Contribuciones a la Estadística
y a la Investigación Operativa
Sicilia et. al. (editores)

ISBN 84-689-8552-X
páginas 1366–1382

Valores composicionales por debajo del límite de detección: un reemplazamiento paramétrico

Javier Palarea-Albaladejo[1], *Josep Antoni Martín-Fernández*[2]

[1]jpalarea@pdi.ucam.edu, Dpto. de Informática de Sistemas, Universidad Católica San Antonio
[2]josepantoni.martin@udg.es, Dpto. de Informática y Matemática Aplicada, Universidad de Girona

Abstract

En este trabajo se presenta un método paramétrico para reemplazar valores por debajo del límite de detección en muestras de datos composicionales. Comparándolo con otros métodos existentes, se analizan sus propiedades y se evalúa su rendimiento mediante la aplicación a conjuntos de datos reales.

Palabras Clave: datos composicionales, algoritmo EM, ceros por redondeo.

AMS: 62Pxx, 62-07, 65C05

1. Introducción

Un dato composicional es la realización de un vector aleatorio D-dimensional $x = [x_1, \ldots, x_D]$ tal que $x_i > 0$, $\forall\ i = 1, \ldots, D$, y $x_1 + x_2 + \ldots + x_D = c$. Nos referiremos a los elementos x_i como partes o componentes del vector x. En cuanto a c, su valor es irrelevante desde un punto de vista matemático, y suele ser 1, 100, 10^6, dependiendo de las unidades de medida. En [1] se establece que el espacio muestral para este tipo de datos es el simplex $\mathcal{S}^D$ definido como

$$\mathcal{S}^D = \left\{[x_1, \ldots, x_D] : x_1 > 0, x_2 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = c\right\}.$$

La operación *perturbación*, $x \oplus x^* = \mathcal{C}\,[x_1 x_1^*, \ldots, x_D x_D^*]$, definida sobre $\mathcal{S}^D \times \mathcal{S}^D$, y la *transformación potencia*, $\alpha \otimes x = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$, definida sobre $\mathrm{R} \times \mathcal{S}^\mathrm{D}$, inducen una estructura de espacio vectorial sobre $\mathcal{S}^D$. El operador *clausura* $\mathcal{C}$ se define como $\mathcal{C}(x) = [x_1/\sum x_i, \ldots, x_D/\sum x_i]$.

Un análisis adecuado de este tipo de datos debe basarse en el estudio de las magnitudes relativas y no en las magnitudes absolutas ([1]). En consecuencia, cualquier aseveración sobre una composición debe realizarse en términos de cocientes entre las componentes. Los log-cocientes $\log(x_i/x_j)$ son más fáciles de manejar matemáticamente que los cocientes, y además la transformación log-cociente proporciona una correspondencia uno a uno entre vectores composicionales y vectores log-cociente transformados en un espacio real. Esto evita el problema de un espacio muestral original restringido, el símplex, pasando a trabajar en un espacio real multivariante no restringido, lo que permite entonces emplear las técnicas multivariantes disponibles para espacios reales. En consecuencia, la metodología log-cociente se ha convertido en el enfoque estándar para el análisis estadístico de datos composicionales. La principal transformación log-cociente empleada hasta la fecha en estudios paramétricos es la *transformación log-cociente aditiva* (alr), definida como

$$\mathrm{alr(x)} = \big[\ln(x_1/x_D), \ldots, \ln(x_{D-1}/x_D)\big]. \tag{583}$$

Podemos definir una distancia log-cociente entre dos composiciones x y x^* de $\mathcal{S}^D$, la distancia de Aitchison, como

$$\mathrm{d_a}(x, x^*) = \mathrm{d_e}(\mathrm{clr}(x), \mathrm{clr}(x^*)), \tag{584}$$

donde $\mathrm{clr}(\cdot)$ se refiere a la *transformación log-cociente centrada* definida como $\mathrm{clr}(x) = [\log(x_1/g(x)), \ldots, \log(x_D/g(x))]$, siendo $g(x) = (x_1 \cdots x_D)^{1/D}$ la media geométrica de la composición x; y $\mathrm{d_e}$ es la distancia Euclídea en $\mathrm{R^D}$. Esta distancia es totalmente compatible con la estructura de espacio vectorial del símplex ([2]). La transformación (583) es asimétrica respecto a sus componentes, por lo que en la práctica deberemos comprobar si nuestra técnica estadística es invariante frente a permutaciones de las partes de una composición.

La presencia de valores nulos en una composición impide la aplicación de cualquier transformación log-cociente. Con frecuencia, dichos valores nulos aparecen en una muestra de datos composicionales porque la parte correspondiente toma valores por debajo del límite de detección de los aparatos de medida, y es habitual referirse a ellos como *ceros por redondeo*. Para reemplazar tales ceros pueden emplearse métodos de imputación, ya que en esencia un cero por redondeo es como un valor perdido que no ha podido observarse ([14]). No trataremos aquí los valores nulos debidos a que una componente toma realmente valor cero (véase [4] y [6]).

