27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8–11 de abril de 2003

EL PROBLEMA DEL ANÁLISIS ESTADÍSTICO DE DATOS COMPOSICIONALES

G. Mateu-Figueras, J.A.Martín-Fernández, V. Pawlowsky-Glahn y C. Barceló-Vidal

Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona, 17071 Girona
E-mail: gloria.mateu@udg.es
josepantoni.martin@udg.es
vera.pawlowsky@udg.es
carles.barcelo@udg.es

RESUMEN

El problema del análisis estadístico de datos composicionales ha sido y es una fuente de preocupación para muchos científicos desde que en 1897 Karl Pearson pusiera de manifiesto la inadecuación de los métodos estadísticos clásicos para el estudio de los mismos. Los datos composicionales son realizaciones de vectores aleatorios de suma constante. Es incuestionable la frecuencia con que aparecen medidas de esta índole en las ciencias aplicadas —ciencias de la tierra (geoquímica, petrología, ...), biología, química, ciencias ambientales, economía, medicina, sociología, ingeniería— y, por ende, es incuestionable el interés por disponer de herramientas adecuadas para su análisis.

Palabras y frases clave: símplex, diagrama ternario, operador clausura, logcociente.
Clasificación AMS: 62Pxx

1 Los datos composicionales

Cualquier vector $\mathbf{x}$, cuyas componentes representan partes de un todo, está sujeto a la restricción de que la suma de sus componentes sea la unidad, o en el caso general, una constante.

Definición 1 Un *dato composicional* $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ con D partes, es un vector con componentes estrictamente positivas, tal que la suma de todas ellas es

igual a una constante k. Su espacio muestral es el *símplex* $\mathcal{S}^D$, definido por

$$\mathcal{S}^D = \left\{(x_1, x_2, \ldots, x_D)' : x_i > 0; \sum_{i=1}^{D} x_i = k\right\}.$$

Para el caso $D = 3$, el símplex $\mathcal{S}^3$ suele representarse mediante el diagrama ternario, triángulo equilátero de altura k (véase la figura 1). Existe una correspondencia biunívoca entre los datos composiciones con 3 partes y los puntos del diagrama ternario. Un dato composicional $\mathbf{x} = (x_1, x_2, x_3)'$ se corresponde con el punto que dista x_1, x_2 y x_3, respectivamente, de los lados opuestos a los vértices 1, 2 y 3.

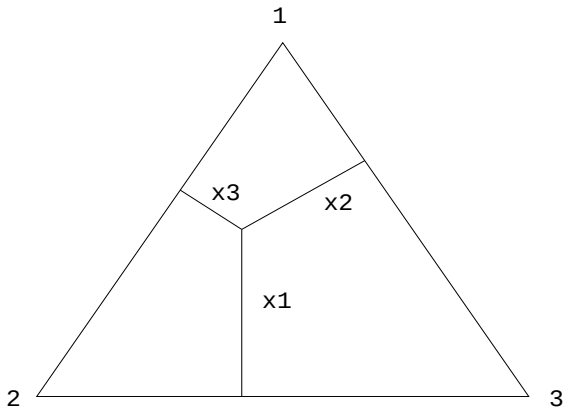

Figura 1: *Representación de un dato composicional* $(x_1, x_2, x_3)'$ en el símplex $\mathcal{S}^3$.

En su formulación estadística, los datos composicionales son realizaciones de una composición, vector aleatorio cuyo recorrido está en $\mathcal{S}^D$. En estos casos se plantea la necesidad de aplicar técnicas estadísticas para el estudio e interpretación de este tipo de datos.

Éstos aparecen en áreas muy diversas. Así, por ejemplo, en Geología al expresar la composición geoquímica de una roca como el porcentaje en peso de los óxidos más abundantes. Encontramos en la literatura numerosos trabajos con diferentes objetivos. Por ejemplo, Thomas y Aitchison (1998) estudian qué óxidos son más efectivos a la hora de discriminar entre dos tipos de calizas. Tolosana et al (2002) presentan un análisis discriminante de basaltos y rocas afines basándose en los elementos traza presentes en los mismos. Buccianti et al (2002) utilizan el porcentaje de los componentes químicos de los gases en fumarolas de volcanes para estudiar las constantes de equilibrio en diferentes reacciones químicas. Weltje (2001) construye regiones de confianza en el símplex para rocas detríticas. También es frecuente encontrar datos composicionales de naturaleza granulométrica provenientes de sedimentos marinos. En estos casos se separan las componentes arenosas de los sedimentos según el tamaño del grano y se mide el porcentaje en peso de cada tamaño de grano respecto del peso total de la muestra recogida. Martín-Fernández et al (1997) analizan la base de datos Darss Sill, que contiene muestras de sedimentos en diferentes puntos

