminante lineal log-cociente, la separación observada en el diagrama clr-biplot se confirma numéricamente al obtener una tasa de clasificación errónea del 3.96%. Para el análisis de sensibilidad de esta tasa haremos variar el valor δ entre 0.00001 y 0.0001 de acuerdo con las sugerencias de Aitchison (1986). Al repetir el cálculo de la tasa de clasificación errónea para cien valores de δ incluidos en este rango se obtienen los resultados que muestra la Figura 2.

La tasa de clasificación errónea parece razonablemente estable para valores de δ cercanos a 0.0065%, incrementando su valor cuando δ tiende a cero. Para valores muy pequeños de δ los grupos se entremezclan más debido a que los dos tienen composiciones con ceros y éstas, al disminuir el valor δ utilizado, tienden a disminuir su distancia mútua y, simultáneamente, a alejarse del centro del espacio clr-transformado. Así, los basaltos alcalinos se alejan del centro del espacio en la dirección opuesta a las variables clr(TiO_2), clr(MgO) y clr(P_2O_5) mezclándose con el otro grupo de rocas y incrementando la tasa de clasificación errónea.

La ejecución del análisis de sensibilidad de los resultados obtenidos en un estudio puede llegar a ser una tarea laboriosa y compleja. Sin embargo, este análisis es imprescindible en los estudios en los que se haya aplicado el reemplazamiento multiplicativo. Esta dificultad, añadida al hecho de que la distorsión que el reemplazamiento multiplicativo provoca en la estructura de covarianzas se incrementa al aumentar la cantidad de ceros presentes en el conjunto de datos, hace que en muchas situaciones sea más recomendable utilizar el método alr-EM propuesto en Palarea-Albaladejo y Martín-Fernández (2008).

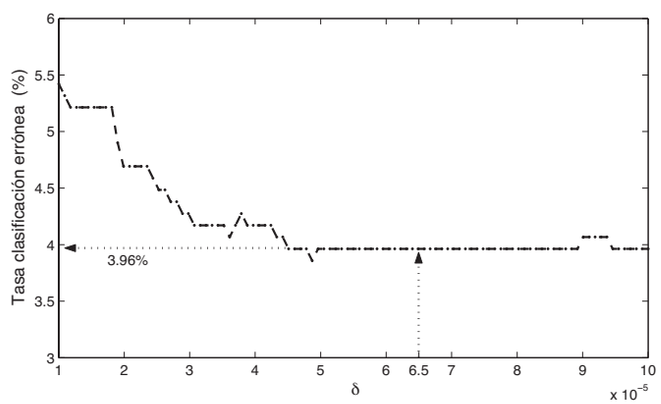

Figura 2. Análisis de sensibilidad en función de δ para la tasa de clasificación errónea en el análisis discriminante lineal log-cociente de las rocas volcánicas.
Figure 2. Sensitivity analysis: misclassification rate in the log-ratio LDA of the volcanic-rock data set.

Reemplazamiento mediante el algoritmo EM

Palarea-Albaladejo *et al.* (2007), en el contexto del tratamiento de ceros por redondeo, introdujeron la idea de aplicar el conocido algoritmo EM –del inglés *Expectation and Maximization*– al conjunto de datos obtenidos al aplicar la transformación log-cociente *alr* –del inglés *additive log-ratio*– a los datos originales. El algoritmo EM es un procedimiento iterativo comúnmente aplicado (ej. Little y Rubin, 2002) en problemas de estimación máximo-verosímil a partir de conjuntos de datos reales multidimensionales con valores perdidos.

En nuestro contexto, si una composición $\mathbf{x}_i$ contiene un cero por redondeo en su *j*-ésima componente, su valor alr-transformado $y_{ij}=\ln(x_{ij}/x_{iD})$ puede interpretarse como un valor perdido en el espacio real del cual se sabe que cumple la desigualdad $y_{ij} \leq \ln(\varepsilon_{ij}/x_{iD})= \psi_{ij}$, donde ψ_{ij} es el valor alr-transformado del umbral ε_{ij}. La contribución teórica de Palarea-Albaladejo *et al.* (2007) consistió en proponer una modificación del algoritmo EM estándar para que el procedimiento aplicado en el espacio alr-transformado tuviese en cuenta que los valores perdidos y_{ij} deben ser reemplazados por valores inferiores al umbral transformado ψ_{ij}.