En este trabajo proponemos un enfoque paramétrico para reemplazar los valores que no superan el límite de detección utilizando la transformación alr. Este enfoque se basa en el conocido algoritmo EM ([8]) y tiene en cuenta la naturaleza especial tanto de los datos composicionales como de los ceros por redondeo.

2. Estrategias de reemplazamiento de ceros por redondeo

Una técnica de imputación o reemplazamiento no debe distorsionar la estructura general de los datos ya que, en caso contrario, los posteriores análisis sobre subpoblaciones podrían llevar a resultados erróneos. En particular, cualquier estrategia de reemplazamiento de valores por debajo del límite de detección en conjuntos de datos composicionales debe ser coherente con las operaciones básicas en el símplex, preservando la estructura de covarianzas y la métrica inducida por la metodología log-cociente.

Desde un punto de vista no paramétrico, [1] propone un método de reemplazamiento aditivo para los ceros por redondeo. En [9] y [13], independientemente, ponen de relieve que el reemplazamiento aditivo no preserva los cocientes entre los valores no nulos y, en consecuencia, distorsiona la estructura de covarianzas del conjunto de datos. En [14] se propone y analiza un *reemplazamiento multiplicativo*. Consideremos una composición $x \in \mathcal{S}^D$ que contiene ceros por redondeo. Mediante el reemplazamiento multiplicativo, x se sustituye por una nueva composición $r = [r_1, \ldots, r_D] \in \mathcal{S}^D$ sin ceros aplicando la regla siguiente:

$$r_j = \begin{cases} \delta_j & \text{si } x_j = 0 \\ \left(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}\right) x_j & \text{si } x_j > 0, \end{cases} \tag{585}$$

donde δ_j es un valor pequeño por debajo del límite de detección dado para la componente x_j. Este procedimiento recupera la *verdadera* composición si los valores de δ_j son idénticos a los valores no observados, es coherente con las operaciones básicas en el símplex y preserva la estructura de covarianzas de las subcomposiciones sin ceros. Sin embargo, aplicando (585) todos los ceros de una componente se imputan con el mismo valor, por lo que introduce una correlación artificial entre las componentes que tengan valores nulos en las mismas observaciones. Este efecto puede distorsionar un análisis multivariante posterior si el número de valores nulos en la muestra de datos es elevado, superior al 10 % ([19]).

Cuando el número de ceros es elevado, son recomendables métodos paramétricos que exploten la información contenida en la estructura de covarianzas. En este caso, un valor imputado tendrá en cuenta los valores del resto de componentes del dato composicional observado y, por lo tanto, no tendrá por qué ser siempre el mismo. Desde un punto de vista paramétrico, el algoritmo EM ([8]) y el método de imputación múltiple (IM) ([18]) se han revelado como las técnicas más fiables para abordar problemas de datos incompletos en espacios reales, como se ilustra mediante simulación en [10]. Si se aplica directamente cualquiera de ellas a datos composicionales, esto es, sin haber aplicado previamente una transformación log-cociente sobre los mismos, su estructura quedará seriamente distorsionada. Siguiendo a [1], si se aplica la transformación alr a un conjunto de datos composicionales podemos considerar un modelo normal multivariante

para los datos alr-transformados en un espacio real. Esta metodología de la transformación se aplica en [7]. Desde un punto de vista empírico, los autores describen el rendimiento del algoritmo EM y de su extensión, el método de Sandford ([19]), sobre un conjunto de datos alr-transformados. Por otra parte, en [15] se hace una primera aproximación a la imputación múltiple de datos alr-transformados y se describe su comportamiento.

3. Un algoritmo EM modificado para tener en cuenta el límite de detección

3.1. Algoritmo EM y datos incompletos

El algoritmo EM es un procedimiento paramétrico iterativo diseñado para obtener estimaciones máximo-verosímiles (MV) cuando una muestra de datos es parcialmente observada. Sea $Y = (Y_{obs}, Y_{per})$ una muestra multivariante incompleta, donde Y_{obs} y Y_{per} denotan, respectivamente, la parte observada y no observada de Y. Sea θ el vector de parámetros desconocidos de la distribución de probabilidad P para los datos completos. Denotemos mediante $\mathcal{L}(\theta|Y)$ a la función de log-verosimilitud asociada. En la t-ésima iteración del algoritmo EM se ejecutan dos pasos:

- **Paso E:** Dada una estimación $\theta^{(t)}$ de θ, calcular $Q(\theta;\theta^{(t)})$, donde

$$Q(\theta;\theta^{(t)}) = \int \ln \mathcal{L}(\theta|Y) P[Y_{per}|Y_{obs},\theta^{(t)}] dY_{per}.$$

- **Paso M:** Encontrar $\theta^{(t+1)}$ tal que $\theta^{(t+1)} = \arg\max_\theta Q(\theta;\theta^{(t)})$.