geográficos del fondo del Mar Báltico, con el objetivo de realizar un mapa del fondo marino con diferentes zonas según el tipo de sedimento.
En Economía encontramos datos de tipo composicional al estudiar, por ejemplo, la distribución de un presupuesto entre las diferentes partidas (Andries, 1999). También en Arqueometria aparecen datos composicionales al estudiar, por ejemplo, la composición geoquímica de las cerámicas procedentes de excavaciones arqueológicas con el objetivo de determinar su origen (Buxeda, 1999).
Antes de indicar la problemática específica que comporta el análisis estadístico de los datos composicionales introducimos dos definiciones de gran importancia: el operador clausura y la noción de subcomposición.
A partir de un vector con componentes positivas siempre podemos obtener un dato composicional de $\mathcal{S}^D$. Basta con dividir cada una de sus componentes por la suma de todas ellas. Este hecho conduce a dar las definiciones siguientes.

Definición 2 El *operador clausura* $\mathcal{C}$ es una transformación que hace corresponder a cada vector $\mathbf{w} = (w_1, w_2, \dots, w_D)'$ de $\mathbb{R}_+^D$ su dato composicional asociado $\mathcal{C}(\mathbf{w}) = k\mathbf{w}/(w_1 + w_2 + \cdots + w_D)$ de $\mathcal{S}^D$, con k la constante de clausura.

En algunos casos puede interesarnos analizar únicamente el valor de las magnitudes relativas de un subconjunto de partes de unos datos composicionales. Es pues necesario disponer de un procedimiento de formación de subcomposiciones.

Definición 3 Si S es un subconjunto cualquiera de las partes $1, 2, \dots, D$ de un dato composicional $\mathbf{x} \in \mathcal{S}^D$ y $\mathbf{x}_S$ simboliza el subvector formado por las correspondientes partes de $\mathbf{x}$, entonces $\mathbf{s} = \mathcal{C}(\mathbf{x}_S)$ recibe el nombre de *subcomposición* de las S partes de $\mathbf{x}$.

Si bien la formación de una subcomposición es en esencia una trasformación de $\mathcal{S}^D$ a un símplex de dimensión inferior, obsérvese que tiene la buena propiedad de conservar la magnitud relativa entre las partes.

2 Principales problemas

La restricción de la suma constante ha sido considerada la fuente de todos los problemas pues impide la aplicación de los procedimientos estadísticos habituales que se utilizan para datos que no presentan esta restricción. Nótese, por ejemplo, que el cambio en una de las partes provoca el cambio en como mínimo una de las otras partes.
Una de las dificultades más relevantes es la imposibilidad de interpretar correctamente las covarianzas y los coeficientes de correlación. En 1897 Pearson ya puso de manifiesto esta dificultad en su artículo "On a form of spurious correlation", donde advertía a la comunidad científica de la existencia de una falsa correlación entre las partes una composición. La matriz de correlaciones habitual no puede anali-

zarse en el estudio de vectores de suma constante porque presenta necesariamente correlaciones negativas no nulas, determinadas precisamente por la mencionada restricción. Pearson calificó estas correlaciones como espurias ya que falsean la imagen de las relaciones de dependencia y pueden conducir a interpretaciones erróneas. En particular, si analizamos la matriz de covarianzas usual entre las partes de una composición, $\mathbf{K} = \{\text{cov}(x_i, x_j) : i, j = 1, 2, \dots, D\}$, obtenemos que

$$\text{cov}(x_i, x_1) + \text{cov}(x_i, x_2) + \cdots + \text{cov}(x_i, x_D) = 0 \qquad i = 1, 2, \dots, D$$