En sus fundamentos teóricos el algoritmo iterativo EM-modificado (Palarea-Albaladejo *et al.*, 2007) asume que el conjunto de datos a tratar son realizaciones de un vector aleatorio composicional **x** cuya distribución es la aditiva logístico normal o modelo *aln* (Aitchison, 1986). Es decir, el vector aleatorio alr-transformado **y**=alr(**x**) tiene por modelo de probabilidad el normal (D-1)-dimensional, con vector de medias μ y matriz de covarianzas Σ. En la etapa de estimación (*E-step*) de su *k*-ésima iteración, el algoritmo alr-EM reemplaza los valores perdidos y_{ij} del conjunto de datos transformados **Y** mediante la expresión

$$y_{ij}^{(k)} = \begin{cases} y_{ij} & \text{if } y_{ij} \geq \psi_{ij} \\ E\left(y_{ij} \mid \mathbf{y}_{i,-j}, y_{ij} < \psi_{ij}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) & \text{if } y_{ij} < \psi_{ij}, \end{cases} \quad (2)$$

donde $\mathbf{y}_{i,-j}$ representa el subvector de datos no perdidos de la *i*-ésima fila de **Y**. Nótese que con el subíndice "-j" se indica el conjunto de columnas que contienen valores observados en la *i*-ésima fila.

Si ϕ y Φ son, respectivamente, la función de densidad y de distribución de la ley de probabilidad normal estándar, el valor esperado en (2) se calcula mediante la fórmula

$$E\left(y_{ij} \mid \mathbf{y}_{i,-j}, y_{ij} < \psi_{ij}\right) = \mathbf{y}_{i,-j}^{T}\beta_j - \sigma_j \frac{\phi\left(\frac{\psi_{ij} - \mathbf{y}_{i,-j}^{T}\beta_j}{\sigma_j}\right)}{\Phi\left(\frac{\psi_{ij} - \mathbf{y}_{i,-j}^{T}\beta_j}{\sigma_j}\right)}, \qquad (3)$$

donde β_j es el vector de coeficientes que se obtiene al estimar mediante regresión lineal el valor perdido y_{ij} a partir del subvector $\mathbf{y}_{i,-j}$ de valores observados, y σ_j es la desviación estándar condicional de la variable aleatoria $\mathbf{y}_j$. Una vez reemplazados los valores perdidos y_{ij}, el algoritmo alr-EM, en su etapa de maximización (*M-step*), calcula las nuevas estimaciones máximo-verosímiles del vector de medias μ y de la matriz de covarianzas Σ. Se sigue iterando hasta alcanzar la convergencia o criterio de parada cuya formulación se basa en detectar en las estimaciones de los parámetros μ y Σ una diferencia en valor absoluto menor que un nivel de tolerancia prefijado (ej. 0.0001).

En Palarea-Albaladejo y Martín-Fernández (2008) se presentó un estudio práctico con datos reales y datos sintéticos donde se ponen de manifiesto las mejoras que representa el algoritmo EM-modificado en relación al método de reemplazamiento multiplicativo. Se constató que a medida que aumenta la cantidad de ceros por redondeo presentes en el conjunto de datos composicionales, el algoritmo EM-modificado introduce menos sesgo y proporciona mejores estimaciones de la variabilidad de los datos. Si bien en los dos métodos el valor imputado se decide teniendo en cuenta la información de los umbrales, el algoritmo EM-modificado no deja la decisión en manos del investigador si no que realiza una estimación máximo-verosímil a partir de los datos observados en el conjunto de la base de datos. Es decir, el cálculo de los valores imputados incorpora la información contenida en los valores no perdidos del conjunto de datos. Una consecuencia de este hecho es que, cuando se aplica el algoritmo EM-modificado, se hace innecesaria la realización de un análisis de sensibilidad de los resultados de posteriores análisis multivariantes. Por lo tanto, se simplifica el estudio a realizar cuando los ceros en el conjunto de datos presentan un patrón complejo. Sin embargo, como se constató en Palarea-Albaladejo y Martín-Fernández (2008), para aquellos conjuntos de datos con patrones de ceros sencillos –pocos ceros y localizados en unas pocas componentes– el reemplazamiento multiplicativo produce resultados muy similares a los del algoritmo EM-modificado.