Partiendo de un punto inicial $\theta^{(0)}$, los pasos E y M se repiten alternativamente, con lo que se genera una sucesión de estimadores $\{\theta^{(t)}\}$. El algoritmo EM tiene la propiedad de que en cada iteración se incrementa $\mathcal{L}(\theta|Y_{obs})$, la función de log-verosimilitud para datos observados, esto es,

$$\mathcal{L}(\theta^{(t+1)}|Y_{obs}) \geq \mathcal{L}(\theta^{(t)}|Y_{obs}),$$

dándose la igualdad si y sólo si $Q(\theta^{(t+1)};\theta^{(t)}) = Q(\theta^{(t)};\theta^{(t)})$. Podemos encontrar propiedades de convergencia del algoritmo en [8] y [20]. Bajo ciertas condiciones de regularidad, $\{\theta^{(t)}\}$ converge hacia la estimación MV de θ. Para más detalles sobre los aspectos teóricos y prácticos del algoritmo EM remitimos al lector a [16] o [12].

Casi todos los métodos para datos incompletos utilizados en la práctica descansan, al menos implícitamente, en un supuesto denominado *ignorabilidad*. Esto es, el analista puede ignorar el mecanismo que genera los datos incompletos y considerar unicamente la verosimilitud para datos observados. Se suelen

distinguir dos situaciones ignorables: MAR (*missing at random*), cuando la probabilidad de que un dato no se observe depende de la parte observada Y_{obs} pero no de Y_{per}; y MCAR (*missing completely at random*), cuando un dato no observado surge de forma totalmente aleatoria. La hipótesis MCAR es un caso particular del supuesto MAR, un caso más restrictivo y menos frecuente. El algoritmo EM estándar, el método de Sandford y el método IM asumen la hipótesis MAR, y sus desarrollos se basan en la función de verosimilitud para datos observados. Por otra parte, se habla de hipótesis NMAR (*not missing at random*) cuando se asume que la probabilidad de que un valor no se observe depende de Y_{per}. Este caso representa la situación no ignorable y requiere modelos y métodos especiales. Para datos continuos, un grupo de métodos se basan en los llamados *modelos de selección* ([5];[11]). En un contexto composicional, el problema de los ceros por redondeo es un caso NMAR particular: los datos no se observan porque su valor se encuentra por debajo del límite de detección. Esto explica los resultados poco satisfactorios del algoritmo EM estándar, del método de Sandford y del método IM.

3.2. Paso E modificado

Supongamos que $y = [y_1, \ldots, y_{D-1}]$ es el vector alr-transformado de la composición $x = [x_1, \ldots, x_D]$. Asumimos que y se distribuye según una distribución normal $(D-1)$-dimensional con vector de medias μ y matriz de covarianzas Σ, esto es, x se distribuye según una normal logística aditiva (aln) ([1]). Como es bien conocido, la función de log-verosimilitud para μ y Σ con datos completos basada en una muestra Y de tamaño n viene dada por

$$\mathcal{L}(\mu, \Sigma|Y) = -n\ln(2\pi) - \frac{1}{2}n\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^T\Sigma^{-1}(y_i - \mu).$$

Como el modelo normal pertenece a la familia exponencial, el algoritmo EM toma una forma especialmente simple. En concreto, la función de log-verosimilitud para datos completos es lineal en los datos no observados. Por ello, en la t-ésima iteración, el paso E se reduce a calcular la esperanza condicionada de la parte no observada, $E[Y_{per}|Y_{obs}, \theta^{(t)}]$, dada la parte observada Y_{obs} y una estimación $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$.

Para nuestro propósito, consideraremos una matriz $\Psi = (\psi_{ij})$, con $i = 1, \ldots, n$ y $j = 1, \ldots, D-1$, donde $\psi_{ij} = \ln(\gamma_j/x_{iD})$, siendo γ_j el límite de detección para la componente x_j. Destacar que Ψ contiene la información sobre el límite de detección en relación a los datos alr-transformados. Modificaremos el paso E del algoritmo EM para incorporar esta información, considerando la esperanza condicionada $E[Y_{per}|Y_{obs}, Y_{per} < \Psi, \theta^{(t)}]$ para la parte no observada. Entonces, en la t-ésima iteración, el algoritmo EM *modificado* reemplazará los valores en

