MODELIZACIÓN Y ANÁLISIS DE DATOS SOBRE PROPORCIONES

JAVIER PALAREA ALBALADEJO
e-mail: jpalarea@pdi.ucam.edu
Departamento de Informática de Sistemas
UNIVERSIDAD CATÓLICA SAN ANTONIO

JOSEP ANTONI MARTIN FERNÁNDEZ
e-mail: josepantoni.martin@udg.edu
Departamento de Informática y Matemática Aplicada
UNIVERSIDAD DE GIRONA

JUAN GÓMEZ GARCÍA
e-mail: jgomezg@um.es
Departamento de Métodos Cuantitativos para la Economía
UNIVERSIDAD DE MURCIA

Área temática: Métodos Cuantitativos.

Resumen

En determinados ámbitos es frecuente trabajar con datos que representan partes de un total: proporciones, porcentajes, partes por millón, o similar. Es el caso de, por ejemplo, la distribución del presupuesto familiar, la composición de una cartera de inversión, el empleo del tiempo diario en distintas actividades, la distribución de ventas en distintas regiones, y otros. Las restricciones de no negatividad y de suma constante que caracterizan este tipo de datos implican que las técnicas multivariantes habitualmente utilizadas no son adecuadas para su análisis y modelización. La cuestión clave es que la geometría del espacio muestral sobre el que se definen, el símplex, es diferente de la clásica geometría Euclídea del espacio real. En el presente trabajo revisaremos los fundamentos teóricos, la geometría del símplex, la metodología log-cociente, e importantes aspectos prácticos como el problema de las componentes nulas. El apartado teórico se acompaña de ejemplos que motivan la exposición y ponen de relieve la naturaleza del problema que se está tratando. Finalmente, se ilustra la aplicación práctica de las ideas y métodos estudiados a través de ejemplos del campo económico.
Palabras clave: Proporciones, Datos Composicionales, Símplex, Análisis Log-cociente.

Abstract

In some scopes analysts frequently work with data that represent parts of a whole: proportions, percentages, parts per million, or similar. This is the case of household budgets, investment portfolios, time budgets, sales by region, and others. The no-negativity and constant-sum constraints characterizing this kind of data imply that common multivariate techniques are not suitable for modelling purposes. The key feature is that the geometry of the sample space on which data are defined, the simplex, differs from the classical Euclidean geometry of real space. In this paper, the theoretical foundations, the geometry of the simplex, the log-ratio methodology and some other relevant practical issues will be reviewed. The theoretical aspects are illustrated by examples which motivate the exposition. Finally, some applications from the economic field are presented.
Key words: Proportions, Compositional Data, Simplex, Log-ratio Analysis.

1. Introducción

Ya en 1897 Karl Pearson (Pearson, 1897) advertía de las dificultades para interpretar la relación entre variables que representaran partes de un total. Habitualmente, dichas variables vienen expresadas en proporciones, porcentajes, partes por millón, o similar. Pongamos como ejemplos en el campo económico la distribución del presupuesto familiar en distintas partidas de gasto, la composición relativa de una cartera de inversión, el patrón de actividad de una cadena de producción, la composición étnica de una población, el empleo del tiempo diario en distintas actividades, la distribución de las ventas de un producto en distintas regiones, la distribución de trabajadores en distintos sectores de actividad, y otros. La característica fundamental de un vector de datos de este tipo es que la suma de sus componentes, siempre positivas, es una constante. Así, un cambio en una componente conlleva un cambio en, al menos, una de las demás componentes. Por lo que hay que replantearse, entre otros, el concepto de independencia. Además, una fila cualquiera de la matriz de covarianzas de una muestra de vectores de proporciones siempre tiene al menos un elemento negativo y su suma es igual a 0. Esto implica que la matriz de covarianzas es singular y que las correlaciones no varían libremente en el habitual intervalo [-1,1].

La cuestión clave es que la geometría del espacio muestral sobre el que se define un vector de proporciones es diferente de la clásica geometría Euclídea de R^D, por ello las técnicas multivariantes habitualmente utilizadas, y fundamentadas en esta geometría, no son directamente aplicables. Para llamar la atención del lector proponemos el siguiente ejemplo. Consideremos los patrones de actividad de dos operarios (A, B) entre tres actividades mutuamente excluyentes (X, Y, Z) en dos periodos (1, 2) de 120 minutos. La asignación de tiempo a cada actividad en el periodo 1 es [10, 80, 30] para el operario A y [20, 80, 20] para el B. En el periodo 2 estos tiempos son, respectivamente, [20, 70, 30] y [30, 70, 20]. La dedicación del operario A a la actividad X se dobla de un periodo a otro, mientras que para la Y se reduce en un 12.5% y no cambia para la Z. El operario B incrementa el tiempo dedicado a X en un 50%, mientras que se reduce en un 12.5% el dedicado a Y y tampoco se modifica el dedicado a Z. Una medida de diferencia adecuada debería detectar un mayor cambio en el patrón de actividad de A que en el de B. Sin embargo, si medimos esta diferencia entre periodos utilizando la distancia Euclídea habitual obtenemos un valor de 14.1421, el mismo para los dos operarios.

Pensemos en las consecuencias de este hecho en, por ejemplo, un análisis cluster. De manera que cualquier medida basada de forma más o menos explícita en la distancia Euclídea no se comportará bien en este contexto.

Aún así, durante el pasado siglo se han aplicado, y aún se aplican, técnicas estadísticas multivariantes para datos reales sobre este tipo de variables, sin tener en cuenta su especial naturaleza. Cómo trabajar con la restricción de suma constante ha sido durante largo tiempo un problema aparentemente irresoluble, y el hecho de no tenerla en cuenta, o de interpretarla de forma inadecuada, ha llevado a conclusiones erróneas en muchos análisis de datos. Algunas actuaciones particulares en este línea se han orientado hacia intentar salvar los problemas técnicos y no hacia un estudio ni del tipo de relación relevante entre las partes de un total ni de su peculiar espacio muestral (véase p. ej. Bohling y otros, 1998). No es hasta los años 80 con la publicación de la monografía de Aitchison (1986) cuando se dispone de una obra de referencia sobre los fundamentos teóricos y sobre una metodología específica para el análisis estadístico de tales vectores de proporciones, que desde entonces suelen denominarse *datos composicionales*.

En la sección 2 resumiremos los fundamentos sobre los que se asienta el análisis de datos composicionales. A continuación, en la sección 3, destacaremos los aspectos más relevantes de la metodología basada en transformaciones para trasladar los datos al espacio real donde poder aplicar las técnicas habituales. Al mismo tiempo, pondremos de relieve algunos de los problemas prácticos que pueden surgir. Finalmente, dedicaremos la sección 4 a destacar la aplicación de las herramientas del análisis de datos composicionales en el campo de la economía.

