Tolosana-Delgado, R., 2011. Guía para el análisis espacial de datos composicionales. *Boletín Geológico y Minero,* 122 (4): 469-482
ISSN: 0366-0176

Guía para el análisis espacial de datos composicionales

Raimon Tolosana-Delgado

Laboratorio de Ingenería Marítima, Dept. Enginyeria Hidràulica, Marítima i Ambiental, Universitat Politènica de Catalunya, c/Jordi Girona 1-3, E-08034 Barcelona. raimon.tolosana@upc.edu

RESUMEN

El tratamiento de bases de datos composicionales (proporciones, porcentajes, concentraciones, etc.) con dependencia espacial debe hacerse atendiendo a las características matemáticas de éstos: una composición válida debe tener todas las componentes positivas y su suma debe ser igual o menor a un total (1, 100%, etc.). En general, esto se consigue de forma razonablemente fácil transformando la composición mediante series de log-cocientes de componentes. Para estudiar la variabilidad espacial de una composición se recomienda estimar y modelar el variaciograma: el conjunto de variogramas de todos los log-cocientes de un par de componentes. El variaciograma contiene toda la información necesaria para caracterizar una composición con estacionaridad intrínseca, y se puede modelar con herramientas habituales de la geoestadística, como el modelo de coregionalización lineal. Además, se pueden estudiar las propiedades del modelo e inferir relaciones entre componentes y posibles procesos vinculados a alguna escala espacial concreta. Finalmente, interpolar la composición y generar mapas es tarea sencilla con las herramientas existentes de krigeado y simulación: estas técnicas y conceptos deben aplicarse a un conjunto de log-cocientes cualquiera, tal que exista una transformación invertible entre éllos y las componentes de la composición original.

Palabras clave: composición, función de covarianza, geometría de Aitchison, símplex

Guide for the spatial analysis of compositional data

ABSTRACT

Dealing with spatially-dependent compositional databases (including proportions, data in percentages, concentrations etc) should pay heed to the mathematical properties of these kinds of data: a valid composition must have positive components whose sum is at most a constant (1, 100% etc.). Generally speaking this is easily done by working on a set of log-ratios of components rather than using the raw data. To study the spatial variability of these databases it is best to estimate and model the lr-variograms, i.e. the set of variograms of all possible pairwise log-ratios of components in the composition. Such lr-variograms contain all the information necessary to deal with intrinsic stationary compositions and may be modelled with standard geostatistical tools such as the linear model of coregionalization. Moreover, the properties of the model can be studied and relationships inferred between components and possible processes linked to a given spatial scale. Finally, component-by-component interpolation and mapping is straightforward with existing kriging and simulation techniques: these tools and concepts should be applied to any set of invertible component log-ratios, i.e. log-ratio transformations, in such a way that the original composition can be recovered from the transformed data and vice versa.

Keywords: Aitchison geometry, composition, covariance function, simplex

ABRIDGED ENGLISH VERSION

Introduction: basic concepts and methods

Spatially dependent compositional data sets are quite common in geology, especially in geochemical surveys. Apart from this, one can also find compositional data when dealing with sand-silt-clay textural compositions, or with petrographic or mineral compositions.

A composition is a vector of positive components summing up to a constant, the closure*, typically 1, 100 or* 10^6 *(Eq. 1), in which each component shows the* relative *importance of a part in a total (Aitchison, 1986). It is frequently found, however, when dealing with compositions, that the constant-sum constraint induces a negative bias in the covariances of a compositional data set, leading to the appearance of spurious correlations unrelated to any natural process of exclusion or affinity between components (Chayes, 1960). To overcome this limitation one should return to the mainstay of the definition, the relative character of compositions: this calls for the use of log-ratios as ubiquitous tools in compositional data analysis (Aitchison, 1997; Barceló-Vidal, 2000) since these transformations introduce relative quantities (a set of percentages, depending on which parts have been taken into account in the analysis) into absolute magnitudes. In general terms, if one has a* D *–part composition one needs* (D–1) *log-ratios to express the same information: Eq. (2) shows a general expression to compute log-ratios from compositions and vice-versa. Once in log-ratios, classical statistical tools can be applied with no difficulty (to compute covariances between log-ratios for example). But to express the variability of a compositional data set one can also make use of the variation matrix* $\mathrm{T} = [t_{ij}]$ *(Eq. 3), the set of variances of each possible pairwise log-ratio of two components. Eq. (4) gives a fundamental relation between the variation matrix and the log-ratio covariance matrix, showing that these matrices afford exactly the same information and can thus be treated in the same way by statistical methods. In particular, principal component analysis (PCA) and its graphic representation, the biplot (Aitchison, 1997), is a useful tool to describe dependence between components. More on compositional data analysis*

can be found in other contributions to this volume, or in the works by Pawlowsky-Glahn y Egozcue (2001); Aitchison (2002); Egozcue and Pawlowsky-Glahn (2005) and Tolosana-Delgado et al. *(2005).*

*The regionalized character of a variable may intuitively appear as a larger similitude or dependence between observations from neighbouring sampling locations in the geographical space. This quite natural effect generates a correlation between observations that typically decreases concomitantly with the distance between the sampling locations. Geostatistics deals with this issue by introducing a function (the autocovariance function or the semivariogram) to model this fading dependence (*cf.*, for example, Journel and Huijbregts, 1978; Isaaks and Srivastava, 1989; Wackernagel, 1998; Chilès and Delfiner, 1999; Clark and Harper, 2000). In the case of compositional data these concepts and tools should be applied to the log-ratio transformed composition because the classical covariance function or variogram will be spurious (Pawlowsky-Glahn and Olea, 2004). Consequently, the same strategy of working with log-ratios is applied. To characterize the spatial correlation structure one can make use of the covariance function for the log-ratios or else characterize the variograms of each pairwise log-ratio, a matrix-valued function called* variation-variograms *(Eq. 9). Once again these two tools are related through Eq. (10), which allows one to study the spatial structure by making use of the simple pairwise variograms. In the modelling phase, the commonly used linear model of coregionalization (LMC) (Eq. 6) allows one to combine a set of spatial correlation structures* $p^k(\vec{h})$ *with their own shapes or ranges of influence or anisotropy (Table 1), together with some weighting covariance matrices* C_k *(Wackernagel, 1998). For compositions, the LMC can be applied to variation-variograms (Eq. 13), meaning that each spatial structure is weighted with a variation matix* B_k *instead of a covariance matrix. Thus, one can use a logarithmic goodness-of-fit criterion (Eq. 15) in automatic variogram fitting processes, which will enhance the fit at short distances (those more important for interpolation). Another advantage of this approach is that these matrices,* C_k*, may be rank-deficient and thus easier to interpret: in this case one can work with their eigenvectors,* b_k*, to define a set of log-ratios linked to equilibrium reactions or balances between components.*

The final goal of geostatistics is typically to obtain interpolation maps of the interesting quantities. As far as compositional geostatistics is concerned, these quantities will be log-ratios. Any set of (D – 1) *log-ratios can be optimally interpolated with a cokriging technique (Eq. 7) (*cf. *Myers, 1984; Chilès and Delfiner, 1999, among others, for details), using the fitted variogram model. Once all the maps are available they can be combined pixel by pixel through Eq. (2) to obtain maps of the original components. In the same way, one can apply simulation procedures to the log-ratios, and back-transform the simulations to compositions.*

This paper is written with the aim of serving as a guide to the spatial analysis of compositional data. The Spanish version explains and illustrates these issues step by step. A complete account of the theory and concepts of compositional geostatistics can be found in Pawlowsky-Glahn and Olea (2004) and Tolosana-Delgado (2006).

Illustration

These concepts are illustrated with a data set of an environmental survey covering the Grazer Paläozoikum (Austria), studied by Weber y Davis (1990). Figure 1 shows a sketch of the geological background. Samples were taken from stream sediments and analysed for several major and trace elements. For this study we chose only the 7 major elements present: Al, Ca, Fe, K, Mg, Na and Ti. Using the automatic fitting criterion of Eq. (15), the set of variation-variograms in Figure 2 was fitted to a spherical model with nugget (Eq. 16). The variation matrices linked to these structures were analysed with PCA techniques to uncover the processes linked to each scale. The nugget effect is described by two more-or-less uncorrelated balances: one showing the contrast of Ca vs *Ti-Fe, and the second contrasting Mg against K-Na-Al. On the other hand, the spherical structure shows a reasonable one-dimensional pattern in the subcomposition Ca-Mg-Na-(Ti). Ternary diagrams of these subcompositions (Fig. 4) confirm these intuitions and suggest that the Fe/Ti ratio is almost constant at all scales whilst Mg and Ca seem unrelated in the nugget scale (thus possibly suggesting that the balance Mg* vs *K-Na-Al is related to the kind of siliciclastic source) whilst they show a linked behaviour on the spherical scale (thus suggesting a more regional control, such as the presence/absence of dolomitization, for example). These hints are consistent with the cokriged maps of the subcomposition Ca-Fe-Ti (Fig. 5).*