Y mediante

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{si } y_{ij} \geq \psi_{ij} \\ E[y_{ij}|y_{i,-j}, y_{ij} < \psi_{ij}, \theta^{(t)}] & \text{si } y_{ij} < \psi_{ij}, \end{cases} \tag{586}$$

con $i = 1, \ldots, n$ y $j = 1, \ldots, D-1$, donde $y_{i,-j}$ denota el conjunto de variables observadas para el caso i de la matriz de datos. Por analogía con los modelos estudiados en [5], bajo normalidad, la esperanza en (586) se obtiene mediante

$$E[y_j|y_{-j}, y_j < \psi_j] = \frac{1}{P[y_j < \psi_j]} \int_{-\infty}^{\psi_j} y_j (2\pi\sigma_j^2)^{-1/2} \exp -\frac{1}{2\sigma_j^2}(y_j - y_{-j}^T\beta)^2 dy_j,$$

donde σ_j^2 denota la varianza de y_j y β es el vector de coeficientes de la regresión lineal de y_j sobre y_{-j}. Algunos cálculos nos conducen a la expresión

$$E[y_j|y_{-j}, y_j < \psi_j] = y_{-j}^T\beta - \sigma_j \frac{\phi\left(\frac{\psi_j - y_{-j}^T\beta}{\sigma_j}\right)}{\Phi\left(\frac{\psi_j - y_{-j}^T\beta}{\sigma_j}\right)}, \tag{587}$$

donde ϕ y Φ son, respectivamente, la densidad y la función de distribución de la distribución normal estándar. Desde un punto de vista computacional, dado el elevado número de ecuaciones de regresión a estimar, recurrimos a un instrumento matemático para hacerlo de forma eficiente. En particular, en este contexto es habitual utilizar el operador *sweep*.

A partir del conjunto de datos *completado* en el paso E, en el paso M se obtienen las estimaciones máximo-verosímiles ordinarias $\theta^{(t+1)}$ para la siguiente iteración. Los pasos E y M se repiten hasta alcanzar la convergencia. En nuestra implementación el algoritmo se detiene cuando $\max\{|\mu^{(t+1)} - \mu^{(t)}|, |\Sigma^{(t+1)} - \Sigma^{(t)}|\}$ es inferior al nivel de tolerancia prefijado $\varepsilon = 0.0001$. Una vez que el algoritmo converge, tomamos el último conjunto de datos completado en el paso E y le aplicamos la transformación alr inversa para llevarlo de nuevo al símplex, obteniendo así un conjunto de datos composicionales sin ceros.

Puede comprobarse que los resultados obtenidos tras la aplicación del algoritmo EM modificado no dependen del divisor seleccionado en la transformación alr. Además, los ceros por redondeo han sido reemplazados por valores por debajo del límite de detección, y el método es coherente con las operaciones básicas sobre el símplex y con la naturaleza composicional de los datos.

4. Resultados empíricos

El conjunto de datos, previamente utilizado en [14] y [17], se refiere a la composición [Al_2O_3,SiO_2,Fe_2O_3,TiO_2,H_2O,Res_6] de 332 muestras de 34 yacimientos en el depósito de bauxita de Halimba (Hungría). La sexta componente Res_6 es una parte residual de la composición, es decir, es igual a (100-(Al_2O_3+...+H_2O))%. En adelante, denotaremos con X a este conjunto de

datos y los llamaremos datos Halimba. Inicialmente X no contiene ceros y los datos están registrados en porcentajes con una cifra decimal. Para nuestros cálculos, trabajaremos con ellos en tanto por uno, con lo que $c = 1$ en este caso.

Como medidas descriptivas utilizaremos el array de variación composicional ([1]), que en el triángulo superior recoge las varianzas log-cociente, $V(\log(x_j/x_k))$, $j = 1, \ldots, 5$ y $k = j + 1, \ldots, 6$, y en el triángulo inferior las esperanzas log-cociente $E[\log(x_k/x_j)]$, $k = 1, \ldots, 5$ y $j = k + 1, \ldots, 6$; la media geométrica composicional $g(X)$ y la variabilidad total (totvar) definidas como

$$g(X) = \mathcal{C}(g_1, \ldots, g_D) \quad \text{y} \quad \text{totvar}(X) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{d}_\mathrm{a}^2(x_i, g(X)),$$

donde $g_j = (x_{1j} \cdots x_{nj})^{1/n}$ es la media geométrica de los datos de la componente x_j y $\mathrm{d_a}$ es la distancia de Aitchison (584). El Cuadro 1 recoge el valor de estas medidas con los datos Halimba.

$g(X) = (0.5644, 0.0246, 0.2421, 0.0282, 0.1242, 0.0166)$						
$\text{totvar}(X) = 0.9718$						
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.8946	0.1288	0.1793	0.0885	0.6105
SiO_2	3.1314	0	0.9095	0.9703	0.8515	0.9321
Fe_2O_3	0.8464	-2.2850	0	0.1915	0.1519	0.6194
TiO_2	2.9981	-0.1333	2.1516	0	0.2214	0.6603
H_2O	1.5140	-1.6174	0.6676	-1.4841	0	0.5566
Res_6	3.5284	0.3970	2.6819	0.5303	2.0144	0

Cuadro 1: Medidas descriptivas del conjunto de datos Halimba.