2. Fundamentos matemáticos

Formalmente, un dato composicional es un vector $\mathbf{x} = [x_1,\ldots,x_D]$ constituido por D partes o componentes positivas cuyo espacio muestral es el símplex S^D definido como $S^D = \{\mathbf{x} = [x_1,...,x_D] : x_1 > 0,...,x_D > 0; \sum_j x_j = c\}$. En realidad, el valor de c es irrelevante desde un punto de vista matemático, por lo que en adelante consideraremos c = 1, esto es, las componentes x_j están medidas en proporciones. Este espacio muestral tiene una estructura de espacio vectorial inducida por las operaciones *perturbación* y *potenciación*. Para cualquier par de composiciones $\mathbf{x}$ y $\mathbf{x}'$ de S^D la operación

perturbación se define como $\mathbf{x} \oplus \mathbf{x}' = C(x_1 x'_1, ..., x_D x'_D)$ y la operación de potenciación como $\alpha \otimes \mathbf{x} = C(x_1^\alpha, ..., x_D^\alpha)$, donde α es un valor real. La perturbación es una operación fundamental encargada de describir el *cambio* composicional en el símplex. El operador C se denomina *clausura* y normaliza un vector de componentes positivas dividiendo cada una por el valor de la suma de todas ellas. Para tener una idea intuitiva del significado de estas operaciones básicas en el símplex diremos que la perturbación es el equivalente a la traslación o suma en espacios reales y la potenciación es el equivalente al producto por escalar. A partir de ellas podemos definir la *diferencia* composicional como $\mathbf{x} \ominus \mathbf{x}' = \mathbf{x} \oplus (-1) \otimes \mathbf{x}' = C(x_1 / x'_1, ..., x_D / x'_D)$. El papel de marginales en el símplex lo hacen las *subcomposiciones*, que son las proyecciones del símplex S^D, el espacio de las composiciones con D partes, sobre un sub-símplex de dimensión menor, digamos S^d, obtenidas mediante la clausura de un subvector formado por d de las partes de una composición en S^D. Por ejemplo, a partir de la composición **x** = [0.2, 0.6, 0.1, 0.1] de S^4 podemos obtener la subcomposición **s** = [0.2222, 0.6667, 0.1111] de S^3 clausurando el subvector formado por las 3 primeras componentes. Notar que las operaciones elementales en espacios reales no se comportan consistentemente cuando se aplican sobre composiciones. Así, la suma o resta de composiciones no da como resultado otra composición, esto es, pueden producir un elemento que no pertenece al símplex.

Adicionalmente, es posible definir un producto interior en S^D

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \left[\frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{x'_i}{x'_j} \right], \qquad (1)$$

que confiere al símplex estructura de espacio Euclídeo (Billheimer, Guttorp y Fagan, 2001; Pawlowsky-Glahn y Egozcue, 2001, 2002). Desde un punto de vista estadístico esta estructura no es irrelevante, ya que las medidas básicas de tendencia central, variabilidad y distancia deben ser coherentes con la naturaleza de los datos. El producto interior (1) induce una distancia en S^D definida como

$$d_a(\mathbf{x}, \mathbf{x}') = \left[\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x'_i}{x'_j} \right)^2 \right]^{1/2} = \left[\sum_{i=1}^{D} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{x'_i}{g(\mathbf{x}')} \right)^2 \right]^{1/2}, \qquad (2)$$

donde $g(\mathbf{x}) = (x_1 \cdots x_D)^{1/D}$ es la media geométrica de las componentes de la composición **x**. La distancia (2) se conoce como *distancia de Aitchison* y es totalmente compatible con las operaciones básicas en el símplex y con la particular naturaleza de los datos composicionales (Aitchison y otros, 2000). La estructura de espacio Euclídeo implica que sobre el símplex pueden definirse todos los entes y conceptos geométricos usuales tales como líneas, ángulos, ortogonalidad, paralelismo, etc. Sin embargo, aunque matemáticamente equivalentes, las características, y el aspecto, de estos elementos en la geometría del símplex difieren de aquellas a las que estamos habituados en la geometría Euclídea clásica del espacio real. Para evitar confusiones nos referiremos a la geometría del símplex como *geometría de Aitchison*.

3. Análisis log-cociente

Una metodología adecuada para el análisis de datos composicionales debe tener en cuenta algunos principios lógicos necesarios y las características del símplex como espacio muestral sobre el que se definen. La idea principal es que las composiciones sólo proporcionan información sobre la magnitud relativa de sus partes, y no pueden justificarse interpretaciones que involucren a las magnitudes absolutas. Se asume que el valor de la suma de las partes es irrelevante. Por lo tanto, cualquier aseveración sobre una composición debe hacerse en términos de los cocientes entre las partes, los cuales medirán dicha relación relativa. Así, una función aplicable sobre composiciones deberá ser invariante por cambios de escala y expresable en términos de cocientes entre las partes.

Trabajar con cocientes asegura además un principio lógico básico: la *coherencia subcomposicional*. Consideremos dos analistas financieros, A y B, interesados en estudiar la relación entre las proporciones de capital invertido por los clientes en distintos productos financieros. Supongamos que existen cuatro productos, siendo x_1, x_2, x_3 y x_4 las proporciones de capital invertido en cada uno de ellos, y que el analista A tiene acceso a toda la información sobre las cantidades invertidas, trabaja con la composición $[x_1, x_2, x_3, x_4]$, mientras que el analista B sólo tiene acceso a la información sobre los tres primeros, trabaja con la subcomposición $[s_1, s_2, s_3]$. Aún así, lo lógico sería que si B calcula el coeficiente de correlación lineal habitual entre s_1 y s_2 obtenga el mismo resultado que si A lo calcula entre x_1 y x_2, las proporciones se refieren

a los mismos productos financieros. Pero esto no es así cuando la medida se calcula sobre las proporciones absolutas. Recogemos a continuación los datos que maneja cada analista sobre los mismos tres clientes:

Datos analista A: $[x_1\ x_2\ x_3\ x_4]$	Datos analista B: $[s_1\ s_2\ s_3]$
[0.2, 0.1, 0.2, 0.5]	[0.4, 0.2, 0.4]
[0.1, 0.1, 0.1, 0.7]	[0.3333, 0.3333, 0.3333]
[0.6, 0.2, 0.1, 0.1]	[0.6667, 0.2222, 0.1111]

El coeficiente de correlación lineal entre las proporciones de capital invertido en los productos 1 y 2 es 0.982 para el analista A y -0.5291 para el B. Por lo tanto, queda claro que esta medida de dependencia no es coherente cuando se aplica directamente sobre proporciones. Sin embargo, es fácil comprobar que los cocientes entre las partes x_i / x_j no varían al pasar de la composición a la subcomposición. Así, para el cliente 1, $x_1 / x_2 = s_1 / s_2 = 2$. Por lo que si trabajamos con los cocientes evitaremos el problema de la incoherencia subcomposicional.