Conclusions

Instead of finishing with some standard conclusions, this paper provides a kind of recipe for the geostatistical analysis of compositions. The steps to follow are briefly:

1. *characterize the spatial-dependence structure of the regionalized composition using variation-variograms, i.e. the set of variorams of each possible pairwise log-ratio;*
2. *model the empirical variation-variograms with a linear model of coregionalization (LMC), in which each structure can then be treated with compositional PCA (i.e. extracting the eigenvectors of each variation matrix involved in the LMC) to study the dependence between components at each spatial scale;*
3. *the obtained LMC can be expressed in terms of any set of interesting log-ratios by simple matrix multiplication procedures; these can then be used to interpolate log-ratios, which can then be back-transformed to obtain maps of interpolated percentages;*
4. *variability issues can be studied using standard cokriging and simulation tools; simulated log-ratios can be back-transformed as in the preceding point to obtain simulated compositions expressed in percentages.*

Introducción

En geología, tratar con bases de datos composicionales con dependencia espacial es algo muy común, desde las fases de exploración geoquímica de vastos territorios hasta la de control de la explotación de una mina. Además, más allá de las bases de datos geoquímicos, se pueden considerar como datos composicionales las proporciones texturales de arena-limo-arcilla, los porcentajes de clases petrográficas o la composición mineral. Por ello, también va siendo común encontrar datos composicionales con dependencia espacial en investigación básica en geología, como soporte a los estudios clásicos de tectónica, estratigrafía, sedimentología y petrografía.

Una composición es un vector de elementos positivos de suma igual o menor a una constante (normalmente 1, 100 o 10^6), que aportan *información relativa* sobre la importancia de varios componentes de un sistema (Aitchison, 1986). Su naturaleza relativa es lo más característico de una composición (Aitchison, 1997; Barceló-Vidal, 2000): para poder interpretar el sentido de un incremento de una determinada componente *A* de 10 a 80, debemos saber o bien cuál era el total (100 o 10^6), o bien qué le sucede a otra componente (¿*B* pasa de 8 a 2, o de 20 a 160?). Ello conlleva tratar con cocientes de dos o mas variables: independientemente de cuáles fueran las unidades, un paso de 10/8 a 80/2 es enorme, mientras que de 10/20 a 80/160 no ha habido ningún cambio en la contribución relativa de la parte *A* respecto la *B*. Surge por tanto como una opción natural el tratar los datos composicionales siempre mediante cocientes. Sin embargo, las propiedades estadísticas (medias, varianzas) de A / B y las de B / A no muestran ninguna relación, lo cual implicaría cierta arbitrariedad en los resultados dependiendo del sentido del cociente: para simetrizar el problema, tomamos logaritmos, puesto que $\log(A / B) = -\log(B / A)$. La idea clave del análisis composicional es pues aplicar una transformación log-cociente invertible a los datos, previa a cualquier análisis. La invertibilidad permitirá representar de forma unívoca los resultados en log-cocientes a composiciones, e interpretar los porcentajes y proporciones en los términos habituales, es decir, atendiendo a su naturaleza relativa.

La dependencia espacial entre muestras de una base de datos se muestra como una mayor similitud entre las muestras a medida que sus puntos de muestreo se encuentran más cerca. Este efecto, nada sorprendente, genera una correlación entre muestras como una función, en general decreciente con la distancia entre puntos de muestreo. Ello invalida la aplicación de la mayoría de métodos estadísticos clásicos, que requieren muestras independientes entre sí. La geoestadística permite trabajar con este tipo de datos modelando la función de correlación espacial. Una vez que se ha obtenido un modelo satisfactorio de esta auto-correlación, podemos interpolar los datos espacialmente y calcular los errores que cometemos en la interpolación misma. En el caso de tratar datos composicionales, aplicaremos estos métodos a una serie de log-cocientes, seleccionada para simplificar los cálculos.

Este número especial contiene artículos sobre las características particulares de los datos composicionales, y también sobre el estudio de datos con dependencia espacial. En este contexto, el presente artículo presenta una guía para el tratamiento de bases de datos con ambas características. Por ser una guía, esta contribución mezcla la teoría con la aplicación en los sucesivos pasos que deben seguirse en el análisis de este tipo de datos. Así, la próxima sección presenta el conjunto de datos de ilustración. A continuación se resumen los conceptos de geoestadística y de análisis de datos composicionales "clásicos" necesarios. Las secciones siguientes cubren los pasos del análisis geoestadístico adaptado a los datos composiconales:

1. estimación de la estructura espacial,
2. modelado y factorización de ésta,
3. krigeado o cokrigeado
4. y análisis de la incertidumbre espacial (varianza de krigeado y simulación).

No se han incluido aspectos teóricos en profundidad o demostraciones, que pueden encontrarse en Pawlowsky-Glahn y Olea (2004) o Tolosana-Delgado (2006).

Caso de estudio

A efectos ilustrativos se usará una base de datos geoquímicos de una campaña de control multi-objetivo llevada a cabo en la región del Grazer Paläozoikum, el área al norte de la ciudad austríaca de Graz donde aflora el Paleozoico (Weber y Davis, 1990). Dos litologías representan conjuntamente algo más del 50% del área, a partes iguales: las metafilitas del *Tonshiefer* [A] del Devónico, y las dolomías y calizas de la formación *Rannach-Hochlantsch* [B]. En orden de importancia, afloran luego el basamento cristalino [C] y rocas detríticas terciarias [E]. Finalmente, con contribuciones menores se observan las litologías menores de las series detríticas de relleno de las cuencas de *Raasberg* [D] y *Gosau* [F], así como las metapelitas arenosas del *Dornerkogel* [G]. El aprovechamiento minero de la región de estudio se inicia ya en la prehistoria, y ha sido una de sus principales fuentes de riqueza. Más detalles sobre la geología y los motivos de esta campaña de muestreo, así como de los detalles del análisis se pueden encontrar en Weber y Davis (1990).

Estos autores aplicaron un análisis de componentes principales al conjunto de datos, sin tener en cuenta la correlación espacial entre los datos. Se extrajeron 7 componentes principales, y cada una de éllas fue entonces tratada con métodos geoestadísticos: se calculó un variograma teórico, se ajustó un modelo de variograma, y se interpoló la componente principal. Los mapas así obtenidos, interpolando la composición geoquímica de los sedimentos en los ríos, se consideran como proxis de las características geológicas de su entorno. Con las 7 componentes interpoladas se pudo reconstruir aproximadamente los mapas de cada uno de los 34 elementos. Sin embargo, hay que destacar que algunos mapas obtenidos con este procedimiento pueden mostrar interpolaciones negativas.

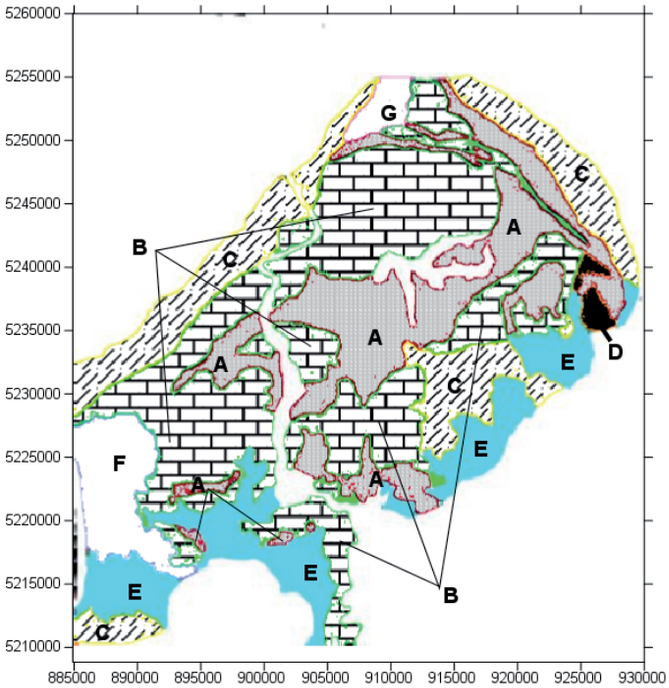

Figura 1. Mapa de las principales unidades del Grazer Paläozoikum, de acuerdo con Weber y Davis (1990). Éstas son: metafilitas de la unidad Tonschiefer (A), calizas y dolomías de las unidades Rannach-Hochlantsch (B), basamento cristalino (C), serie Raasberg (D), sedimentos terciarios (E), serie Gosau (F), y filitas arenosas del Dornerkogel (G).

Figure 1. Map of the main units of the Grazer Paläozoikum, according to Weber and Davis (1990): Tonschiefer shales (A), Rannach-Hochlantsch mudstones and dolostones (B), crystalline basement (C), Raasberg series (D), Tertiary sediments (E), Gosau series (F) and Dornerkogel sandy shales (G).