Con el fin de evaluar el rendimiento del algoritmo EM modificado, cada valor observado de X menor que $\gamma_j = 0.01$, $\forall j$, se transforma en un cero por redondeo, esto es, interpretamos que nuestro instrumento de medida no es capaz de detectar proporciones de un componente químico inferiores al 1 %. Denotaremos con X^* al conjunto de datos composicionales obtenido. El resultado es que 105 de las 332 observaciones tienen algún cero, pero el número total de ceros sólo representa un 6 % de los datos. Se observa que las componentes Al_2O_3, Fe_2O_3 y H_2O no tienen ceros en X^*, y que éstos se concentran en SiO_2 y Res_6.

En este estudio compararemos los resultados obtenidos con el reemplazamiento multiplicativo (585) y con la versión modificada del algoritmo EM que aquí se presenta, ya que son los únicos que ofrecen resultados adecuados al problema que se pretende resolver. Los resultados con el algoritmo EM estándar, el método de Sandford y el método IM se analizan en [15] y [17]. Aunque estos tres

métodos, aplicados sobre los datos alr-transformados para después devolver al símplex los datos completados, son coherentes con las operaciones básicas en el símplex, ninguno de ellos tiene en cuenta que los ceros por redondeo deben ser reemplazados por valores pequeños por debajo del límite de detección.

Primero se reemplazarán los ceros en X^* y, a continuación, se compararán las medidas descriptivas de X con las de los conjuntos de datos completados. Sea r_i la composición que reemplaza a $x_i^* \in X^*$. Como conocemos los datos reales $x_i \in X$, también podemos considerar dos medidas de distorsión:

$$\text{MSD} = \frac{\sum \mathrm{d_a}^2(x_i, r_i)}{332} \quad \text{y} \quad \text{STRESS} = \frac{\sum\limits_{i<j} (\mathrm{d_a}(x_i, x_j) - \mathrm{d_a}(r_i, r_j))^2}{\sum\limits_{i<j} \mathrm{d_a}^2(x_i, x_j)}.$$

Para utilizar el reemplazamiento multiplicativo (585) es necesario imponer un valor de δ_j. Siguiendo a [14], tenemos que los mejores resultados se obtienen con $\delta_j = 0.0065$, $\forall j$, (65 % del umbral de detección $\gamma_j = 0.01$). En el Cuadro 2 tenemos las medidas descriptivas obtenidas.

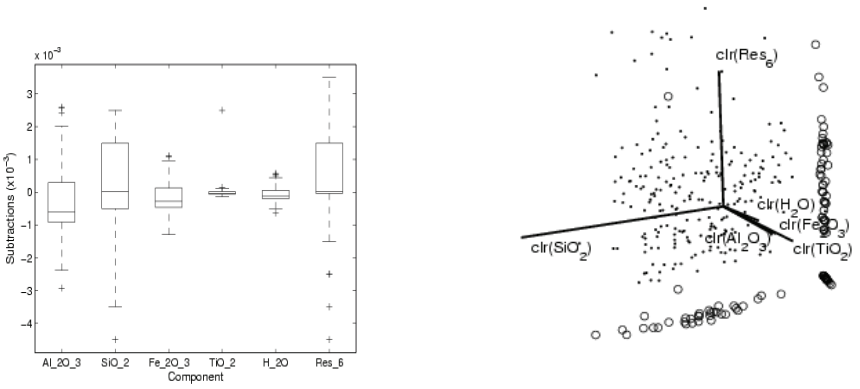
$g(X) = (0.5645, 0.0246, 0.2421, 0.0281, 0.1242, 0.0164)$						
$\text{totvar}(X) = 0.9602$						
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.8829	**0.1288**	0.1864	**0.0885**	0.6166
SiO_2	3.1318	0	0.8984	0.9598	0.8397	0.9153
Fe_2O_3	**0.8464**	-2.2853	0	0.1979	**0.1519**	0.6246
TiO_2	2.9990	-0.1327	2.1526	0	0.2273	0.6727
H_2O	**1.5140**	-1.6178	**0.6676**	-1.4850	0	0.5612
Res_6	3.5411	0.4093	2.6947	0.5421	2.0271	0
MSD: 0.0328						
STRESS: 0.0208						

Cuadro 2: Medidas descriptivas con reemplazamiento multiplicativo.

Los valores de MSD y STRESS están razonablemente cerca de cero, con lo que podemos concluir que la distorsión es mínima. La misma conclusión se obtiene cuando comparamos las medidas descriptivas con las verdaderas (cuadro 1). Se observa que la estructura relativa de las componentes que no contienen valores cero se preserva (valores en negrita en Cuadro 2).