Para destacar un poco más la conveniencia de trabajar con cocientes entre las partes, retomemos el ejemplo de la introducción sobre el patrón de actividad de los dos operarios. Vimos que la distancia Euclídea no se comportaba según lo esperado, proporcionaba la misma diferencia entre periodos para los dos operarios. En la sección 2 se define la distancia de Aitchison, la cual se deduce de la propia geometría del símplex. Si nos fijamos en la expresión (2) vemos que se expresa en términos de cocientes entre las partes. Utilizando (2) como medida de diferencia entre los patrones de actividad de cada operario en los dos periodos obtenemos un valor de 0.6276 para el operario A, y un valor de 0.3970 para el operario B. Esto es, un mayor cambio en el patrón de actividad de A que en el de B, que es lo que efectivamente ocurre.

3.1 Transformaciones log-cociente

Una vez que se ha puesto de relieve la necesidad de centrar la atención en los cocientes entre las partes surge la pregunta sobre qué tipo de cocientes utilizar. La piedra angular de la metodología propuesta por Aitchison (1986) es la transformación de una composición definida sobre S^D en un vector que involucre los cocientes entre las partes y que esté definido sobre el espacio real. Si esa transformación es biyectiva se establece

una correspondencia uno a uno entre las composiciones en el símplex y los correspondientes vectores transformados reales. De esta manera, cualquier problema que afecte a composiciones queda expresado en términos de tales vectores transformados, con lo que se tiene la posibilidad de resolverlo utilizando las técnicas multivariantes habituales en espacios reales. Aitchison propone fundamentalmente dos tipos de transformaciones: la *transformación log-cociente aditiva* (alr) sobre R^{D-1}, definida como

$$\mathrm{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, ..., \ln \frac{x_{D-1}}{x_D} \right], \tag{3}$$

y la *transformación log-cociente centrada* (clr) sobre R^D, definida como

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, ..., \ln \frac{x_D}{g(\mathbf{x})} \right]. \tag{4}$$

El hecho de tomar logaritmos de los cocientes sólo responde a una conveniencia matemática. Los log-cocientes son más manejables y además permiten que se cumplan algunas propiedades sencillas. Por ejemplo, no existe una relación exacta entre las varianzas $\mathrm{var}(x_i/x_j)$ y $\mathrm{var}(x_j/x_i)$ y, sin embargo, sí se cumple que $\mathrm{var}[\ln (x_i/x_j)] = \mathrm{var}[\ln (x_j/x_i)]$.

La transformación alr está más vinculada a modelizaciones paramétricas ya que a partir de ella es posible definir una clase de distribuciones, la *normal logística aditiva*, (Aitchison y Shen, 1980), con densidad

$$f(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} |\Sigma|^{-1/2}}{x_1 \cdots x_D} \exp\left\{ -\frac{1}{2} (\mathrm{alr}(\mathbf{x}) - \mu)\Sigma^{-1} (\mathrm{alr}(\mathbf{x}) - \mu)^T \right\},$$

donde μ y Σ son, respectivamente, el vector de medias y la matriz de covarianzas de los datos alr-transformados. Esta distribución, que denotaremos $\mathrm{aln}(\mu, \Sigma)$, juega un papel central en el estudio de distribuciones en S^D, análogo al de la distribución normal multivariante en R^D. Además resulta que el vector alr-transformado sigue una distribución normal multivariante en R^{D-1}. En Mateu-Figueras (2003) puede encontrase un estudio detallado sobre la definición y caracterización de modelos de probabilidad en el símplex. Un inconveniente de la transformación alr es su asimetría respecto a las partes de la composición. Aquella utilizada como denominador cobra un protagonismo

especial. Por ello, siempre que se trabaje con datos alr-transformados, debe comprobarse que la técnica estadística utilizada es invariante frente a permutaciones de las partes. Sin embargo, el principal inconveniente es que no es una transformación isométrica. En consecuencia, los ángulos y distancias en el símplex, con la métrica de Aitchison, no pueden asociarse con ángulos y distancias en el espacio real, con la métrica Euclídea. A pesar de esto, la consistencia de los resultados está garantizada cuando la inferencia se basa en la función de verosimilitud de la distribución aln (Aitchison, 1986). La transformación clr está más vinculada a contextos no paramétricos, ya que a partir de ella es posible especificar la distancia de Aitchison (2) en términos de la distancia Euclídea entre los vectores clr-transformados, esto es,

$$d_a(\mathbf{x}, \mathbf{x}') = d_e(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{x}')).$$

La transformación clr es simétrica e isométrica, pero la imagen de S^D queda realmente restringida a un subespacio de R^D y la matriz de covarianzas de los datos clr-transformados es singular, del mismo modo que ocurría con la matriz de covarianzas aplicada directamente sobre las composiciones.

La caracterización del símplex como espacio Euclídeo sugiere a Egozcue y otros (2003) proponer la transformación *log-cociente isométrica* (ilr), que salva los principales inconvenientes de las dos anteriores. La transformación ilr de una composición **x** se define como $\mathrm{ilr}(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle]$, donde $\mathbf{e}_1, ..., \mathbf{e}_{D-1}$ es una base ortonormal obtenida a partir del producto interior (1). Esto es, las componentes del vector ilr-transformado son las coordenadas de la composición **x** respecto a la base $\mathbf{e}_1, ..., \mathbf{e}_{D-1}$. Esta transformación es isométrica y los datos ilr-transformados se representan en los habituales ejes ortogonales. En Egozcue y otros (2003) los autores obtienen una expresión explícita para la transformación ilr dada una base ortonormal particular:

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{y} = [y_1, \ldots, y_{D-1}] \in R^{D-1}, \text{ donde } y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i} \right).$$

El principal problema aquí es determinar cuál es la base ortonormal más apropiada para un problema concreto, la que proporciona las expresiones que hacen más fácil la interpretación de los resultados, ya que las coordenadas ilr no suelen ser fáciles de

interpretar. Una posibilidad es utilizar las coordenadas y expresar los resultados en la base canónica de R^D sin abandonar el símplex. Trabajar en coordenadas permite definir distribuciones de probabilidad en el símplex, estimando los parámetros con los procedimientos habituales a partir de la expresión en coordenadas de los datos originales (Pawlowsky-Glahn, 2003; Mateu-Figueras y Pawlowsky-Glahn, 2007). Notar que realmente esta estrategia para *escapar* de un espacio muestral restringido expresando los datos en coordenadas respecto a una base ortonormal es aplicable a cualquier problema definido sobre un espacio Euclídeo (véase Pawlowsky-Glahn, 2003).