a causa de la restricción $\sum_{i=1}^{D} x_i = k$. Sabemos que $\text{cov}(x_i, x_i) = \text{var}(x_i) > 0$, excepto en la situación trivial que la parte x_i sea una constante. Este hecho provoca que necesariamente deba haber una covarianza $\text{cov}(x_i, x_j)$ $(i \neq j)$ de signo negativo. Vemos pues que estas covarianzas no son libres de tomar cualquier valor. Esto invalida la interpretación habitual de las covarianzas, y por ende de las correlaciones, pues a priori suponemos que deberían poder adquirir libremente valores nulos, positivos o negativos. Por el mismo motivo, el hecho que el coeficiente de correlación entre dos partes cualesquiera de una composición sea igual a 0 no puede interpretarse, como es habitual, como indicio de independencia entre ambas partes.
Encontramos otra incoherencia en relación a las subcomposiciones. Intuitivamente esperaríamos encontrar una cierta relación entre la matriz de covarianzas de una subcomposición y la de la composición de procedencia. Sin embargo, no existe ninguna relación. Es posible, incluso, que dos partes estén correlacionadas positivamente en el seno de una composición y en cambio pasen a tener correlación negativa al analizarlas como partes integrantes de una subcomposición. En otras palabras, el signo de la covarianza entre dos partes puede ir fluctuando cuando nos movemos de la composición inicial a subcomposiciones de dimensión cada vez más pequeña. Aitchison (1997) nos muestra este problema mediante un sencillo ejemplo. Consideremos dos científicos A y B que analizan muestras de tierra divididas en partes iguales. Para cada una de las partes de la muestra, el científico A calcula un dato composicional de 4 partes (animal, vegetal, mineral, agua). El científico B procede primero al secado de las muestras eliminando el agua de las mismas y calcula a continuación datos composicionales de 3 partes (animal, vegetal, mineral). Està claro que, los datos composicionales de 3 partes del científico B son subcomposiciones de los datos composicionales del científico A. Sin duda, es obvio que cualquier conclusión sobre las partes comunes debería ser la misma para los dos científicos. Supongamos que se han obtenido los siguientes datos:

$(x_1; x_2; x_3; x_4)$	$(s_1; s_2; s_3)$
$(0.1; 0.2; 0.1; 0.6)$	$(0.25; 0.50; 0.25)$
$(0.2; 0.1; 0.1; 0.6)$	$(0.50; 0.25; 0.25)$
$(0.3; 0.3; 0.2; 0.2)$	$(0.375; 0.375; 0.25)$

Cuando el científico A calcula la correlación entre las partes animal y vegetal, obtiene $\text{corr}(x_1, x_2) = 0.5$ mientras que el científico B obtiene $\text{corr}(s_1, s_2) = -1$.

En general tampoco es correcto aplicar las operaciones clásicas del espacio real vectorial a los datos composicionales. Martín-Fernández et al (1998) muestran un ejemplo que pone en evidencia que la distancia euclidiana no es una medida de diferencia adecuada entre datos composicionales. Consideremos, por ejemplo, dos empresas A y B de las cuales se recoge la proporción de días anuales en los que se paró la producción, la proporción de días en los que se redujo la producción a la mitad y por último, la proporción de días normales. Estos 3 valores forman un dato composicional de 3 partes. Supongamos que los resultados correspondientes a los años 1999 y 2000 para la empresa A son $A_{99} = (0.2; 0.1; 0.7)$ y $A_{00} = (0.1; 0.2; 0.7)$ y para la empresa B son $B_{99} = (0.4; 0.3; 0.3)$ y $B_{00} = (0.3; 0.4; 0.3)$. Si comparamos A_{99} y A_{00} vemos que los días de paro de producción del año 2000 se han reducido a la mitad en relación al año 1999. Esto no se observa en la empresa B. Se concluye pues que hay una diferencia más grande entre los dos datos de la empresa A que entre los dos datos de la empresa B. No obstante, con la distancia euclidiana habitual obtenemos la igualdad, es decir, $d_{eu}(A_{99}, A_{00}) = d_{eu}(B_{99}, B_{00})$. De ello se desprende que la distancia euclidiana no tiene sentido cuando trabajamos con datos composicionales. Esto tiene consecuencias estadísticas importantes porque existen multitud de conceptos y técnicas estadísticas que se fundamentan de forma más o menos explícita en la distancia euclidiana.
Otra de las dificultades importantes es la falta de familias paramétricas suficientemente flexibles para modelar los conjuntos de datos composicionales. Las distribuciones de Dirichlet y sus generalizaciones se obtienen mediante la clausura de vectores aleatorios con componentes independientes. Como consecuencia, sus partes son prácticamente independientes, puesto que su correlación está únicamente motivada por el hecho de haber dividido todas sus componentes por la suma de éstas. Esto impide su uso en la modelización de fenómenos con relaciones de dependencia no inducidas por la suma constante.
Todas estas dificultades ponen de relieve la necesidad de replantear el análisis estadístico de los datos composicionales.