Mediante un diagrama de cajas, analizamos las diferencias entre los valores verdaderos en X y los valores correspondientes en el conjunto de datos completado. En la Figura 1A se observa que la distorsión introducida por el reemplazamiento multiplicativo no es grande (10^{-3} en el eje Y) y es simétrica, esto es, los valores

verdaderos han sido reemplazados por valores grandes o pequeños en, aproximadamente, la misma proporción. Sin embargo, en la Figura 1B, un biplot composicional ([3]) nos permite apreciar cómo la similaridad entre las observaciones completadas se exagera (círculos vacíos en Fig. 1B). Por otra parte, con este reemplazamiento no paramétrico será necesario analizar la sensibilidad de los resultados al valor considerado de δ_j. Este análisis de sensibilidad se realiza en [14].



(A) (B)

Figura 1: Reemplazamiento multiplicativo: (A) Boxplot de las diferencias entre los datos en X y los datos completados; (B) Biplot composicional de los datos completados.

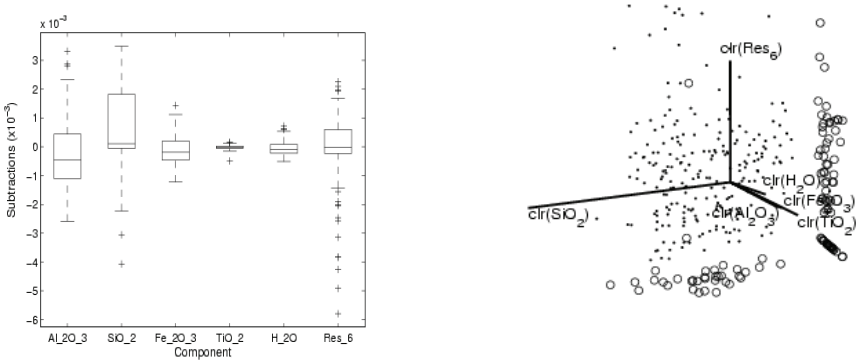
El Cuadro 3 muestra las medidas descriptivas del conjunto de datos completados mediante la aplicación del algoritmo EM modificado. Vemos que se preserva la estructura relativa de las partes sin ceros y que las varianzas log-cociente están algo sobreestimadas (véase Cuadro 1). Las medidas MSD y STRESS toman valores cercanos a cero. La Figura 2A es similar a la Figura 1A obtenida con el reemplazamiento multiplicativo. A diferencia de los métodos paramétricos comentados anteriormente, los ceros son reemplazados adecuadamente por valores por debajo del límite de detección y la distorsión es mínima. En la Figura 2B vemos que las observaciones completadas (círculos vacíos en Fig. 2B) son menos parecidas entre sí que las obtenidas con el reemplazamiento multiplicativo,

aunque siguen siendo similares. Esto puede explicarse por la pequeña proporción de ceros que hay en X^*. En este caso los resultados del reemplazamiento multiplicativo y del algoritmo EM modificado son bastante parecidos, aún así, con el EM modificado se evita el posterior análisis de sensibilidad respecto a δ_j en (585).

Efectivamente, si incrementamos el umbral de detección γ_j del 1 % al 1.5 %, aumentando así el número de ceros en X^* del 6 % al 11.78 %, el Cuadro 4 nos

$g(X) = (0.5646, 0.0241, 0.2422, 0.0282, 0.1242, 0.0168)$						
totvar$(X) = 0.9734$						
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.9171	**0.1288**	0.1783	**0.0885**	0.5809
SiO_2	3.1537	0	0.9334	0.9927	0.8737	0.9217
Fe_2O_3	**0.8464**	-2.3072	0	0.1906	**0.1519**	0.5903
TiO_2	2.9979	-0.1558	2.1515	0	0.2206	0.6373
H_2O	**1.5140**	-1.6397	**0.6676**	-1.4839	0	0.5231
Res_6	3.5165	0.3628	2.6700	0.5186	2.0025	0
MSD: 0.0346						
STRESS: 0.0226						

Cuadro 3: Medidas descriptivas con el algoritmo EM modificado.