3.2. Medidas de posición y variabilidad en un conjunto de datos composicionales

Como sabemos, la medida de posición central más utilizada sobre conjuntos de datos en un espacio real es la media aritmética o centroide del conjunto. Sabemos además que es un buen estimador de la media de una variable aleatoria **x**. Desde un punto de vista geométrico, la media $\mu = \mathrm{E}[\mathbf{x}]$ de una variable aleatoria **x** en el espacio real minimiza la distancia Euclídea al cuadrado esperada $\mathrm{E}[\mathrm{d}_e(\mathbf{x},\xi)^2]$ y la varianza queda expresada como $\mathrm{Var}(\mathbf{x}) = \mathrm{E}[\mathrm{d}_e(\mathbf{x},\mu)^2]$. Trasladando esta caracterización a la geometría del símplex, Aitchison (2001) define el centro cen(**x**) de una composición aleatoria **x** en S^D como el valor que minimiza $\mathrm{E}[\mathrm{d}_a(\mathbf{x},\xi)^2]$. Dada una muestra $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_n)$ de datos composicionales, la *media geométrica composicional* definida como

$$\mathrm{g}(\mathbf{X}) = \mathrm{C}(g_1,...,g_D), \tag{5}$$

donde $g_j = (x_{1j} \cdots x_{nj})^{1/n}$, es un estimador lineal insesgado óptimo del centro cen(**x**) (Pawlowsky-Glahn y Egozcue, 2002). Destacar que en este contexto la linealidad se expresa como $\alpha_1 \otimes \mathbf{x}_1 \oplus \alpha_2 \otimes \mathbf{x}_2 \oplus \cdots \oplus \alpha_n \otimes \mathbf{x}_n$, donde los α_i son coeficientes reales. En la Figura 1 representamos una muestra de puntos del símplex S^3 en un *diagrama ternario*. Puede apreciarse cómo la media geométrica composicional (o), además de ser compatible con las operaciones básicas en el símplex, representa el centro de gravedad de la nube de puntos composicionales de forma más adecuada que la habitual media aritmética (*). En general, esto ocurrirá siempre que la nube de puntos en S^3 tenga una forma aproximadamente cóncava.

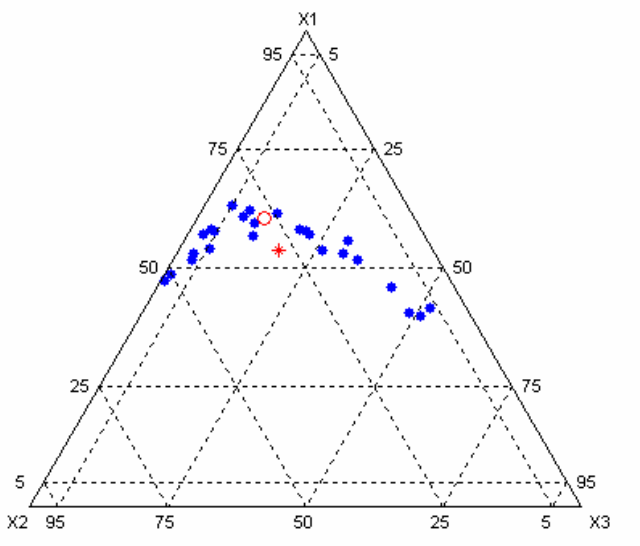


Figura 1. Diagrama ternario. Media geométrica composicional (o) y media aritmética (*) de un conjunto de datos composicionales.

Siguiendo la misma línea, la varianza entorno a cen(**x**) viene dada por $E[d_a(\mathbf{x}, cen(\mathbf{x}))^2]$. Pawlowsky-Glahn y Egozcue (2002) establecen que esta varianza coincide con la medida de *variabilidad total* $\text{tot var}(\mathbf{x})$ definida en Aitchison (1997) como la suma de las *varianzas log-cociente* $\text{var}(\ln x_i / x_j)$ dividida por el número de partes. De manera que su equivalente muestral

$$\text{totvar}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d_a^2(\mathbf{x}_i, g(\mathbf{X})) \qquad (6)$$

es una medida de variabilidad adecuada y coherente con la métrica del símplex.

Las medidas (5) y (6) están relacionadas con las medidas muestrales habituales aplicadas sobre los datos log-cociente transformados. Así, se tiene que:

1. $\text{alr}(g(\mathbf{X})) = \overline{\text{alr}(\mathbf{X})}$, $\text{clr}(g(\mathbf{X})) = \overline{\text{clr}(\mathbf{X})}$ y $\text{ilr}(g(\mathbf{X})) = \overline{\text{ilr}(\mathbf{X})}$, donde $\overline{\text{alr}(\mathbf{X})}$, $\overline{\text{clr}(\mathbf{X})}$ y $\overline{\text{ilr}(\mathbf{X})}$ representan las medias aritméticas de los respectivos conjuntos de datos transformados.

2. $\text{totvar}(\mathbf{X}) = \text{tr}(\text{cov}(\text{clr}(\mathbf{X}))) = \text{tr}(\text{cov}(\text{ilr}(\mathbf{X})))$. Destacar que tal relación no puede establecerse con la trasformación alr debido a que ésta no es isométrica.