Para ilustrar el funcionamiento del algoritmo sobre el mismo conjunto de datos que hemos utilizado en el caso del reemplazamiento multiplicativo, empleare-

mos la función *EMCOMP*, escrita en MatLab® y descrita en Palarea-Albaladejo y Martín-Fernández (2008). La correcta ejecución de este programa, que puede obtenerse libremente en la página www.compositionaldata.com, se ha testeado recientemente con Matlab® R2008b. En la documentación que lo acompaña se encuentra toda la información necesaria para su instalación y uso.

Se aplicó el algoritmo EM-modificado al conjunto de datos de rocas volcánicas considerando, al igual que en el reemplazamiento multiplicativo, un umbral de detección del 0.01%. A continuación, se aplicó la transformación clr a los datos y se representaron en el diagrama clr-biplot (Figura 3), con un porcentaje de explicación de la variabilidad del 77.21%.

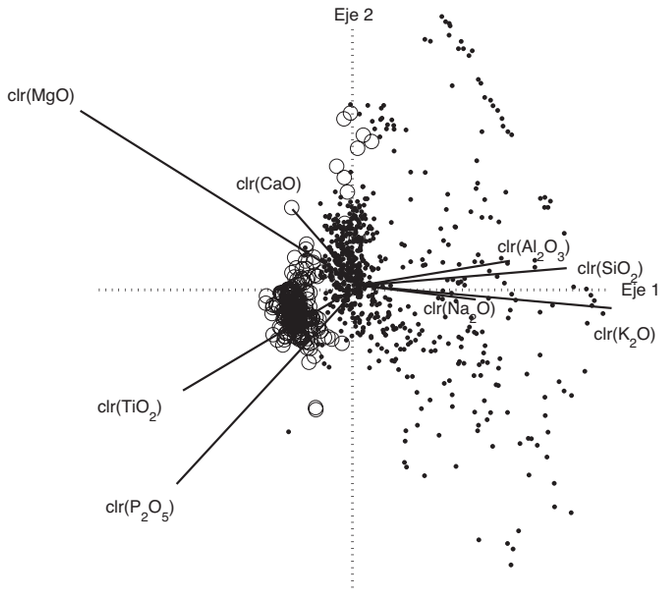

Figura 3. Diagrama clr-biplot de los datos de las rocas volcánicas tratados mediante el algoritmo EM-modificado. Basaltos alcalinos (círculos) y rocas calc-alcalinas (puntos).
Figure 3. clr-biplot of volcanic-rock data via the EM-modified algorithm. Alkaline basalts (circles) and calc-alkaline rocks (dots).

Comparando la Figura 3 con el clr-biplot de los datos producidos por el reemplazamiento multiplicativo (Figura 1) se constata que la diferencia entre los dos métodos es, en este caso, muy pequeña. Este hecho se puede explicar porque sólo un 1.77% de los valores de la matriz del conjunto de datos son iguales a cero. De esta manera, no se aprecia casi ninguna diferencia entre imputar un valor constante para todos los individuos (reemplazamiento multiplicativo) e imputar un valor teniendo en cuenta la variabilidad de los datos observados no nulos (algoritmo EM-modificado).

Para ilustrar qué sucede cuando el porcentaje de ceros es más elevado, seleccionamos un subconjunto

de muestras geoquímicas del conjunto de datos anterior. Primeramente, escogemos las 101 muestras (Tabla 1) que contienen algún cero. A su vez, de entre las 858 muestras que no contienen ningún valor cero, seleccionamos 35 de ellas de forma aleatoria. De este modo, el nuevo conjunto de datos estará formado por 136 muestras con un porcentaje de valores igual a cero (12.5%) mucho mayor que el conjunto anterior, aunque con el mismo patrón (Tabla 1).

A este nuevo conjunto de datos se le aplicaron las dos estrategias de tratamiento de ceros, utilizando el mismo umbral (0.01%). La Figura 4 muestra los diagramas clr-biplot de los dos conjuntos sin valores nulos resultantes, cuyos porcentajes de variabilidad explicada fueron del 69.25% y 74.33%, respectivamente, para el reemplazamiento multiplicativo y el algoritmo EM-modificado.