(A) (B)

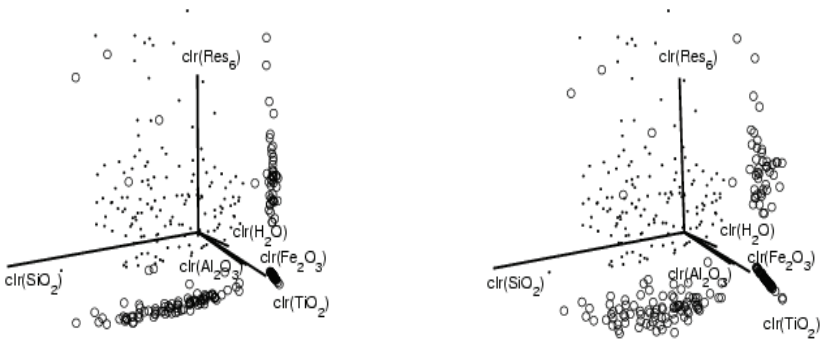
Figura 2: Algoritmo EM modificado: (A) Boxplot de las diferencias entre los datos en X y los datos completados; (B) Biplot composicional de los datos completados.

muestra que el algoritmo EM modificado mejora las medidas descriptivas y de distorsión del reemplazamiento multiplicativo (con $\delta_j = 0.00975$, 65 % del umbral de detección $\gamma_j = 0.015$). Si comparamos las observaciones completadas con el reemplazamiento multiplicativo (Fig. 3A) con las completadas con el algoritmo EM modificado (Fig. 3B) vemos que la variabilidad de estas últimas se incrementa. Todo esto sugiere que el EM modificado será preferible cuando aumenta el número de ceros por redondeo en una muestra.

Finalmente, aplicamos el reemplazamiento multiplicativo y el algoritmo EM a dos conjuntos de datos extremos: uno con 30 observaciones con 4 partes y 5 ceros (datos *Foraminiferal*), y otro de 1281 observaciones con 8 partes y 2852

Multipl.: $g(X) = (0.5633, 0.0264, 0.2416, 0.0280, 0.1239, 0.0167)$		
EM Mod: $g(X) = (0.5641, 0.0254, 0.2420, 0.0283, 0.1241, 0.0162)$		
	Reempl. Multiplicativo	EM Modificado
totvar(X)	0.7714	0.8550
MSD	0.0903	0.0772
STRESS	0.0563	0.0448

Cuadro 4: Medidas descriptivas de los datos completados con $\gamma_j = 0.015$.



(A) (B)

Figura 3: Biplots composicionales de los datos completados con $\gamma_j = 0.015$: (A) reemplazamiento multiplicativo; (B) algoritmo EM modificado.

ceros (datos *Darss Sill*). Los límites de detección son, respectivamente 0.01 y 0.001. El Cuadro 5 resume las medidas descriptivas obtenidas. Las medidas MSD y STRESS son utilizadas en este caso para comparar los dos conjuntos de datos completados. Los resultados ponen de nuevo en evidencia que con pocos ceros (datos Foraminiferal) ambos métodos de tratamiento de los ceros producen resultados muy similares, sin embargo, con muchos ceros (datos Darss Sill) el comportamiento es muy diferente (véase medidas MSD y STRESS en Cuadro 5). Es importante destacar que la variabilidad en los datos completados por el EM modificado es mayor. Este diferente patrón se muestra en la Figura 4, donde se representan las observaciones completadas en el conjunto Darss Sill mediante reemplazamiento multiplicativo y mediante el algoritmo EM modificado. Este biplot composicional explica el 92.3 % de la variabilidad total, que parece deberse principalmente al diferente comportamiento de los dos métodos respecto a la primera componente de la composición. Como se observa, el reemplazamiento multiplicativo exagera la similaridad de las observaciones completadas. Destacar que los dos métodos sólo tienen el mismo comportamiento en aquellas observaciones sin ceros en la primera, segunda y tercera componente, donde los cuadrados y las cruces se superponen.

		Datos Foraminiferal	Datos Darss Sill
totvar(X)	RM	2.4499	16.9151
	EMm	2.6756	64.4153
MSD		0.0260	221.0945
STRESS		0.0072	0.3146

	g(X)
F-RM	(0.6830,0.2465,0.0400,0.0305)
F-EMm	(0.6847,0.2471,0.0379,0.0303)
D-RM	(0.0017,0.0033,0.0079,0.0338,0.3418,0.5540,0.0484,0.0090)
D-EMm	(7×10^{-10},0.0003,0.0038,0.0330,0.3448,0.5603,0.0490,0.0089)

Cuadro 5: Medidas descriptivas de los datos Foraminiferal y Darss Sill completados. (F-RM: Foraminiferal-reempl. mult., F-EMm: Foraminiferal-EM modif., D-RM: Darss Sill-reempl. mult., D-EMm: Darss Sill-EM modif.).

5. Conclusiones

Las dificultades de los aparatos de medida para detectar proporciones por debajo de un cierto umbral implica la aparición de valores nulos en conjuntos de datos composicionales. Su presencia impide realizar un análisis de los datos basado en la metodología log-cociente, por lo que se requieren métodos adecuados para salvar este problema.

En este trabajo introducimos una modificación del algoritmo EM que, en combinación con la transformación log-cociente aditiva, reemplaza los ceros por valores adecuados por debajo del límite de detección, preservando la estructura de covarianzas de las subcomposiciones sin valores nulos y la coherencia con las operaciones básicas en el símplex. Los valores se obtiene haciendo uso de la información contenida en los datos observados, y los resultados no dependen del denominador utilizado en la transformación alr.