3 Principios fundamentales

Muchos han sido los autores que han intentado afrontar los problemas del análisis estadístico de los datos composicionales. La solución no aparece hasta 1982 cuando Aitchison presenta, por primera vez, una forma de evitar la restricción de la suma constante.
Aitchison argumenta que todas las dificultades de interpretación vienen motivadas por centrar nuestra atención en las magnitudes absolutas de las partes $x_1, x_2, \ldots, x_D$ de una composición. Nuestra atención debe centrarse en la magnitud relativa de las partes, es decir, en los cocientes x_i/x_j $(i, j = 1, 2, \ldots, D; i \neq j)$. Por lo tanto diremos que un problema es composicional cuando reconozcamos que el valor en términos absolutos de las partes es irrelevante. Esto es un principio fundamental

del análisis de datos composicionales que Aitchison (1997) denomina *invariancia por cambios de escala*. Una importante consecuencia que se deduce de este principio es que

"*cualquier función aplicada sobre datos composicionales debe poder expresarse en términos de cocientes entre sus partes o componentes*".

Trabajando con los cocientes desaparecen los problemas de las correlaciones espurias. Por otra parte, la magnitud relativa entre las partes de una subcomposicion no cambia en relación a la magnitud relativa entre las partes de la composición original, es decir, $s_i/s_j = x_i/x_j$. Por lo tanto, cuando trabajamos con funciones invariantes por cambios de escala, somos "subcomposicionalmente"coherentes.

La metodología de Aitchison se basa en la trasformación de los datos composicionales al espacio real multivariante. Notemos que el espacio muestral para los cocientes entre las partes es el octante positivo de $\mathbb{R}^{D-1}$. Si tomamos los logaritmos de los cocientes, el espacio final es todo $\mathbb{R}^{D-1}$ y por lo tanto podemos aplicar cualquier técnica estadística clásica. Esta estrategia se remonta al trabajo de McAlister (1879) quien desarrolló los fundamentos de la ley lognormal univariante tomando el logaritmo de los datos.

Tenemos diversas posibilidades de transformación de los datos, todas ellas basadas en los logaritmos de cocientes entre las partes de un dato composicional

Definición 4 La trasformación logcociente aditiva (alr) de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{y} \in \mathbb{R}^{D-1}$ se define como $\mathbf{y} = \text{alr}(\mathbf{x}) = \big(\ln(x_1/x_D), \ln(x_2/x_D), \ldots, \ln(x_{D-1}/x_D)\big)'$.

Esta trasformación es biyectiva pero no es simétrica en las partes de $\mathbf{x}$ ya que la parte del denominador adquiere un protagonismo especial respecto al resto. Este hecho condujo a Aitchison (1986) a introducir la transformación logcociente centrada

Definición 5 La trasformación logcociente centrada (clr) de $\mathbf{x} \in \mathcal{S}^D$ a $\mathbf{z} \in \mathbb{R}^D$ se define como $\mathbf{z} = \text{clr}(\mathbf{x}) = \big(\ln(x_1/g(\mathbf{x})), \ln(x_2/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x}))\big)'$, donde $g(\mathbf{x})$ es la media geométrica de las D partes de $\mathbf{x}$.

Esta trasformación es biyectiva, y simétrica entre las partes. Su imagen es el hiperplano de $\mathbb{R}^D$ que pasa por el origen y es ortogonal al vector de unidades, es decir, la suma de las componentes del vector trasformado es igual a cero. Nos encontramos pues ante una nueva dificultad ya que la matriz de covarianzas del vector trasformado será singular.

Por esta razón Aitchison aplica una estrategia doble en sus trabajos. En las aplicaciones que exigen simetría en el tratamiento de sus componentes utiliza la trasformación clr. Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, utiliza la trasformación alr. De esta forma evitamos trabajar con distribuciones degeneradas.

Los desarrollos de la propuesta de Aitchison fueron publicados en diferentes artículos, resumidos en su monografía (Aitchison, 1986). A esta monografía le siguen numerosos trabajos entre los cuales destacamos Aitchison (1997, 2001). Siguiendo la línea

de Aitchison se han adaptado al análisis de datos composicionales diversas técnicas estadísticas, entre ellas, la predicción de observaciones multivariantes con dependencia espacial o cokrigeado (Pawlowsky, 1986, 2003), el análisis discriminante (Barceló, 1996) o la clasificación no paramétrica (Martín-Fernández, 2001). Por otra parte esta metodología ha permitido ampliar las familias de distribuciones sobre el símplex. Destacamos el modelo normal logístico aditivo (Aitchison, 1996), el modelo normal asimétrico logístico aditivo (Mateu-Figueras et al, 1998) o los modelos basados en las trasformaciones Box-Cox (Barceló, 1996).
Actualmente se ha desarrollado el fundamento matemático del símplex (Barceló, 2001). Sabemos que $\mathcal{S}^D$ tiene una estructura de espacio vectorial euclidiano, con unas operaciones, un producto escalar y una distancia propias y diferentes a las clásicas del espacio real (Pawlowsky y Egozcue, 2001). Esta estructura de geometría euclidiana ha permitido reformular conceptos fundamentales de los estimadores y obtener sus primeras propiedades con referencia al sesgo o a la variancia (Pawlowsky y Egozcue, 2002). En esta línea, han surgido los primeros trabajos en relación al estudio de las técnicas de regresión lineal multivariante sobre datos composicionales (Daunis-i-Estadella et al, 2002).