Conceptos previos

Datos composicionales

Sean $x = [x_1,\ldots,x_D]$ e $y = [y_1,\ldots,y_D]$ dos vectores(-fila) composicionales, i.e. cuyas componentes muestran la importancia relativa de una serie de partes de un total. Dada esa información relativa, es habitual clausurar las composiciones a suma constante, por ejemplo, a porcentajes:

$$C[x] = \frac{100}{\sum_{i=1}^{D} x_i}\mathrm{x}. \qquad (1)$$

Denotamos por S^D el espacio muestral de las composiciones, el *símplex*. Sea $\lambda \in \Re$ un valor escalar real. Podemos dar al símplex una estructura de espacio euclídeo mediante las operaciones de perturbación (denotado por $\oplus$, y definido como el producto directo por componentes de los dos vectores, y clausurado), potenciación por λ (denotado por $\odot$, y definido como el vector clausurado obtenido a partir de potenciar cada componente de la composición por λ) y el producto escalar de Aitchison (Pawlowsky-Glahn y Egozcue, 2001; Aitchison, 2002; Tolosana-Delgado *et al.*, 2005).

Como en cualquier espacio euclídeo, la manera más conveniente de *representar* un vector en el símplex es mediante sus coordenadas respecto a un sistema de referencia ortonormal. El paso entre composiciones y vectores de coordenadas es immediato mediante la *transformación log-cociente isométrica*,

$$\mathrm{ilr}(\mathbf{y}) = \ln(\mathbf{y})\cdot\mathbf{V}^t = \boldsymbol{v}, \quad \mathrm{ilr}^{-1}(\boldsymbol{v}) = C[\exp(\boldsymbol{v}\cdot\mathbf{V})], \qquad (2)$$

donde $\mathbf{V}$ es una matriz de *contrastes*, i.e. una matriz de $(D-1)\times D$ elementos tal que

$$\mathbf{V}\cdot\mathbf{V}^t = \mathbf{I}_{D-1},$$
$$\text{y}$$
$$\mathbf{V}\cdot\mathbf{1}^t = \mathbf{0}^t,$$

o en otras palabras, cuyas $(D-1)$ filas suman cero, y forman un sistema de vectores ortonormales. Una posibilidad es basar esta matriz en una partición binaria secuencial de las partes en grupos, que describan patrones de asociación entre las partes, según Egozcue y Pawlowsky-Glahn (2005).

Respecto al tratamiento estadístico de datos composicionales, ya Chayes (1960) advirtió de los peligros de interpretar la matriz de correlaciones de una composición. Debido a la clausura a suma constante, la correlación entre algunas componentes debe ser negativa, sin que ello implique un proceso natural de exclusión mutua. Este efecto se conoce como *negative bias* (*sesgo negativo*). Por otro lado, la correlación entre dos partes cualesquiera cambia arbitrariamente en función de si se analiza la composición completa o una subcomposición (por ejemplo, en un diagrama ternario o cuaternario). Esto se conoce como *correlación espúrea*: dado que la correlación entre dos partes depende de las demás, y ya no sólo de ellas dos, uno no puede interpretarla con las reglas habituales. En resumen, las correlaciones (y covarianzas, así como toda técnica basada en éllas) de una composición son arbitrarias, y de éllas no se puede desprender presencia o ausencia de dependencia mediante un proceso natural.

Como alternativa al uso de correlaciones/covarianzas clásicas, Aitchison (1986) propuso una serie de medidas de dispersión y codependencia, basadas en la "regla de oro" del análisis composicional (*usa sólo log-cocientes de partes*). Para los objetivos de este artículo, usamos la matriz de variaciones y la covarianza de las coordenadas.

Sea $\mathbf{Y} = [Y_1,\ldots,Y_D]$ una composición aleatoria. La *matriz de variaciones* es una matriz de $D \times D$ elementos, denotada por $\mathbf{T} = [t_{ij}]$, donde

$$t_{ij} = \mathrm{Var}\left[\ln\frac{Y_i}{Y_j}\right]. \quad (3)$$

La covarianza de las coordenadas es simplemente la clásica matriz de covarianzas calculada para la composición log-cociente-transformada: $\Sigma = \mathrm{Cov}[\mathrm{ilr}(\mathbf{Y})]$. Nótese que Σ tiene $(D-1)\times(D-1)$ elementos, y presenta todas las propiedades de una matriz de covarianzas. Estas dos formas de expresar la dispersion/codependencia están relacionadas por (Aitchison, 1986; Pawlowsky-Glahn y Olea, 2004):

$$\Sigma = -\frac{1}{2}\mathbf{V}\cdot\mathbf{T}\cdot\mathbf{V}^t. \quad (4)$$

Ello implica que los autovectores de ambas matrices $\mathbf{T} = [t_{ij}]$ y $\Sigma = [\sigma_{ij}]$ son los mismos (dado que $\mathbf{V}$ es una matriz ortonormal, es decir, una *rotación* del espacio, y por tanto no puede cambiar los autovectores). Además, los autovalores de la primera matriz deben ser (-2) –veces aquellos de la segunda. Por tanto, $\mathbf{T}$ es una matriz semi-definida negativa, puesto que Σ debe ser definida positiva.

Finalmente, se puede definir también una varianza escalar como descriptor de la dispersión total de la composición: la *varianza métrica*, denotada por $\mathrm{Mvar}[\mathbf{Y}]$. Esta se define como la distancia promedio entre la composición y su media, y equivale a la traza de la matriz de covarianzas de las coordenadas, o la variación promedio

$$\mathrm{Mvar}[\mathbf{Y}] = \sum_{i=1}^{D-1}\sigma_{ii} = \frac{1}{2D}\sum_{i,j=1}^{D}t_{ij}. \quad (5)$$

Geoestadística

En esta sección presentamos brevemente los conceptos y pasos habitualmente aplicados en el análisis geoestadístico multivariable. Otros artículos de este número especial, o algunos manuales clásicos, como Journel y Huijbregts (1978); Isaaks y Srivastava (1989); Chilès y Delfiner (1999); Clark y Harper (2000), pueden ser útiles para profundizar en ello. En esta sección usamos una aproximación al tema similar a la de Myers (1984). Los conjuntos de datos regionalizados se modelizan con el concepto de la *función vectorial aleatoria*, denotada por $\mathbf{Z}(\vec{x})$. Esto es una colección infinita de variables aleatorias vectoriales indexadas mediante $\vec{x}\in\Re^p$, una localización en el espacio-tiempo real (aunque típicamente es suficiente con $p = 2$, para datos sobre un mapa). Casi siempre asumimos que la función aleatoria presenta *estacionariedad de segundo orden*, es decir,

$$\mathrm{E}[\mathbf{Z}(\vec{x})] = \boldsymbol{\mu}(\vec{x}) = \boldsymbol{\mu},$$

y

$$\mathrm{Cov}[\mathbf{Z}(\vec{x}),\mathbf{Z}(\vec{y})] = \mathbf{C}(\vec{x}-\vec{y}),$$

o en otras palabras, el valor esperado del vector aleatorio es una constante μ no dependiente del espacio, y la covarianza $\mathbf{C}(\cdot)$ entre los vectores aleatorios ligados a dos localizaciones $\vec{x}$ e $\vec{y}$ solo depende del desplazamiento entre ellas, $\vec{h} = \vec{x} - \vec{y}$. La función $\mathbf{C}(\cdot)$ se conoce por *función de covarianza*. Típicamente estas condiciones son demasiado estrictas, y se relajan a *estacionariedad intrínseca*, definida como

$$\mathrm{E}[\mathbf{Z}(\vec{x}) - \mathbf{Z}(\vec{y})] = \mathbf{0},$$

y

$$\mathrm{Var}[\mathbf{Z}(\vec{x}) - \mathbf{Z}(\vec{y})] = \boldsymbol{\Gamma}(\vec{x}-\vec{y}),$$

o en otras palabras, el incremento medio entre dos vectores ligados a dos posiciones es nulo y su varianza sólo depende del desplazamiento entre ellos. La función $\Gamma(\cdot)$ se conoce por *(semi)-variograma*. En el caso que tanto ésta como la función de covarianza existan, están relacionadas por

$$\boldsymbol{\Gamma}(\vec{h}) = 2\mathbf{C}(\vec{0}) - (\mathbf{C}(\vec{h}) + \mathbf{C}(-\vec{h})),$$

lo que permite pasar de una a otra cuando $\mathbf{C}(\vec{h}) = \mathbf{C}(-\vec{h})$, es decir, cuando hay simetría espacial. Esta simetría es a menudo otra hipótesis necesaria para una inferencia satisfactoria de las propiedades de la función aleatoria en problemas bi- y tridimensionales.

$\mathbf{C}(\cdot)$ y $\Gamma(\cdot)$ son funciones matriciales. Los términos de la diagonal se conocen por autocovarianzas y variogramas directos, y muestran la continuidad espacial de una variable particular. Los términos de fuera de la diagonal se denominan covarianzas/variogramas cruzados, y explican cómo se relacionan dos variables distintas tomadas en dos posiciones distintas. Para trabajar con estas funciones, uno estima las versiones empíricas para varios desplazamientos, y se ajusta un modelo adecuado a ellas.