El patrón de variabilidad de una composición **x** en S^D queda completamente determinada por la *matriz de variación* $\mathbf{T} = [\tau_{ij}]$ formada por las varianzas log-cociente $\tau_{ij} = \text{var}[\ln(x_i / x_j)]$, con i = 1,…,D-1 y j = i+1,…,D. Esta matriz es simétrica y tiene ceros en la diagonal principal. Aunque no puede expresarse como la matriz de covarianzas estándar de un vector, sí que está relacionada con las matrices de covarianzas de los vectores alr-, clr-, y ilr-transformados mediante simples operaciones matriciales (Aitchison, 1986; Egozcue y otros, 2003). Además, tiene la ventaja de que la matriz de variación de cualquier subcomposición se obtiene simplemente extrayendo las varianzas log-cociente asociadas a las partes que forma la subcomposición. En consonancia con la metodología log-cociente, las varianzas log-cociente miden la variabilidad relativa de una componente x_i respecto a otra x_j. Así, $\text{var}[\ln(x_i / x_j)] = 0$ implica que el cociente x_i / x_j es constante, por lo que existe una relación de proporcionalidad perfecta entre x_i y x_j, mientras que $\text{var}[\ln(x_i / x_j)] = \infty$ implica que una relación de proporcionalidad inexistente. Esta relación de proporcionalidad sustituye a la habitual correlación entre variables. Es sencillo obtener una medida que esté acotada en el intervalo [0,1] considerando $\nu_{ij} = \exp(-\sqrt{\tau_{ij}})$. De esta manera, un valor $\nu_{ij} = 0$ indica ausencia de relación de proporcionalidad y un valor $\nu_{ij} = 1$ indica relación de proporcionalidad perfecta. Destacar que si el interés es la independencia, ésta debe expresarse en términos de independencia de subcomposiciones. Por ejemplo, dada una composición de 5 partes, la independencia entre las subcomposiciones [s_1, s_2, s_3] y [s_4, s_5] se expresa como $\text{cov}[\ln(x_1 / x_3); \ln(x_4 / x_5)] = 0$ y $\text{cov}[\ln(x_2 / x_3); \ln(x_4 / x_5)] = 0$. Estas medidas son simplemente estimadas a partir de sus análogas muestrales aplicadas sobre los correspondientes log-cocientes observados.

3.3 El problema de los ceros y su tratamiento

En los puntos anteriores hemos puesto de manifiesto que el correcto análisis estadístico de vectores de proporciones gira entorno a los cocientes entre sus partes. Sin embargo, si un vector presenta alguna componente nula resulta imposible considerar todos los cocientes de la forma x_i / x_j. De manera que ni las transformaciones log-cociente, ni la distancia de Aitchison, ni las medidas descriptivas son aplicables en este caso. En la

literatura sobre datos composicionales se han distinguido dos tipos de ceros: los *ceros esenciales* y los *ceros por redondeo*. Los ceros esenciales se corresponden con componentes que realmente toman un valor nulo. A veces surgen por una excesiva desagregación o bien como un indicador de la existencia de diferentes submuestras dentro del conjunto de datos. Así, en la distribución de un presupuesto familiar en distintas partidas de gasto pueden distinguirse las familias que consumen tabaco o bebidas alcohólicas de las que no lo hacen. La fusión, *amalgamamiento*, de algunas de las partes o el análisis de submuestras de forma separada puede salvar el problema en algunos casos. En Aitchison y Kay (2003) y Bacon-shone (2003) pueden encontrarse algunas propuestas para tratar este tipo de ceros.

Un problema distinto es el de los ceros por redondeo, ceros que se corresponden con valores que no han podido observarse porque limitaciones en los instrumentos de medida o en el procedimiento de recogida y tratamiento de datos, o incluso políticas o culturales, impiden que se registren cuantías pequeñas que no superan cierto umbral de detección. El dato composicional con ceros por redondeo es un tipo especial de dato incompleto y, por lo tanto, es susceptible de ser tratado mediante técnicas de imputación o reemplazamiento. El procedimiento elegido debe preservar la estructura de covarianzas y ser coherente con las propiedades métricas específicas de los datos composicionales. En Martín-Fernández, Barceló-Vidal y Pawlowsky-Glahn (2003) se propone un procedimiento no paramétrico de reemplazamiento que satisface estas condiciones. Consiste en imputar de forma multiplicativa los valores nulos con un valor pequeño prefijado proporcional al umbral de detección. Con este procedimiento se obtienen buenos resultados cuando el número de ceros es reducido o en conjuntos de datos pequeños donde no es posible, o es arriesgado, establecer hipótesis distribucionales sobre los datos. El inconveniente lo encontramos en la necesidad medir la sensibilidad de los resultados al valor elegido para imputar, además de que tiende a subestimar la variabilidad. Como alternativa paramétrica cuando se dispone de conjuntos de datos relativamente grandes, Palarea-Albaladejo, Martín-Fernández y Gómez-García (2007) proponen una versión del algoritmo EM basada en la distribución de probabilidad normal logística aditiva. Este procedimiento imputa los ceros teniendo en cuenta la información suministrada por el resto de componentes y corrige sustancialmente la tendencia a la subestimación de la variabilidad del reemplazamiento

multiplicativo. Para agilizar la exposición de los ejemplos de la sección 4 asumiremos que los datos no contienen ceros o que éstos han sido previamente tratados con alguno de los métodos anteriores.

4. Algunos ejemplos seleccionados en economía

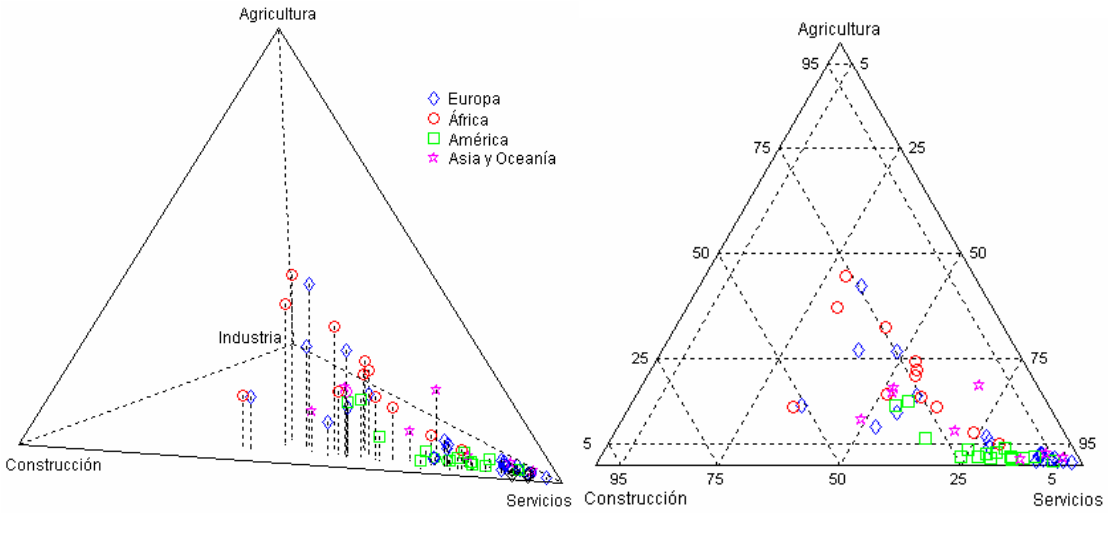
Hasta la fecha, el volumen de trabajos que hacen uso de la metodología log-cociente en el campo de la economía y las ciencias sociales es relativamente pequeño. Destacamos los trabajos de Fry, Fry y McLaren (2000), donde se aplica a la especificación de un modelo microeconométrico para el análisis de la distribución del presupuesto familiar; Katz y King (1999), donde se propone un modelo para estudiar cómo la distribución geográfica de los resultados electorales en un sistema multipartito dependen de una serie de variables socioeconómicas; Larrosa (2003), que estudia los patrones de la composición del stock de capital en distintos países; Anyadike-Danes (2003), que analiza la composición de la población no trabajadora en el Reino Unido y su relación con el nivel de desempleo. En las subsecciones siguientes ilustraremos la aplicación práctica del análisis log-cociente a través de unos ejemplos del ámbito económico.