4 Conclusiones

Los datos composicionales aparecen frecuentemente y en disciplinas muy dispares. Es, por lo tanto, necesario disponer de herramientas adecuadas para su análisis estadístico.
La restricción que la suma de las partes de un dato composicional sea igual a una constante provoca la inadecuación de los métodos estadísticos clásicos. En consecuencia, es necesario desarrollar nuevas metodologías compatibles con el carácter composicional de los datos. Estas se fundamentan en la línea iniciada por Aitchison (1982) basada en la trasformación, mediante logcocientes, de los datos composicionales al espacio real multivariante.

Referencias

Aitchison, J. (1982): The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc. B*, 44, 2, 139–177.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman and Hall, London (UK). 416 p

Aitchison, J. (1997): The one-hour course in compositional data analysis or compositional data analysis is simple. *Proceedings of IAMG'97*, Barcelona (E), Vol 1, 3-35.

Aitchison, J. (2001): Simplicial inference. *Algebraic Methods in Statistics, Contemporary Mathematics Series of the American Mathematical Society*, 287, 1-22.

Andries van der Ark, L. (1999). Contributions to latent budget analysis: a tool for the analysis of compositional data, Ph. D. thesis, Universiteit Utrecht, 221 p.

Barceló-Vidal, C. (1996): Mixturas de datos composicionales. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E) 291 p.

Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. (2001): Mathematical foundations of compositional data analysis. *Proceedings of IAMG'01*, Cancun (Mexico), CD-ROM, 20 p.

Buccianti, A., Montegrossi, G., Tassi, F. and Vaselli, O. (2002): Log-contrast analysis of volcanic fluid composition: a way to check equilibrium conditions?. *Terra Nostra, special issue: IAMG'02*, vol 3, 405–410.

Buxeda, J., (1999): Alteration and contamination of archaelogical ceramics: the perturbation problem. *J Archeol Sci* 26, 295-313.
Daunis-i-Estadella, J., Egozcue, J.J. and Pawlowsky-Glahn, V. (2002): Least squares regression in the simplex. *Terra Nostra, special issue: IAMG'02*,vol 3, 411–416.

Martín-Fernández, J. A. (2001): Medidas de diferencia y clasificación no paramétrica de datos composicionales, Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E) 233 p.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1997): Different classifications of the Darss Sill data set based on mixture models for compositional data. *Proceedings of IAMG'97*, vol 1, 151-158.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998): Measures of difference for compositional data and hierarchical clustering methods. *Proceedings of IAMG'98*, vol 2, 526-531.

Mateu-Figueras, G., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (19989: Modeling compositional data with multivariate skew-normal distributions. *Proceedings of IAMG'98*, vol 2, 532-537.

McAlister, D. (1879): The law of the geometric mean. *Proceedings of the Royal Society of London*, 29, 367-376.

Pawlowsky, V., 1986, Räumliche Strukturanalyse und Schätzung ortsabhängiger Kompositionen, Ph. D. thesis, Freie Universität Berlin, 180 p.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001): Geometric approach to statistical

analysis on the simplex. *SERRA*, 15, 5, 384–398.

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

Pawlowsky-Glahn, V. and Olea, R. A. (en imprenta): Geostatistical Analysis of compositional data. Oxford University Press, New York, NY (USA). 200p.

Pearson (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.

Thomas, C. W. and Aitchison, J. (1998): The use of log-ratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central scottish highlands. *Proceedings of IAMG'98*, vol 2, 549-554.

Tolosana-Delgado, R., Palomera-Roman, R. Gimeno-Torrente, D. Pawlowsky-Glahn, V. and Thió-Henestrosa, S.(2002): A first approach to classification of basalts using trace elements. *Terra Nostra, special issue: IAMG'02*,vol 3, 435–440.

Weltje, G.J. (2002): Quantitative analysis of detrical modes: statistically rigorius confidence regions in ternary diagrams and their use in sedimentary petrology. *Earth-Science Reviews*, 57, 211-253