En el ajuste de un modelo a las versiones experimentales, se exige que la función modelo sea simétrica definida positiva para las covarianzas (condicionalmente definida negativa para los variogramas), una condición de difícil manejo. En aplicaciones prácticas, uno más bien se restringe a usar el *modelo de corregionalización lineal* (LMC, e.g. Wackernagel, 1998), en el que los variogramas empíricos son modelados como una combinación de distintos autocorrelogramas $\rho^k(\vec{h})$,

$$\boldsymbol{\Gamma}(\vec{h}) = \sum_{k=0}^{k}(1-\rho^k(\vec{h}))\cdot\mathbf{C}_k, \quad (6)$$

donde $\mathbf{C}_k$ son matrices (semi-)definidas positivas. Respecto a los correlogramas, estos son funciones esca-

lares que describen cómo se desvanece la correlación en una variable tomada en dos puntos a medida que estos puntos se alejan. La Tabla 1 muestra algunos ejemplos, incluyendo los usados en este artículo. Es habitual que el primer correlograma sea la función por casos $\rho^0(\vec{0}) = 1$ y $\rho^0(\vec{h} \neq \vec{0}) = 0$, el llamado *efecto pepita* (*nugget effect*).

Los pasos precedentes de estimación y modelado del variograma se conocen como *análisis estructural*.

Se puede obtener exactamente el mismo resultado si se toman los bloques como $\mathbf{\Gamma}_{ij} = \mathbf{C}(\vec{x}_i - \vec{x}_j)$, usando covarianzas en vez de variogramas.

La principal ventaja del cokrigeado sobre otros métodos de interpolación es su capacidad de producir una medida del error cometido en la estimación: la varianza de cokrigeado. Esta se obtiene como

forma	nombre	fórmula normalizada
—	esférico	$(1 - 3r/2 + r^3/2)I(r < 1)$
$0 < v < 2$	lineal generalizado	$1 - r^v$
1	lineal	$1 - r$
2	cuadrático	$1 - r^2$
	exponencial	$\exp(-r/3)$
	Gaussiano	$\exp(-r^2/3)$

Tabla 1. Correlogramas $\rho(r)$ más comúnmente usados, como funciones de la distancia anisótropa adimensional $r^2 = \vec{h}^t \times A^{-1} \times \vec{h}$, donde A es una matriz describiendo la elipse (o elipsoide, en 3D) de influencia donde la correlación es notable.
Table 1. Most commonly used correlograms, $\rho(r)$ as a function of an anisotropic distance r, where A is a matrix describing an ellipse (or ellipsoid, in 3D) in which the correlation influence is considerable.

Una vez se tiene un modelo para el variograma, éste se puede usar para interpolar las observaciones disponibles y estimar con ello la función aleatoria en una localización no muestreada. La interpolación geoestadística multivariante se conoce como *cokrigeado* (*cokriging*). El estimador de *cokrigeado ordinario* para una localización $\vec{x}_0$ se denota por $\hat{\mathbf{z}}_0$ y se estima mediante una función lineal de los datos $\{\mathbf{z}_i = \mathbf{z}(\vec{x}_i), i = 1,\ldots, n\}$:

$$\hat{\mathbf{z}}_0 = \sum_{i=1}^{n} \mathbf{\Lambda}_i \cdot \mathbf{z}_i \tag{7}$$

restringida mediante laa condición de asuencia de sesgo $\sum_{i=1}^{n} \mathbf{\Lambda}_i = \mathbf{I}$. Los pesos $\mathbf{\Lambda}_i$ son matrices del mismo tamaño que $\mathbf{C}(.)$ o $\mathbf{\Gamma}(.)$. Cada uno de estos pesos muestra la influencia de la muestra z_i sobre la predicción de $\mathbf{Z}(\vec{x}_0)$, y se obtienen resolviendo el sistema de ecuaciones

$$\mathbf{\Lambda} = \mathbf{S}^{-1} \cdot \mathbf{S}_0,$$

donde las matrices se definen por bloques como sigue, tomando $\mathbf{\Gamma}_{ij} = \mathbf{\Gamma}(\vec{x}_i - \vec{x}_j)$:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \vdots \\ \mathbf{\Lambda}_n \\ \boldsymbol{\nu} \end{bmatrix}, \mathbf{S} = \begin{bmatrix} \mathbf{\Gamma}_{11} & \cdots & \mathbf{\Gamma}_{1n} & \mathbf{I} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Gamma}_{n1} & \cdots & \mathbf{\Gamma}_{nn} & \mathbf{I} \\ \mathbf{I} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}, \mathbf{S}_0 = \begin{bmatrix} \mathbf{\Gamma}_{10} \\ \vdots \\ \mathbf{\Gamma}_{n0} \\ \mathbf{I} \end{bmatrix}.$$

$$\hat{\mathbf{\Sigma}}_{OK} = \mathbf{\Gamma}(\vec{0}) - \mathbf{\Lambda}^t \cdot \mathbf{S}_0, \tag{8}$$

Entre otras utilidades, ello permite generar regiones predictivas para las interpolación o simular diferentes versiones alternativas de la función aleatoria en los puntos no muestreados. Ambas utilidades permiten evaluar y propagar la incertidumbre espacial a otros aspectos ulteriores del análisis, como podrían ser cálculos de costes de explotación o remediación. En estos casos, la hipótesis de Gaussianidad de la función aleatoria se torna casi imprescindible. Bajo estas circunstancias, se puede afirmar que $\mathbf{Z}(\vec{x}_0)$ sigue una distribución normal multivariante con media $\hat{z}_0$ y matriz de covarianza $\hat{\mathbf{\Sigma}}_{OK}$.

Tal y como se ha mencionada anteriormente, ninguna de estas técnicas debería aplicarse directamente a datos composicionales regionalizados. El motivo principal es la naturaleza espúrea de las correlaciones espaciales descritas por variogramas y funciones covarianza: dado que la suma de proporciones o porcentajes es fija, las matrices $\mathbf{\Gamma}(\vec{h})$ o $\mathbf{C}(\vec{h})$ son singulares *para cualquier desplazamiento* $\vec{h}$, y todas sus filas y columnas deben sumar siempre cero: la presencia de varianzas positivas en la diagonal de estas matrices obliga a alguna covarianza o variograma cruzado a ser negativo. Es más, el cokrigeado presenta severos problemas derivados de esta singularidad. El estimador de cokrigeado requiere la inversión (generalizada) de una matriz múltiplemente singular, lo que no es posible con casi ningún software, comercial o de código libre. Además, nada garantiza que los resultados ob-

tenidos sean positivos, y por tanto interpretables. Si para esquivar el problema de singularidad se decide aplicar kriging a cada componente independientemente, el resultado es subóptimo y además las interpolaciones de las distintas componentes no respetarán la suma constante: pueden ser negativas, y sumar más o menos que 1 (o que 100%). Para un mayor detalle en la descripción de estos problemas véase Pawlowsky-Glahn y Olea (2004) o Tolosana-Delgado (2006). Los próximos apartados aportan soluciones a estos problemas, basadas en el uso de log-cocientes.

Estimación de la estructura espacial

Para explorar la codependencia de las partes de una composición en el caso no regionalizado podemos optar por tratar la matriz de variaciones (Eq. 3) o bien por la matriz de covarianzas de un vector de coordenadas. De la misma manera, para explorar la estructura espacial de una composición regionalizada $\mathbf{Y}(\vec{x})$, usaremos el equivalente a una matriz de variaciones $\mathbf{T}(\vec{h}) = [t_{ij}(\vec{h})]$, donde

$$t_{ij}(\vec{h}) = \mathrm{Var}\left[\ln\frac{y_i(\vec{x})}{y_j(\vec{x})} - \ln\frac{y_i(\vec{x}+\vec{h})}{y_j(\vec{x}+\vec{h})}\right]. \quad (9)$$

La función matricial $\mathbf{T}(\vec{h})$ se llama lr-variograma (Pawlowsky-Glahn y Olea, 2004) o variaciograma (*variation-variogram*). Alternativamente, podemos escoger una transformación ilr, calcular las coordenadas asociadas como $\mathbf{Z}(\vec{x}) = \mathrm{ilr}(\mathbf{Y}(\vec{x}))$, y obtener funciones de covarianza $\mathbf{C}(\vec{h})$ y variogramas $\mathbf{\Gamma}(\vec{h})$ de las coordenadas, como se describe en la sección sobre geoestadística. Es directo demostrar que para cada desplazamiento $\vec{h}$,

$$\mathbf{\Gamma}(\vec{h}) = -\frac{1}{2}\mathbf{V}\cdot\mathbf{T}(\vec{h})\cdot\mathbf{V}^t, \quad (10)$$

estableciendo un paralelismo con la Eq. (4).