4.1. Distribución de trabajadores inmigrantes por sectores de actividad.

En los últimos años, el dinamismo de nuestra economía ha provocado una importante entrada de mano de obra inmigrante para cubrir las necesidades del mercado laboral español. Es apreciable que el peso de las distintas nacionalidades no ha sido el mismo, destacando las americanas y norteafricanas, en concreto Ecuador y Marruecos, aunque se detecta una importancia creciente de otros grupos como es el caso de los ciudadanos europeos fuera del ámbito de la UE (informe 2/2004 del CES). Cabe preguntarse cómo varía la demanda de trabajadores por sectores de actividad según sea la nacionalidad de los mismos. Por ejemplo, contrastaremos si hay diferencias significativas entre la distribución por sectores de actividad de los trabajadores procedentes de países latinoamericanos y africanos (grupo 1, 27 países) y la de los trabajadores de otras nacionalidades (grupo 2, 31 países). Utilizaremos datos elaborados a partir de la información suministrada por el Boletín Estadístico de Extranjería e Inmigración, nº 6, de Julio de 2005, editado por el Ministerio de Trabajo y Asuntos Sociales.

Se consideran cuatro sectores: agricultura, industria, construcción y servicios. La Figura 2a muestra los datos de cada país en S^4 en un *diagrama cuaternario* distinguiendo los

continentes de origen. Se observa que el sector servicios es el que más trabajadores extranjeros atrae, especialmente los procedentes del continente americano. El sector industria es el que menos trabajadores inmigrantes contrata. Si nos quedamos con la subcomposición formada por los tres sectores principales, los datos se proyectan sobre el símplex S^3. Esto nos permite visualizar más claramente la distribución sectorial de los trabajadores de los distintos países en la Figura 2b.



(a) (b)

Figura 2. Representación gráfica de la distribución porcentual de los trabajadores extranjeros en los distintos sectores de actividad distinguiendo continente de procedencia.

En la Tabla 1 tenemos un resumen descriptivo de los datos (véase subsección 3.2.). Observamos que la gran mayoría del trabajo inmigrante se concentra en el sector servicios, tanto de forma global como cuando se distingue entre inmigrantes latinoamericanos y africanos y el resto.

		[agricul, industr, constru, servici]
Global	g(**X**)	[5.3027, 6.6366, 16.3326, 71.7281]
	totvar(**X**)	1.5072
Países Latinoamericanos y africanos (grupo 1)	g($\mathbf{X}_1$)	[6.4736, 6.9263, 21.1360, 65.4641]
	totvar($\mathbf{X}_1$)	1.3548
Otras nacionalidades	g($\mathbf{X}_2$)	[4.3880, 6.2955, 12.8461, 76.4704]

(grupo 2)	totvar($\mathbf{X}_2$)	1.5637

Tabla 1. Medidas descriptivas de la distribución de trabajadores inmigrantes según sector de actividad, a nivel global y distinguiendo nacionalidad de origen.

Asumiendo, respectivamente, modelos $\text{aln}(\mu_1, \Sigma)$ y $\text{aln}(\mu_2, \Sigma)$ para los datos de los grupos 1 y 2, contrastamos la hipótesis $\mu_1 = \mu_2$ mediante el estadístico Traza de Hotelling para muestras independientes aplicado sobre las muestras alr-transformadas. Se obtiene un valor del estadístico de 0.193 con un p-valor asociado igual a 0.022. Destacar que estos valores son independientes de la componente utilizada como denominar en la transformación alr. Al habitual 95% de confianza rechazaríamos la hipótesis de igualdad de medias, esto es, existe una diferencia significativa en la distribución sectorial de los dos grupos de trabajadores inmigrantes considerados. Un buen indicador de la naturaleza de esta diferencia puede obtenerse como $g(\mathbf{X}_1) \ominus g(\mathbf{X}_2)$ = [29.0591, 21.6707, 32.4081, 16.8621]. Este vector diferencia composicional muestra que el grupo 1 se caracteriza frente al grupo 2 por un mayor peso relativo en los sectores construcción y agricultura, respecto a los sectores industria y servicios.

4.2. Empleo del tiempo según la edad.

En las últimas décadas, con el diseño e implementación de políticas de bienestar en los países desarrollados, hay un creciente interés por el estudio del uso que los ciudadanos hacen de su tiempo. Esta información sirve de apoyo a la formulación de políticas familiares, de igualdad de género, etc.

En este ejemplo utilizaremos datos elaborados a partir de la Encuesta de Empleo del Tiempo realizada por el INE en 2004 con el fin de construir un modelo de regresión que explique la relación entre el uso del tiempo y la edad de los individuos. Para simplificar la exposición, se agrupan las actividades en 3 grupos: 1. estudios o trabajo; 2. cuidados personales, hogar y familia; 3. ocio y vida social. Hemos considerado una submuestra aleatoria de 442 individuos sin proporciones nulas en los grupos de actividades considerados. Destacar que la distribución del tiempo es una variable composicional en origen. El total, 24 horas, es el mismo para todos. De hecho, podríamos trabajar directamente en unidades de tiempo y no con proporciones.