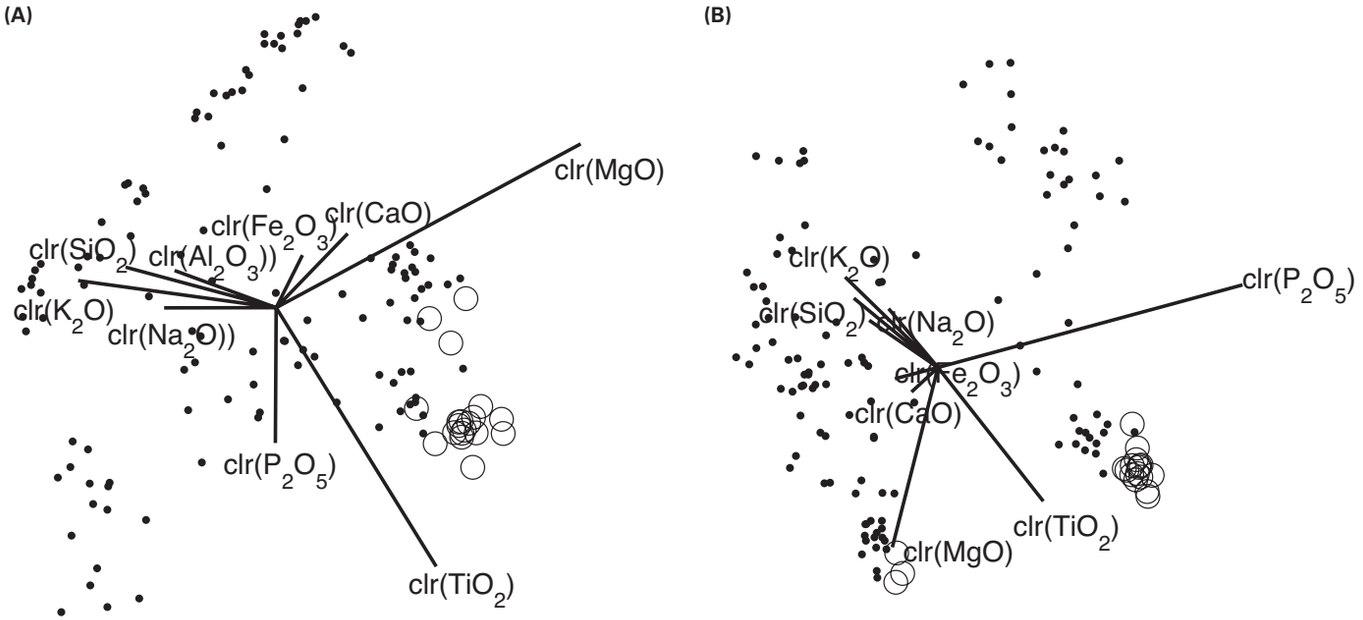

Figura 4. Diagrama clr-biplot de los datos de las rocas volcánicas tratados mediante: (A) reemplazamiento multiplicativo; (B) algoritmo EM-modificado. Basaltos alcalinos (círculos) y rocas calc-alcalinas (puntos).
Figure 4. clr-biplot of volcanic-rock data via: (A) multiplicative replacement; (B) EM-modified algorithm. Alkaline basalts (circles) and calc-alkaline rocks (dots).

Al contrario de lo que acontecía en la comparación entre la Figura 1 y 3, al observar los clr-biplot de la Figura 4 se aprecian mayores diferencias entre los dos métodos de reemplazamiento. Por ejemplo, en la Figura 4(B) se observa que la variable clr(P_2O_5) sigue teniendo más variabilidad que la variable clr(TiO_2). No ocurre lo mismo en la Figura 4(A). Nótese que la componente P_2O_5 es la que posee una cantidad mayor de ceros (Tabla 1). Puesto que al aplicar el reemplazamiento multiplicativo estos ceros se sustituyen todos por un mismo valor –65% del umbral– conlleva una reducción de la variabilidad. También se aprecia en la Figura 4(A) que algunas muestras alejadas del centro, correspondientes a rocas alcalinas (puntos), adoptan un cierto patrón lineal. En concreto, se trata de aquellas muestras que contenían ceros y que han sido sustituidos por valores pequeños. Este comportamiento se debe a que el reemplazamiento multiplicativo imputa un valor constante para todas las muestras y, por tanto, provoca linealidad en el espacio clr-transformado. El hecho de que la variabilidad sea menor para el caso del reemplazamiento multiplicativo se confirma observando que la variabidad total calculada sobre los datos tratados con el reemplazamiento multiplicativo es igual a 14.51 y, en cambio, es igual a 27.17 cuando los datos se tratan mediante el algoritmo EM-modificado.