El procedimiento de estimación de estas varianzas es el mismo que en geoestadistica no composicional:

1. seleccionar un desplazamiento $\vec{h}$,
2. buscar todos los pares de puntos de muestreo $\{\vec{x}_n, \vec{x}_m\}$ tales que $\vec{x}_n, \vec{x}_m \approx \vec{h}$ con una cierta tolerancia;
3. estimar la función de estructura espacial deseada,

$$\hat{t}_{ij}(\vec{h}) = \frac{1}{2N(\vec{h})}\sum_{n,m}\left[\ln\frac{y_i(\vec{x}_m)}{y_j(\vec{x}_m)} - \ln\frac{y_i(\vec{x}_n)}{y_j(\vec{x}_n)}\right]^2,\ 1 \le i < j \le D \quad (11)$$

$$\hat{\gamma}_{ij}(\vec{h}) = \frac{1}{2N(\vec{h})}\sum_{n,m}(\mathrm{ilr}_i(\mathbf{y}(\vec{x}_m)) - (\mathrm{ilr}_i(\mathbf{y}(\vec{x}_n))) \times \times (\mathrm{ilr}_j(\mathbf{y}(\vec{x}_m)) - (\mathrm{ilr}_j(\mathbf{y}(\vec{x}_n)))\ \ 1 \le i < j \le D-1 \quad (12)$$

4. regresar al punto 1.

En el cálculo de algunas coordenadas intervienen todas las partes; estas coordenadas pueden por tanto acumular errores analíticos notables. Si alguna de estas partes es un valor perdido o por debajo del límite de detección, las coordenadas no podrán calcularse, y los variogramas tampoco. Por el contrario, cada variaciograma sólo requiere dos partes en cada par de puntos, lo cual reduce el impacto de los errores, ceros o valores perdidos, y permite calcular todos los variaciogramas con el máximo de pares posible. Por este motivo parece razonable usar variaciogramas en lugar de variogramas de coordenadas. Otra razón en la misma línea aparece en el modelado de la estructura espacial.

Usando los datos de ilustración presentados anteriormente, se han calculado los variaciogramas experimentales omnidireccionales (fig. 2), para 20 desplazamientos equiespaciados entre 0 y 25 km.

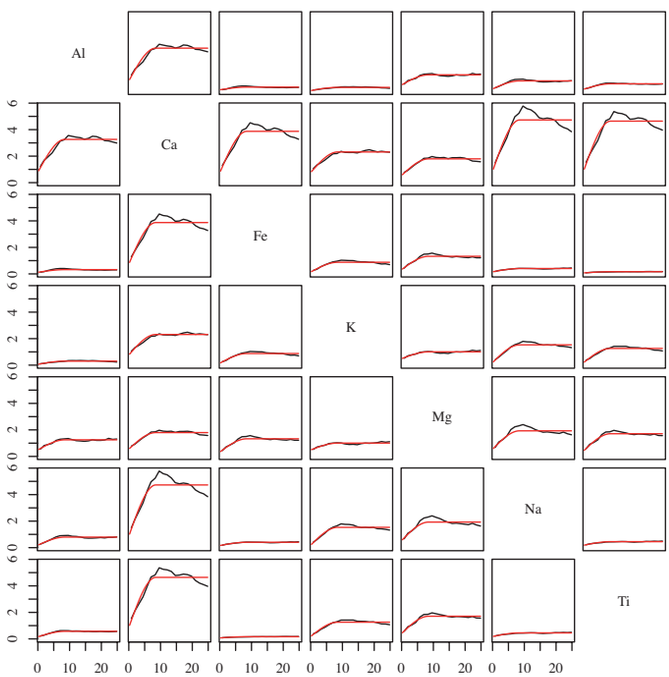

Figura 2. Variaciogramas del sistema Al-Ca-Fe-K-Mg-Na-Ti (en %), distancias en km. La linea quebrada (negra) muestra el variograma experimental, la curva suave (roja) el modelo ajustado.
Figure 2. Variation-variograms of the system Al-Ca-Fe-K-Mg-Na-Ti (in %), lag distances in km. The broken line (black) shows the empirical variograms, whilst the smooth function (red) shows the fitted model.

Modelado y factorización de la estructura espacial

El siguiente paso del procedimiento estándar en geoestadística no composicional es el ajuste (manual o semi-automático) a las estimaciones de un modelo de variograma válido: eso es una función matricial condicionalmente negativa definida. Sin embargo, tí-

picamente uno se limita a ajustar un modelo lineal de coregionalización (LMC), como el de la Eq. (6). En el caso de datos composicionales, dada la Eq. (10), podemos expresar el LMC para el variaciograma y para el variograma en coordenadas como

$$\mathbf{T}(\vec{h}) = \sum_{k=0}^{k} (1-\rho^k(\vec{h}))\cdot \mathbf{B}_k,\ \boldsymbol{\Gamma}(\vec{h}) = \sum_{k=0}^{k} (1-\rho^k(\vec{h}))\cdot \mathbf{C}_k \qquad (13)$$

donde $\rho^k(\vec{h})$ son una serie de correlogramas como en la Eq. (6), B_k representan matrices de variaciones (i.e., matrices simétricas semi-definidas negativas con valores diagonales cero) y

$$\mathbf{C}_k = -\frac{1}{2}\mathbf{V}\cdot\mathbf{B}_k\cdot\mathbf{V} \qquad (14)$$

las correspondientes matrices de covarianza en coordenadas. Esta relación viene impuesta por la Eq. (10), e implica que las descomposiciones en autovalores y autovectores de $\mathbf{C}_k$ y de $\mathbf{B}_k$ estan ligadas como lo estaban en la Eq. 4.

Una de las ventajas de usar el variaciograma proviene del hecho de que todas sus componentes sean variogramas, es decir, funciones estrictamente positivas: eso permite usar un criterio logarítmico de optimización del ajuste, en el que la discrepancia del modelo respecto al variaciograma experimental se mide como

$$D(\boldsymbol{\theta}) = \sum_k N(\vec{h}_k) \sum_{i<j}^{D} \ln^2 \frac{\hat{t}_{ij}(\vec{h}_k)}{t_{ij}(\vec{h}_k \mid \boldsymbol{\theta})}, \qquad (15)$$

donde $\mathbf{T}(\vec{h} \mid \boldsymbol{\theta})$ representa un LMC con un vector de parámetros $\boldsymbol{\theta}$, que incluye matrices de anisotropía y demás parámetros de forma de los correlogramas $\rho^k(\vec{h})$ así como las matrices $\mathbf{B}_k$. Un criterio como el de la Eq. (15) prioriza el ajuste del modelo a desplazamientos $\vec{h}$ pequeños sobre desplazamientos más largos. Ello equivale a dedicar más esfuerzos a las propiedades del variaciograma cerca del origen (que condicionan más los valores de la interpolación, según Chilès y Delfiner, 1999), y menos esfuerzos a capturar el valor de las mesetas (valroes mucho más variables, y muy relacionadas con la varianza de krigeado).

Las estimaciones de todas las funciones de estructura espacial (funciones de covarianza, variogramas, variaciograma) tienen comportamientos bastante erráticos: es habitual sub- o sobre-estimar las mesetas con unos factores del 50% al 200% (Wackernagel, 1998). Por ello, es a menudo deseable reducir el rango de las distintas $\mathbf{B}_k$, que por defecto es $D-1$. Es incluso posible que cada $\mathbf{B}_k$ sea una matriz de rango 1, por tanto computable como

$$\mathbf{C}_k = \mathbf{b}_k^t \cdot \mathbf{b}_k$$

donde $\mathbf{b}$ es un vector-fila de $(D-1)$ coordenadas y norma uno. Si ello es así, los distintos vectores $\mathbf{b}_k$ pueden tomarse como los vectores de una base composicional, que al ser multiplicados por la matriz $\mathbf{V}$ de la Eq. (2), darán una nueva base $\mathbf{V}$'. Esta base simplificará el proceso de interpolación, como veremos en la sección siguiente. Así mismo, es posible que alguno de estos vectores esté ligado a un proceso concreto (e.g., reacciones de equilibrio) que ocurre a una escala descrita por el correlograma asociado.

