Journal of Geochemical Exploration xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo

Spatial analysis of compositional data: A historical review

Vera Pawlowsky-Glahn [a,*], Juan José Egozcue [b]

[a] Dept. of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain
[b] Dept. of Applied Mathematics III, Universitat Politècnica de Catalunya, Barcelona, Spain

ARTICLE INFO

Article history:
Received 18 October 2015
Revised 17 December 2015
Accepted 18 December 2015
Available online xxxx

Keywords:
Compositional data analysis
geostatistics
Simplex
Variation-variogram
Simplicial indicator kriging

ABSTRACT

Like the statistical analysis of compositional data in general, spatial analysis of compositional data requires specific tools. A historical overview of their development is presented in three steps: (a) the recognition of the problem, known as spurious spatial covariance, (b) first attempts to use the logratio approach, and (c) the application of the principle of working in coordinates using isometric logratio representations. Also mentioned are the use of matrix-valued variation-variograms as a tool to model crossvariograms, and the simplicial approach to indicator kriging, that solves inconsistencies in the standard approach to indicator kriging.

1. Introduction

According to Chilès and Delfiner, (2012), the term *geostatistics* was introduced by Matheron, (1962) to designate his own methodology for ore reserve estimation. Since then, geostatistics expanded amazingly, as the methodology finds application in many fields, not only in geo- and environmental sciences. Independently, in the 1980's, J. Aitchison started developing *compositional data analysis* (CoDa) (Aitchison and Shen, 1980; Aitchison, 1982; Aitchison, 1986) introducing what nowadays is known as *the log-ratio approach*. Although most type of data to which geostatistics is applied are compositional, like ore grade, chemical or mineralogical composition of rocks, contaminants in air or water, it was not recognised until 1984 that spurious spatial correlation might be at work (Pawlowsky, 1984). We summarise in what follows the steps that have been undertaken since then to solve the problems derived from the compositional character of some spatially dependent data. We limit our contribution to the historical development, omitting most formal derivations which can be found in the references cited.

2. Spurious spatial covariance

The problem of spurious spatial covariance of regionalized compositions, or *r-compositions* for short, was first stated in Pawlowsky, (1984). The results are illustrative, and are therefore briefly exposed.

According to our present understanding, a random vector, **Z**, with D strictly positive components representing parts of a whole, is a composition if it carries only relative information (Pawlowsky-Glahn et al., 2015c). Note that the term *relative information* is equivalent to *information lies in the ratios between components*, not in the absolute values. The same definition holds for a spatially distributed random vector, Z(x), at any point x of a spatial domain $\mathcal{R}$.

In 1984, r-compositions were still understood as random vectors subject to a constant sum constraint, or *closed r-compositions*. We know now that compositions in general, and r-compositions in particular, are equivalence classes, and that a closed composition is just a representation. This means, that the results obtained under this assumption hold for any representation of the equivalence classes.

For the understanding of spurious spatial covariance or correlation, it is mathematically easier to work with a closed representation. Therefore, in what follows, we work with a *closed r-composition*, i.e. with a spatially distributed random vector, Z (x), with D strictly positive parts or components, that is subject to a constant sum constraint for all $x \in \mathcal{R}$,

$$\sum_{i=1}^{D} Z_i(x) = \kappa, \tag{1}$$

with κ a given positive constant depending on the units of the random vector. The constant κ is usually 1 (parts per unit), 100 (percentages), or 10^6 (parts per million).

Following Matheron, (1965), geostatistics can be used with regionalized variables satisfying stationarity conditions. Second order stationarity requires regionalized variables to have a constant mean and the autocovariance only depending on the lag between pairs of variables **Z** (x_i) and **Z** (x_j); a less stringent condition is the *intrinsic hypothesis*, which assumes that the first order differences are second order stationary. Under these kind of assumptions, geostatistics builds on modelling the mean and the spatial

* Corresponding author.
E-mail address: vera.pawlowsky@udg.edu (V. Pawlowsky-Glahn).

http://dx.doi.org/10.1016/j.gexplo.2015.12.010
0375-6742/

autocovariance, or related parameters, like the variogram. The following development handles the components of the closed r-composition $\mathbf{Z}(x)=(Z_1(x),Z_2(x),\ldots,Z_D(x))$ at two spatial locations, say x and $x+h$ in $\mathcal{R}$, where h denotes the lag between them.