El modelo a estimar puede expresarse en términos de los datos alr-transformados como

$$\begin{cases} \log(\text{act}_1 / \text{act}_3) = \beta_{01} + \beta_{11} \log(\text{edad}) + \varepsilon_1 \\ \log(\text{act}_2 / \text{act}_3) = \beta_{02} + \beta_{12} \log(\text{edad}) + \varepsilon_2, \end{cases} \tag{7}$$

donde act_1, act_2 y act_3 denotan, respectivamente, las proporciones de tiempo dedicadas a las actividades de los grupos 1, 2 y 3. Destacar que este modelo de regresión es invariante frente a permutaciones de las partes de la composición **act** = [act_1, act_2, act_3]. Esto es, si se toma como divisor de la transformación alr una parte distinta a act_3 se obtienen resultados compatibles. A continuación, se estiman los parámetros del modelo mediante el habitual procedimiento de mínimos cuadrados ordinarios y se llevan al símplex utilizando la inversa de la transformación alr (3). Tenemos entonces la *recta* de regresión composicional dada por

$$[\text{act}_1, \text{act}_2, \text{act}_3] = [0.1129, 0.6036, 0.2836] \oplus [0.4354, 0.3270, 0.2376] \otimes \log(\text{edad}).$$

Para facilitar la interpretación de los resultados, calculamos las distribuciones de tiempo estimadas por el modelo y las ordenamos de menor a mayor edad. Denotemos mediante $\hat{g}_{\max}$ y $\hat{g}_{\min}$ las estimaciones de las distribuciones de tiempo asociadas, respectivamente, a la mayor y menor edad de la muestra. Si calculamos la diferencia composicional $\hat{g}_{\max} \ominus \hat{g}_{\min}$ se obtiene [0.5484, 0.2992, 0.1524], mientras que $\hat{g}_{\min} \ominus \hat{g}_{\max}$ es igual a [0.1555, 0.2849, 0.5596]. Estos resultados ponen claramente de manifiesto el cambio en la distribución del tiempo según la edad. Conforme se incrementa ésta aumenta el tiempo dedicado a trabajo o estudios y disminuye el dedicado a ocio y vida social. Podemos decir que el intercambio de tiempo se produce básicamente entre estos dos grupos. El valor de la media geométrica composicional, [0.2996, 0.6046, 0.0957], revela que el grupo cuidados personales, hogar y familia es al que más tiempo se dedica. Puede comprobarse también que esto es así independientemente de la edad. Destaca que la relación de este grupo de actividad con el segundo y tercer grupo en importancia, que como hemos visto cambian según la edad, se mantiene casi constante en términos relativos. Así, $0.5484 / 0.2992 \cong 0.5596 / 0.2849$ y $0.1524 / 0.2992 \cong 0.1555 / 0.2849$.

En la Figura 3 representamos los datos observados y la recta de regresión composicional en el símplex S^3 correspondiente a la composición **act**. Las flechas indican el sentido del desplazamiento conforme se incrementa la edad.

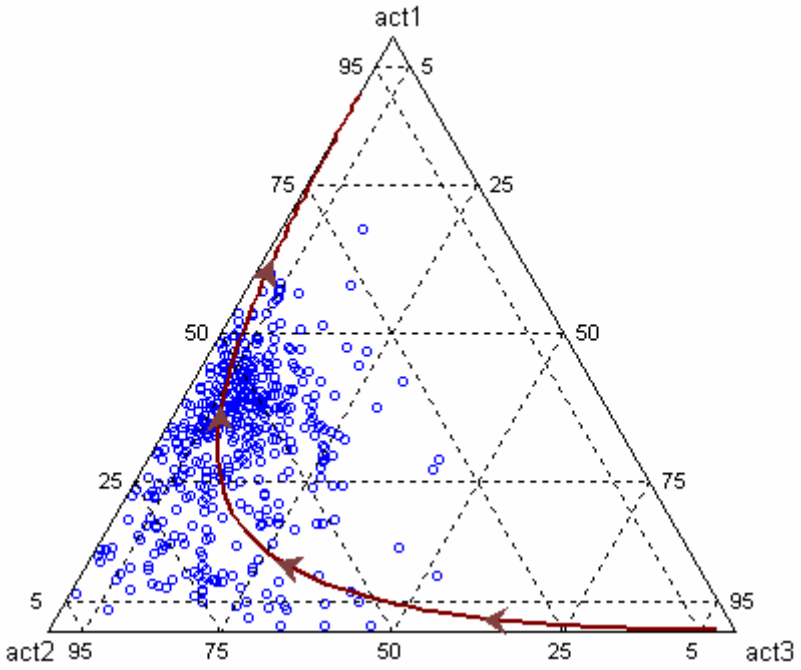


Figura 3. Distribución del tiempo diario entre tres grupos de actividades: valores observados y recta de regresión composicional estimada en función de la edad.

Finalmente, destacaremos que también es posible realizar un análisis semejante en el plano alr-transformado. Así, las rectas de regresión estimadas a partir de (7) pueden interpretarse como las ecuaciones paramétricas de la recta que relaciona $\log(\text{act}_2/\text{act}_3)$ y $\log(\text{act}_1/\text{act}_3)$, siendo $\log(\text{edad})$ la variable paramétrica. La ecuación de dicha recta vendrá dada entonces por

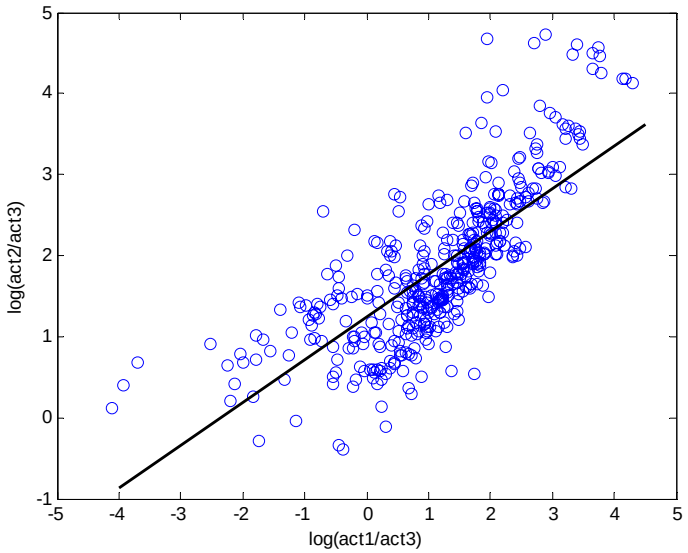
$$\log(\text{act}_2/\text{act}_3) = \frac{\hat{\beta}_{02}\hat{\beta}_{11} - \hat{\beta}_{12}\hat{\beta}_{01}}{\hat{\beta}_{11}} + \frac{\hat{\beta}_{12}}{\hat{\beta}_{11}}\log(\text{act}_1/\text{act}_3),$$

que en este caso resulta ser igual a $\log(\text{act}_2/\text{act}_3) = 1.2412 + 0.5274\log(\text{act}_1/\text{act}_3)$. En la Figura 4 representamos los puntos en el plano alr junto a la recta estimada. La pendiente de la recta es aproximadamente 1/2, por lo que un incremento de 2 unidades en $\log(\text{act}_1/\text{act}_3)$ se traduce aproximadamente en un incremento de 1 unidad en $\log(\text{act}_2/\text{act}_3)$. Realizando algunas operaciones llegamos a la expresión equivalente:

$$\text{act}_1/\text{act}_3 \propto (\text{act}_2/\text{act}_3)^{1.8961}. \qquad (8)$$

De esta manera se observa de forma más evidente una relación cuadrática entre los incrementos en $\text{act}_1/\text{act}_3$ y los incrementos en $\text{act}_2/\text{act}_3$. Sabemos que la proporción

act_2 se mantiene aproximadamente constante con la edad, por lo que el incremento *cuadrático* de act_1 / act_3 en relación al incremento de act_2 / act_3 se explica necesariamente por el efecto conjunto de la disminución en act_3 junto al incremento en act_1. De modo que llegamos a la misma conclusión que cuando realizamos el análisis sobre el símplex.