La figura 2 muestra los valores experimentales del variaciograma para los datos de ejemplo, así como el modelo esférico siguiente:

$$\mathbf{T}(h) = \rho^0(h)\cdot\mathbf{B}_0 + \left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right)\cdot\mathbf{B}_1 \qquad (16)$$

ajustado automáticamente con el criterio (Eq. 15), lo cual da un alcance isótropo $a = 8.49$ km y unas matrices de pepita y meseta de:

$$\mathbf{B}_0 = \begin{pmatrix} Al & Ca & Fe & K & Mg & Na & Ti \\ 0 & 0.782 & 0.120 & 0.060 & 0.486 & 0.186 & 0.156 \\ 0.782 & 0 & 0.732 & 0.762 & 0.544 & 0.851 & 0.857 \\ 0.120 & 0.732 & 0 & 0.140 & 0.322 & 0.156 & 0.073 \\ 0.060 & 0.762 & 0.140 & 0 & 0.487 & 0.186 & 0.181 \\ 0.486 & 0.544 & 0.322 & 0.487 & 0 & 0.535 & 0.404 \\ 0.186 & 0.851 & 0.156 & 0.186 & 0.535 & 0 & 0.178 \\ 0.156 & 0.857 & 0.073 & 0.181 & 0.404 & 0.178 & 0 \end{pmatrix}$$

$$\mathbf{B}_1 = \begin{pmatrix} 0 & 2.482 & 0.202 & 0.243 & 0.764 & 0.609 & 0.417 \\ 2.482 & 0 & 3.141 & 1.552 & 1.251 & 3.878 & 3.785 \\ 0.202 & 3.141 & 0 & 0.736 & 1.006 & 0.250 & 0.093 \\ 0.243 & 1.552 & 0.736 & 0 & 0.517 & 1.346 & 1.079 \\ 0.764 & 1.251 & 1.006 & 0.517 & 0 & 1.393 & 1.306 \\ 0.609 & 3.878 & 0.250 & 1.346 & 1.393 & 0 & 0.275 \\ 0.417 & 3.785 & 0.093 & 1.079 & 1.306 & 0.275 & 0 \end{pmatrix}$$

Estas matrices ofrecen la posibilidad de explorar la estructura de codependencia de las variables entre sí. Como si se tratara de un análisis exploratorio de datos composicionales no regionalizados (Aitchison, 1997), un análisis de autovectores y autovalores de estas matrices puede contribuir a dilucidar qué procesos están jugando un papel en el caso de estudio, y a qué escalas. De los autovalores de $\mathbf{B}_0$ y $\mathbf{B}_1$ se obtiene

el gráfico de sedimentación del análisis de componentes principales, tras multiplicarlos por un factor de –2, como impone la Eq. (4), mientras que los autovectores definen las componentes principales del conjunto de datos, y se pueden representar en un biplot. La Fig. 3 muestra esos biplots, de los que se deduce que

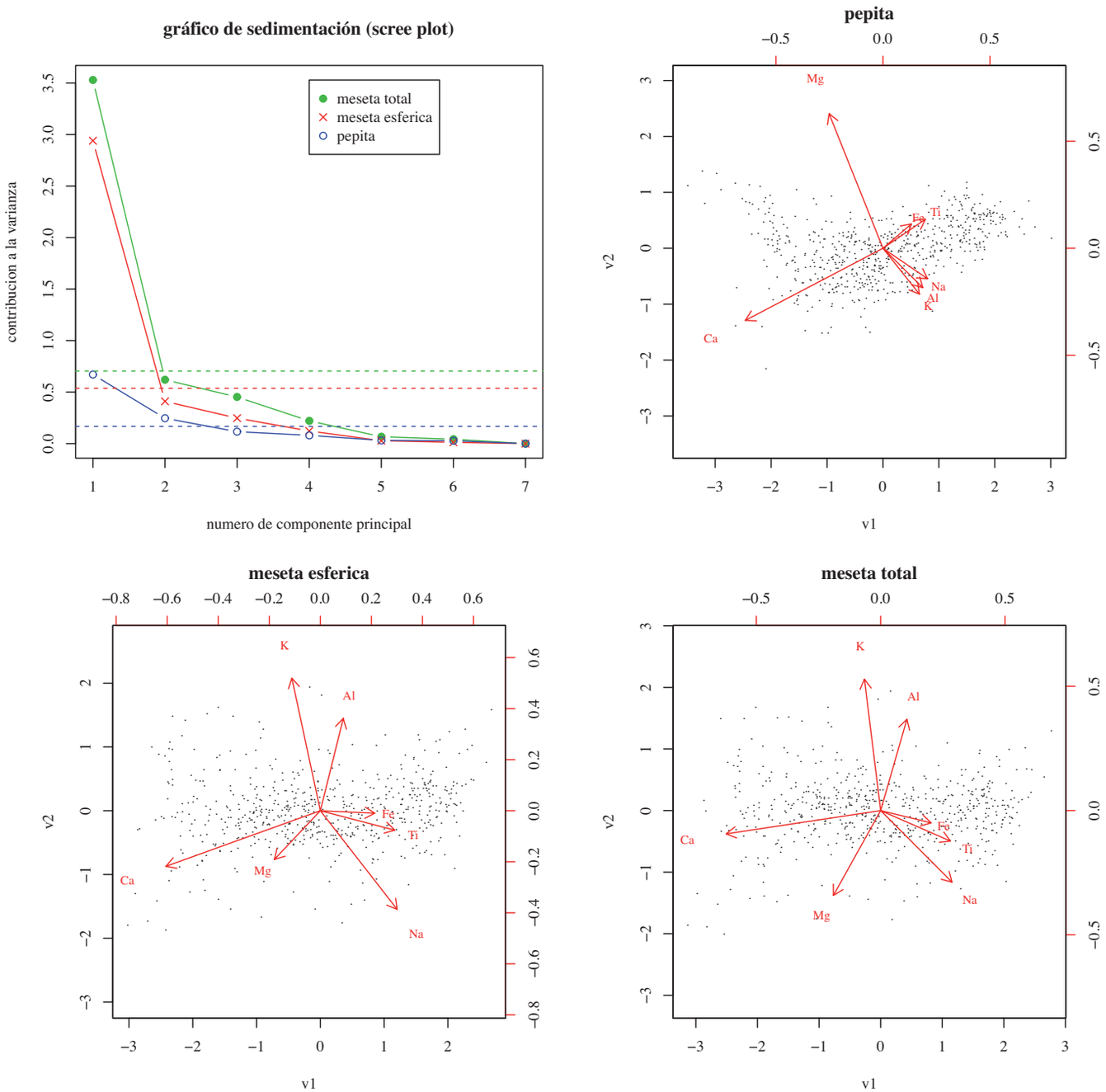

Figura 3. Gráfico de sedimentación (diagrama superior izquierda), y biplots del efecto pepita (sup. derecha), de la estructura esférica (inf. izq.) y de la meseta total (inf. dcha.), mostrando las asociaciones de variables en cada una de las dos escalas. El gráfico de sedimentación incluye tres líneas horizontales a trazos, que indican el número aproximado de componentes principales significativas de cada estructura (aquéllos cuya varianza es mayor que la varianza métrica promedio de la estructura). La meseta total está dominada por la estructura esférica, puesto que el efecto pepita representa aproximadamente un 25% de la variabilidad total.

Figure 3. Scree plot (top left) and biplots of the nugget effect (top right) of the spherical structure (bottom left) and of the total sill (bottom right), showing the association structures between variables in each of the two scales. The scree plot shows three horizontal dashed lines, indicating the approximate number of significant principal components in each structure (a significant component is defined here as one having a larger variance than the average variance of data). Note that the total sill is dominated by the spherical structure, as the nugget represents approximately 25% of the total variability.

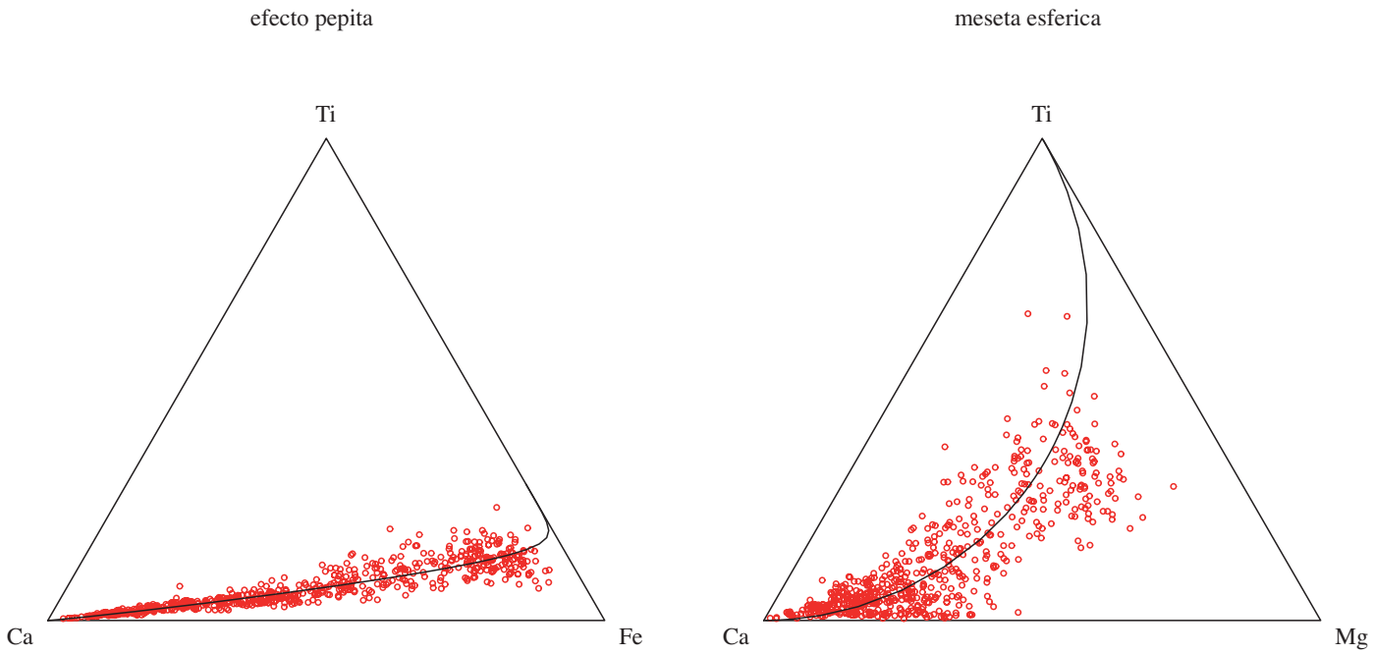

Figura 4. Diagramas ternarios de las subcomposiciones representativas de los procesos dominantes a escala pepita y a escala esférica (con radio de influencia ~8.5 km), con indicación de la curva de la primera componente principal en cada caso.
Figure 4. Ternary diagrams of some subcompositions representative of the dominant processes at microscale (linked to the nugget) and at a ~8.5 km-mesoscale (linked to the spherical structure). Each ternary diagram also shows a curve following the first principal component of each subcomposition.