Cuando el número de ceros es pequeño, su comportamiento es similar al del reemplazamiento multiplicativo no paramétrico. Sin embargo, cuando el número de ceros es elevado, el algoritmo EM modificado proporciona mejores resultados.

6. Bibliografía

[1] Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London. Reprinted in 2003 by Blackburn Press.

[2] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., y Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Math. Geol.* 32, 271-275.

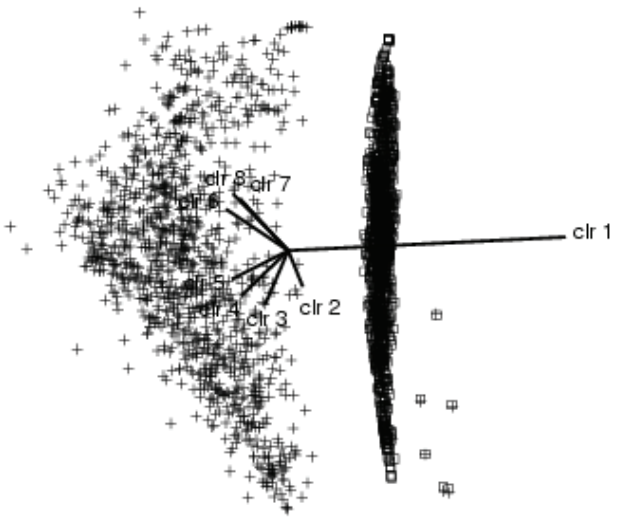


Figura 4: Biplot composicional de las observaciones completadas en el conjunto de datos Darss Sill. (□) para reemplazamiento multiplicativo; (+) para el algoritmo EM modificado.

[3] Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Appl. Stat.* 51, 375-392.

[4] Aitchison, J. and Kay, J. W. (2003). Possible solutions of some essential zero problems in compositional data analysis. En: Thió-Henestrosa and Martín-Fernández (eds.) *CoDaWork'03, Proceedings*, Universitat de Girona, CD-ROM, ISBN: 84-8458-111-X, disponible en http://ima.udg.es/Activitats/CoDaWork03/.

[5] Amemiya, T. (1984). Tobit models: a survey. *J. Econometrics* 24, 3-61.

[6] Bacon-Shone, J. (2003). Modelling structural zeros in compositional data. En: Thió-Henestrosa and Martín-Fernández (eds.) *CoDaWork'03, Proceedings*, Universitat de Girona, CD-ROM, ISBN: 84-8458-111-X, disponible en http://ima.udg.es/Activitats/CoDaWork03/.

[7] Buccianti, A. and Rosso, F. (1999). A new approach to the statistical analy-

sis of compositional (closed) data with observations below the detection limit. *Geoinformatica* 3, 17-31.

[8] Dempster, A. P., Laird N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc.*, Series B, 39, 1-38.

[9] Fry, J. M., Fry, T. R. L., and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Appl. Economics* 32, 953-959.

[10] Gómez-García, J., Palarea-Albaladejo, J., y Martín-Fernández, J. A., (2006). Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. *Revista Estadística Española* 162, (en prensa).

[11] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economics and Social Measurement* 5, 475-492.

[12] Little, R. J. A. and Rubin, D.B. (2002).*Statistical analysis with missing data*. Wiley & Sons, New York.

[13] Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2000). Zero replacement in compositional data sets. En: H. Kiers, J. Rasson, P. Groenen and M. Shader (eds.) *Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000)*, Berlin, Springer-Verlag, 155-160.

[14] Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Math. Geol.* 35, 253-278.

[15] Martín-Fernández, J. A., Palarea-Albaladejo, J. and Gómez-García, J. (2003). Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach. En: Thió-Henestrosa and Martín-Fernández (eds.) *CoDaWork'03, Proceedings*, Universitat de Girona, CD-ROM, ISBN: 84-8458-111-X, disponible en http://ima.udg.es/Activitats/CoDaWork03/.

[16] McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley, New York.

[17] Palarea-Albaladejo, J., Martín-Fernández, J.A. y Gómez García, J. (2004). Dificultades en la aplicación de técnicas paramétricas para el problema de los ceros composicionales: un estudio de caso. *Actas del XXVIII Congreso*

Nacional de Estadística e Investigación Operativa, Universidad de Cádiz, CD-ROM, ISBN: 84-609-0438-5.

[18] Rubin, D. B. (1987). *Multiple imputation for nonresponse in survey*. Wiley & Sons.

[19] Sandford, R. F., Pierson, C. T. and Crovelli, R. A. (1993). An objective replacement method for censored geochemical data. *Math. Geol.* 25, 59-80.

[20] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* 11, 95-103.