5. Comentarios finales

Al iniciar un proceso de modelización y análisis de datos es necesario prestar especial atención a las características del espacio muestral sobre el que están definidos. Esto es imprescindible para la elaboración de modelos estadísticos apropiados. Cuando se trabaja con datos numéricos, habitualmente, se asume que se trata de datos reales. Sin embargo, los datos que se refieren a partes de un total, entre otros, no se corresponden con valores continuos sobre todo el espacio real, sino que se definen sobre un subconjunto de este último sujeto a las restricciones de no negatividad y de suma constante. Sin embargo, la mayoría de las veces se analizan recurriendo a las técnicas multivariantes estándar para datos reales no restringidos, quizás por su apariencia de vectores de números reales. Aunque en algunos casos prácticos puede que se llegue a conclusiones similares, lo importante es que, de base, no se está actuando correctamente. Las medidas de posición, variabilidad, similaridad, etc. utilizadas en el espacio real no son coherentes con la estructura geométrico-algebraica del símplex, por

lo que nada garantiza que las inferencias tengan algún significado y se correspondan en cierta medida con la realidad subyacente.

A lo largo de este trabajo se han revisado los fundamentos que determinan el marco formal dentro del cual desarrollar técnicas y medidas coherentes con las características del símplex como espacio muestral. Reconociendo que los datos contienen sólo información sobre las magnitudes relativas, la metodología log-cociente permite trasladar el problema al más familiar espacio real, y aplicar entonces las herramientas de análisis usuales. La implementación de esta estrategia no está exenta de problemas prácticos. En concreto, han sido comentados los problemas derivados de la existencia de observaciones con partes nulas y algunas de las soluciones propuestas. Como se ha destacado, la caracterización del símplex como espacio Euclídeo abre la posibilidad de circunscribir la modelización estadística al símplex, sin necesidad de recurrir a transformaciones de los datos, redefiniendo los conceptos fundamentales de forma adecuada. Esta idea marca una de las principales líneas de investigación actuales en esta área.

Finalmente, para ilustrar la aplicación práctica del análisis log-cociente, se han considerado dos ejemplos en el campo económico empleando técnicas de uso habitual entre los investigadores. Estos ejemplos ponen de manifiesto las diferencias en el modo de actuar y de interpretar los resultados cuando se trabaja con composiciones.

Bibliografía

Aitchison, J. (1986): *The statistical analysis of compositional data*, Chapman & Hall, London. Reimpresión por Blackburn Press en 2003.

Aitchison, J. (1997): "The one-hour course in compositional data analysis or compositional data analysis is easy", en Pawlowsky-Glahn, V. (ed.), *Proceedings of the Third Annual Conference of the IAMG*, Barcelona, pp. 3-35.

Aitchison, J., Shen, S. M. (1980): "Logistic normal distributions: some properties and uses", *Biometrika*, 67, pp. 261-272.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., Pawlowsky-Glahn, V. (2000): "Logratio analysis and compositional distance", *Mathematical Geology*, 32, pp. 271-275.

Aitchison, J., Kay, J. W. (2003): "Possible solutions of some essential zero problems in compositional data analysis", en Thió-Henestrosa, S., Martín-Fernández, J. A. (eds.), *First Compositional Data Analysis Workshop - CoDaWork'03*. Universitat de Girona.

Anyadike-Danes, M. (2003): "The allometry of- non-employment", en Thió-Henestrosa, S., Martín-Fernández, J. A. (eds.), *First Compositional Data Analysis Workshop - CoDaWork'03.* Universitat de Girona.

Bacon-Shone, J. (2003): "Modelling structural zeros in compositional data", en Thió-Henestrosa, S., Martín-Fernández, J. A. (eds.), *First Compositional Data Analysis Workshop - CoDaWork'03.* Universitat de Girona.

Billheimer, D., Guttorp, P., Fagan, W. F. (2001): "Statistical interpretation of species composition", *J. Am. Stat. Assoc.*, 96, pp. 1205-1214.

Bohling, G. C., Davis, J. C., Olea, R. A. Harff, J. (1998): "Singularity and nonnormality in the classification of compositional data", *Mathematical Geology*, 30, pp. 5-20.

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

Fry, J. M., Fry, T. R. L., McLaren, K. R. (2000): "Compositional data analysis and zeros in micro data", *Applied Economics*, 32, pp. 953-959.

Katz, J. N., King, G. (1999): "A statistical model for multiparty electoral data", *American Political Science Review*, 93, pp. 15-32.

Larrosa, J. M. (2003): "A compositional statistical analysis of capital stock", en Thió-Henestrosa, S., Martín-Fernández, J. A. (eds.), *First Compositional Data Analysis Workshop - CoDaWork'03.* Universitat de Girona.

Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V. (2003): "Dealing with zeros and missing values in compositional data sets", Mathematical Geology, 35, pp. 253–278.

Mateu-Figueras, G. (2003): *Modelos de Distribución sobre el Símplex*, tesis doctoral, Universitat Politècnica Catalunya.

Mateu-Figueras, G. Pawlowsky-Glahn, V. (2007): "The skew-normal distribution on SD", *Comm. Statist. Theory Methods*, 36, (en prensa).

Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J. (2007): "A parametric approach for dealing with compositional rounded zeros", *Mathematical Geology*, (en prensa).

Pawlowsky-Glahn, V. (2003): "Statistical modelling on coordinates", en Thió-Henestrosa, S., J. A. Martín-Fernández (eds.), *First Compositional Data Analysis Workshop - CoDaWork'03.* Universitat de Girona.

Pawlowsky-Glahn, V., Egozcue, J. J. (2001): "Geometric approach to statistical analysis on the simplex", *Stoch. Environ. Res. Risk Assess.* ", 15, pp. 384-398.

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

Pearson, K. (1897): "Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurements of organs", *Proc. R. Soc.*, 60, pp. 489-498.