la variabilidad pepita (debida a procesos ligados a una escala espacial inferior a la distancia de muestro) está dominada por dos factores más o menos ortogonales, y por tanto razonablemente independientes: uno representa el balance de Ca contra Fe-Ti, y otro es el balance de Mg contra K-Na-Al. Así mismo, la variabilidad asociada a la estructura del variograma esférico está dominada por una sola componente, con elevados coeficientes positivos para Ti-Fe-Na y negativos para Ca. Un diagrama ternario del sistema Ca-Fe-Ti representa la primera componente de variabilidad pepita (fig. 4): éste muestra una relación cuasi-constante del cociente Ti/Fe (véase los pequeños valores de este cociente en $\mathbf{B}_0$ y $\mathbf{B}_1$) con respecto a la gran variabilidad de la proporción de Ca. Un vistazo a los biplots de meseta muestra que las variables Ca-Mg-Ti-Na se encuentran aproximadamente alineadas, lo que sugiere un patrón unidimensional en su conducta: un diagrama ternario de a subcomposición Ca-Mg-Na lo confirma razonablemente. En su conjunto, estos diagramas sugieren que el mayor control sobre la variabilidad del conjunto de datos es el enriquecimiento relativo en calizas (fuente de Ca) vs. componentes terrígenos (con aportes de Fe/Ti constantes). A escala pepita una segunda fuente de variabilidad la ofrece el intercambio de Mg por elementos félsicos (Al, K, Na), lo que sugiere un contraste entre rocas más félsicas y menos félsicas (la geología nos indica que no hay aportes de rocas básicas notables). Por el contrario, a la mesoescala esférica Mg se asocia preferentemente con Ca, lo que podría sugerir que esta componente principal también contrasta dominios con distinto grado de dolomitización.

Interpolación

Una vez se ha obtenido un modelo de variaciograma satisfactorio, la interpolación de la composición es una consecuencia immediata. A nivel de cálculo, podemos:

- usar una base ilr arbitraria,
- usar una base sugerida por el análisis estructural (es decir, definida a partir de los vectores $\{\mathbf{b}_k\}$)
- o bien krigear las distintas estructuras separademente con krigeado univariante, según la filosofía del llamado krigeado factorial (Chilès y Delfiner, 1999; Wackernagel, 1998), y recomponer luego la composición mediante los autovectores de cada estructura.

En los primeros dos casos, podemos calcular las coordenadas de nuestras observaciones con la Eq. (2) y expresar el variaciograma en el mismo sistema de coordenadas mediante la Eq. (10). Con estos variogramas y observaciones podemos aplicar el sistema de cokrigeado ordinario (Eq. 7) y obtener predicciones

para las coordenadas ilr. Si hemos usado una ilr interpretable, los mapas de estas predicciones pueden mostrar estructuras interesantes. En cualquier caso, una vez disponemos de las coordenadas interpoladas, podemos usar la transformación ilr inversa (Eq. 2) para recuperar composiciones en porcentajes o proporciones, que podrán ser representadas en un mapa junto a los datos originales.

En el caso de ejemplo que nos ocupa, podemos escoger una base arbitraria de cálculo, y obtener las $(D-1)=6$ coordenadas de nuestros datos mediante la Eq. (2). Con la misma matriz $\mathbf{V}$ y la Eq. (14) aplicada a $\mathbf{B}_0$ y $\mathbf{B}_1$, podremos expresar el modelo de variograma en esa base. Cualquier paquete de interpolación mediante cokriging nos servirá entonces para obtener 6 mapas interpolados para las coordenadas, que

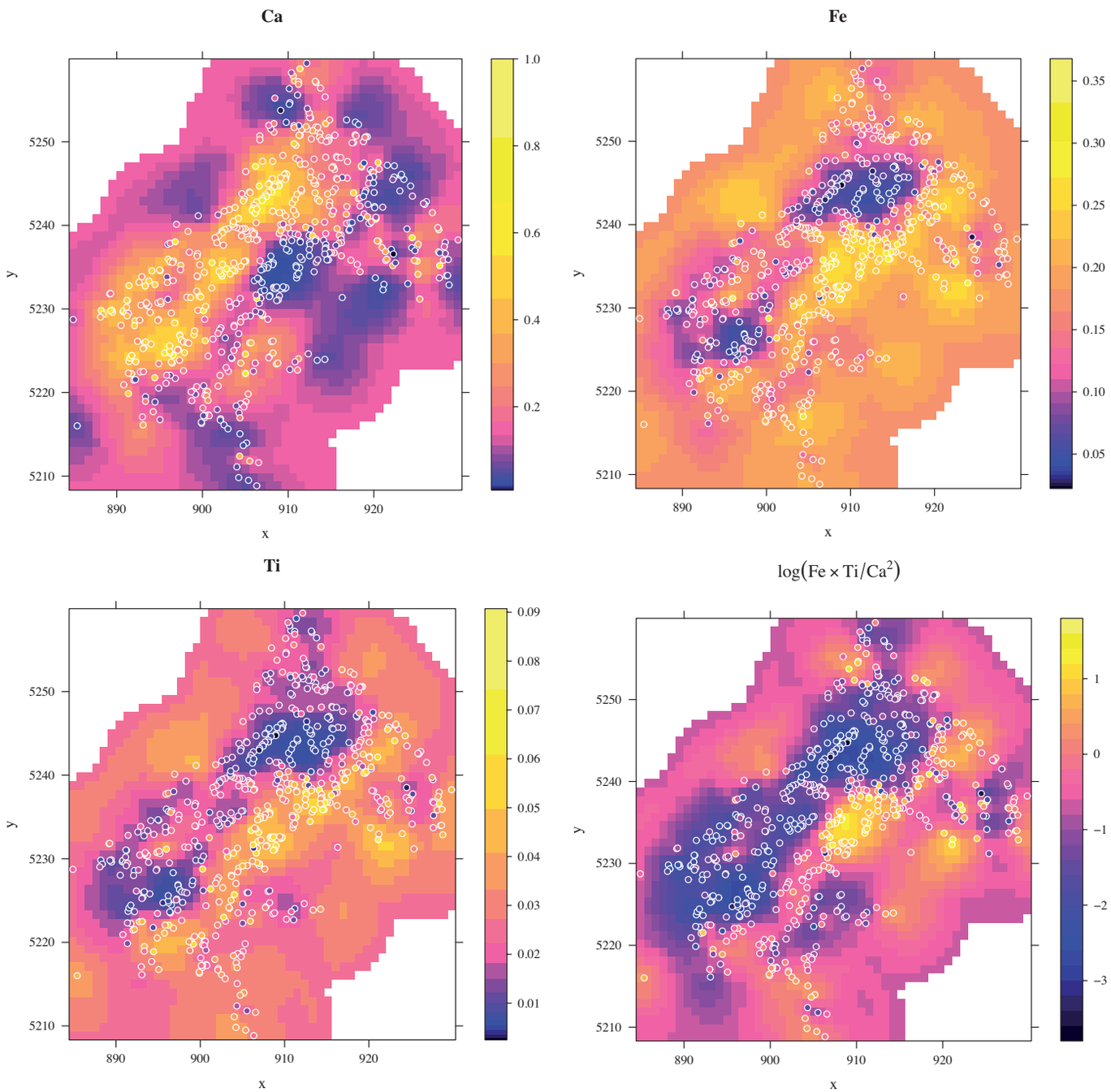

Figura 5. Mapas interpolados de la proporción de Ca, Fe y Ti, así como del balance de Ca contra los otros dos elementos.
Figura 5. Interpolated maps of the proportions between Ca, Fe and Ti, as well as of the balance of Ca against the other two components.

podremos luego anti-transformar con la Eq. (2) para obtener interpolaciones para la composición original. El resultado para las componentes Ca, Fe y Ti, así como uno para un balance proporcional a $\ln(Fe \cdot Ti / Ca^2)$, se pueden ver en la Fig. 5, comparados con los datos originales en la misma escala. Este log-cociente mostraba aproximadamente la primera componente principal de ambas estructuras (nugget y variograma esférico), y describe por tanto una de las componentes de más alta variabilidad del sistema. Comparando estas figuras con la geología (fig. 1), parece claro que este balance está controlado por la litología dominante: las metafilitas del Tonschiefer muestran valores altos del contraste (Ti-Fe) vs. Ca, mientras que los valores más bajos se muestran en las zonas de dominio dolomítico de la formación Rannach-Hochlantsch.