Conclusiones

Hemos visto que el problema práctico causado por la presencia de valores por debajo del límite de detección en estudios geoquímicos con datos composicionales podría ser satisfactoriamente solventado en una variedad de casos mediante la aplicación tanto del reemplazamiento multiplicativo como del algoritmo alr-EM. Ambos métodos están diseñados para no distorsionar la estructura de covariancias del conjunto de datos.

La decisión sobre qué método aplicar en un determinado caso práctico la debe valorar en último término el propio investigador, principalmente en base a las características particulares de los datos de los que dispone, el tamaño de la matriz de datos, y los supuestos teórico-estadísticos que está dispuesto a asumir. En general, el método de reemplazamiento multiplicativo sería recomendable para conjuntos de datos con relativamente pocos valores por debajo del límite de detección y que siguen un patrón simple. Si este patrón se complica y la cantidad de valores no observados es moderadamente alta, es de esperar que el algoritmo EM-modificado proporcione resultados más adecuados, especialmente si se dispone de grandes matrices de datos donde se satisfacen los supuestos estadísticos en los que se basa el algoritmo.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto "CODA-RSS"' (Ref. MTM2009-13272) del Ministerio de Ciencia e Innovación y por el proyecto "RGCODA" (Ref. 2009SGR424) de la Agència de Gestió d'Ajuts Universitaris i de Recerca de la Generalitat de Catalunya.

Referencias

Aitchison, J. 1986. *The statistical analysis of compositional data*. London, Chapman and Hall Ltd; reprinted in 2003 at Caldwell, NJ: Blackburn Press, 416 pp.

Aitchison, J. and Kay, J. 2003. Possible solution of some essential zero problems in compositional data analysis. En: *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, Thió-Henestrosa, S. and J.A. Martín-Fernández (Eds.), Universitat de Girona, CD-ROM (ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/), October 15-17, 6 pp.

Bacon-Shone, J. 2008. Discrete and continuous compositions. In: *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop*, Daunis-i-Estadella, J. and J. A. Martín-Fernández (eds.), Universitat de Girona, CD-ROM (ISBN 84-8458-272-4, http://ima.udg.es/Activitats/CoDaWork08/), May 27-30, 11 pp.

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

Little, R. J. A. y Rubin, D. B. 2002. *Statistical Analysis with Missing Data* (2nd ed.). John Wiley and Sons, New York (USA). 381 pp.

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.

Martín-Fernández, J.A., Palarea-Albaladejo, J. and Olea, R.A. 2011. Dealing with Zeros. In: Pawlowsky, V. and Buccianti, A. (eds) *Compositional Data Analysis: Theory and Applications*. Chichester (UK), John Wiley & Sons. Chapter 4, 43-58.

Martín-Fernández, J. A. and Thió-Henestrosa, S. 2006. Rounded zeros: some practical aspects for compositional data. In: A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn (eds.), *Compositional data analysis from theory to practice*. London: The Geological Society, Special Publications 264, 191–201.

Ó.Kovács, L. and Kovács, G.P. 2001. Petrochemical database of the Cenozoic volcanites in Hungary: structure and statistics. *Acta Geologica Hungarica* 44 (4), 381–417.

Palarea-Albaladejo, J., Martín-Fernández, J.A. and Gómez-García, J. 2007. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology* 39(7), 625–645.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computer&Geosciences* 34(8), 902–917.

Pierotti, M.E.R., Martín-Fernández, J.A. y Seehausen, O. 2009. A mapping individual variation in male mating preference space: multiple choice in a colour polymorphic cichlid fish. *Evolution* 63 (9), 2372-2388.

Walley, P. 1996. Inferences from multinomial data: learning about a bag of marbles (with discussion). *Journal Royal Statistical Society B* 58, 3-57.

Recibido: noviembre 2010
Revisado: abril 2011
Aceptado: julio 2011
Publicado: octubre 2011