Varianza de krigeado y simulación

El cálculo y uso de la varianza de krigeado sigue las mismas líneas que la interpolación. Podemos usar la Eq. (8) para obtener una matriz de covarianzas de las coordendadas de la composición interpolada. Junto con la interpolación de las coordenadas descrita en la sección anterior, especificaremos la distribución del valor real del vector aleatorio de coordenadas como una normal multivariante,

$$\mathrm{ilr}[\mathbf{Y}(\vec{x}_0)) \sim N(\hat{\mathbf{z}}_0, \Sigma_{0K}).$$

Sabiendo la distribución de las coordenadas ilr de $\mathbf{Y}(\vec{x}_0)$ podemos calcular regiones de probabilidad para ellas, o simular valores alternativos. Ambos resultados pueden anti-transformarse con la transformación ilr inversa (Eq. 2), y así obtener regiones de probabilidad o simulaciones alternativas para las composiciones $\mathbf{Y}(\vec{x}_0)$. Así mismo, tambien se puede usar la distribución para calcular la probabilidad de que cumpla $\mathbf{Y}(\vec{x}_0)$ ciertas condiciones, como por ejemplo $Y_1 < Y_3$, o $Y_1 > 10\%$ y a la vez $Y_3 > 5\%$, o lo que se derive del problema práctico. Condiciones del primer tipo se pueden calcular directamente con la teoría de la distribución normal, ya que describen relaciones expresables de forma lineal en las coordenadas (piénsese en la coordenadas ad-hoc $z = \ln(Y_1 / Y_3)$: si la condición se cumple, entonces $z < 0$). En el caso de relaciones más complejas, como las del segundo ejemplo, se puede recurrir a la simulación para obtener aproximaciones de Monte Carlo de las probabilidades buscadas. Supongamos que las condiciones deseadas describen un campo dentro del simplex denotado por $\mathcal{G} \in S^D$, y que se ha obtenido K simulaciones $\{\mathbf{z}_0^1, \ldots, \mathbf{z}_0^k\}$ posibles de $\mathbf{Z}(\vec{x}_0)$; la probabilidad buscada se puede aproximar por

$$\widehat{\Pr}[\mathbf{Y}(\vec{x}_0) \in \mathcal{G}] = \frac{1}{K}\sum_{i=1}^{K} I(\mathrm{ilr}^{-1}(\mathbf{z}_0^i) \in \mathcal{G}),$$

donde la función $I(\cdot)$ vale 1 si la condición argumento se cumple y 0 en caso contrario.

Conclusiones

Tratar bases de datos con dependencia espacial y variables en porcentajes, concentraciones o proporciones (composiciones en general) es sencillo atendiendo a su naturaleza relativa. Los pasos a seguir son los siguientes:

1. Se estima la estructura espacial de la composición mediante el variaciograma. El variaciograma es el conjunto de variogramas convencionales de todos los posibles log-cocientes de dos variables de la composición.
2. Se modela el variaciograma, preferentemente con un modelo de coregionalización lineal, que considera el variaciograma como una combinación lineal de matrices de variaciones multiplicadas por correlogramas. Los autovectores de estas matrices se pueden tratar e interpretar como en el análisis de componentes principales, mostrando balances entre grupos de variables o relaciones de tipo constante de equilibrio. Como cada uno de estos autovectores se asocia a un correlograma con un alcance (y posiblemente una anisotropía) propio, dichos balances/constantes de equilibrio se pueden interpretar como procesos que ocurren a una escala ligada al alcance de ese correlograma.
3. Una vez se tiene un modelo de coregionalización, éste se puede usar para interpolar la composición, y con ello obtener mapas de las variables, o bien de ciertos log-cocientes escogidos: por ejemplo, los balances o las constantes de equilibrio ligadas a los autovectores del paso anterior como proxis de los procesos inferidos. Para ello, basta seleccionar una base de cálculo, calcular las coordenadas de la composición en esa base (como log-cocientes de variables), expresar el modelo de coregionalización para esas coordenadas (mediante productos de matrices) y aplicar programas convencionales de cokrigeado. Las coordenadas interpoladas se pueden antitransformar para obtener interpolaciones para la composición original.
4. Si se desea estudiar la variabilidad espacial, la matriz de covarianzas de krigeado y el estimador de krigeado obtenidos para las coordenadas en la base seleccionada se pueden tomar como la covarianza y la media de una distribución normal multivariante que describe la incertidumbre de la

interpolación de las coordenadas. Este resultado permite simular vectores de coordenadas, que luego se podrán antitransformar para obtener composiciones simuladas.

Agradecimientos

Este trabajo forma parte de una tesis doctoral, financiada por la Universitat de Girona, dentro de su programa de becas de investigación (Ref: BR01/03). Así mismo, se agradece la financiación de los proyectos "Modelado estadístico sobre el simplex" (MESS, Ref: BFM2003-05640) y "Modelado estadístico sobre el simplex y otros espacios restringidos" (MEASURE; Ref: MTM2006-03040), así como de "Corrientes, Oleaje y Viento: mejora del Análisis de RIesgos mediante Asimilación en esquemas Numéricos de la Costa y su Entorno" (COVARIANCE, Ref: CTM2010-19709). El autor quiere finalmente agradecer al Prof. J. Davis el acceso a la base de datos de ejemplo y a los estudios anteriores, y a los profesores V. Pawlowsky-Glahn, J.J. Egozcue y K.G. van den Boogaart la guía y las fructíferas discusiones que llevaron a este trabajo. Finalmente, quisiera agradecer a los revisores de este documento, Ricardo Olea y especialmente Carolina Guardiola, las detalladas revisiones y comentarios a la versión original del manuscrito.

References

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability.* Chapman & Hall Ltd., London (UK). (reimpreso en 2003 con material adicional por The Blackburn Press). 416 pp.

Aitchison, J. 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. En: Pawlowsky-Glahn, V. (ed.), *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology,* Centro Internacional de Métodos Numéricos en la Ingeniería (CIMNE), Barcelona, 3-35.

Aitchison, J. 2002. Simplicial inference. En: M. A. G. Viana y Richards, D. S. P. (eds.), *Algebraic Methods in Statistics and Probability,* American Mathematical Society, Providence, Rhode Island, 1-22.

Barceló-Vidal, C. 2000. *Fundamentación matemática del análisis de datos composicionales. Technical Report IMA 00-02-RR,* Departament d'Informática i Matemática Aplicada, Universitat de Girona, Spain. 77 pp.

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

Chilès, J.-P. y Delfiner, P. 1999. *Geostatistics — modeling spatial uncertainty.* Series in Probability and Statistics. John Wiley and Sons, Inc., New York, NY, 695 pp.

Clark, I. y Harper, W. V. 2000. Practical Geostatistics 2000. *Ecosse North America Llc,* Columbus Ohio, 342 pp.

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

Isaaks, E. H. y Srivastava, R.M. 1989. *An Introduction to Applied Geostatistics.* Oxford University Press, New York, 561 pp.

Journel, A. G. y Huijbregts, C. J. 1978. *Mining Geostatistics.* Academic Press, London, 600 pp.

Myers, D. E. 1984. Co-kriging: New developments. En: *Geostatistics for Natural Resources Characterization,* 2nd NATO-ASI, Stanford, 479-484.

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

Pawlowsky-Glahn, V. y Olea, R. A. 2004. *Geostatistical Analysis of Compositional Data.* Oxford University Press, USA. 204 pp.

Tolosana-Delgado, R. 2006. Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring. Tesis de doctorado, Universitat de Girona. Disponible online.

Tolosana-Delgado, R., Otero, N. y Pawlowsky-Glahn, V. 2005. Some Basic Concepts of Compositional Geometry. *Mathematical Geology,* 37 (7), 673-680.

Wackernagel, H. 1998. *Multivariate Geostatistics, An Introduction With Applications* (2.ª edición). Springer Verlag, Berlin, 291 pp.

Weber, L. y Davis, J. 1990. Multivariate statistical analysis of stream-sediment geochemistry in the Grazer Paläozoikum, Austria. *Mineralium Deposita,* 25, 213-220.

Recibido: enero 2011
Revisado: marzo 2011
Aceptado: julio 2011
Publicado: octubre 2011