From Eq. (1), for any lag h it holds

$$\sum_{i=1}^{D}(Z_i(x)-Z_i(x+h))=\sum_{i=1}^{D}Z_i(x)-\sum_{i=1}^{D}Z_i(x+h)=\kappa-\kappa=0. \tag{2}$$

Hence, multiplying both sides of Eq. (2) by $(Z_j(x)-Z_j(x+h))$,

$$\sum_{i=1}^{D}(Z_i(x)-Z_i(x+h))(Z_j(x)-Z_j(x+h))=0.$$

for any $j=1,2,\ldots,D$. Taking expectations,

$$\sum_{i=1}^{D}cov\left[(Z_i(x)-Z_i(x+h)),(Z_j(x)-Z_j(x+h))\right]=0. \tag{3}$$

Given that a variance is always positive, Eq. (3) can be rewritten for any $j=1,2,\ldots,D$, as

$$var\left[(Z_j(x)-Z_j(x+h))(Z_j(x)-Z_j(x+h))\right]=-\sum_{i\neq j}cov\left[(Z_i(x)-Z_i(x+h))(Z_j(x)-Z_j(x+h))\right]. \tag{4}$$

Note that Eq. (4) depends only on the fact that $\mathbf{Z}(x)$ is the closed representation of an r-composition, and not on the type of spatial dependence of its components. Eq. (4) implies that non-stochastic factors determine the value of cross-covariances. They cannot be all null simultaneously, as the variance is, by definition, always positive. Also, if the closed r-composition was generated by closure of independent random variables, a dependence will appear, which is spurious, as it is not generated by the phenomenon itself (Pawlowsky, 1984). This result is well known for compositional data in general as the *closure problem* (Chayes, 1960). It has many implications in standard multivariate analysis which can be directly extended to r-compositions.

For a closed intrinsic r-composition $\mathbf{Z}(x)$, Eq. (4) can be written in terms of variograms, $\gamma_j(h)$, and crossvariograms, $\gamma_{ij}(h)$,

$$\gamma_j(h)=-\sum_{i\neq j}\gamma_{ij}(h),\quad j=1,2,\ldots,D. \tag{5}$$

for any lag h. As stated in Pawlowsky, (1984), the obvious conclusion is the need of non-zero cross-variograms for r-compositions, some of which have to be negative—as the variogram is, by definition, positive. It is clear that the only case in which cross-variograms could be all null or all positive is that the variogram is null, i.e. the r-composition is constant. The fact that variograms and cross-variograms of r-compositions are subject to non-stochastic controls leads to the conclusion that, when based on raw data, they are spurious.

Under the assumption that the sample space is the whole real space endowed with the standard Euclidean space structure and geometry, or a subset with the induced structure and geometry, for $\mathbf{Z}(x)$ satisfying the second order stationary hypothesis, the following equalities hold:

$$\begin{aligned}\sum_{i=1}^{D}Z_i(x)&=\kappa,\\ \sum_{i=1}^{D}E(Z_i(x))&=\sum_{i=1}^{D}m_i=\kappa,\\ \sum_{i=1}^{D}(Z_i(x)-m_i)&=0,\end{aligned} \tag{6}$$

with $E(Z_i(x))=m_i$, the expected value of $Z_i(x)$, $i=1,2,\ldots,D$. Multiplying both sites of Eq. (6) by $(Z_j(x)-m_j)$ and taking expectations, it holds

$$\sum_{i=1}^{D}cov\left[(Z_i(x)-m_i)(Z_j(x)-m_j)\right]=0,\quad j=1,2,\ldots,D, \tag{7}$$

and therefore, for any lag h,

$$C_j(h)=-\sum_{i\neq j}C_{ij}(h),\quad j=1,2,\ldots,D, \tag{8}$$

where $C_j(h)$ stands for the auto-covariance of component j, and $C_{ij}(h)$ for the cross-covariance of components i and j. Consequently, also the cross-covariances cannot be all null, and some of them have necessarily to be negative. Being subject to algebraic, non-stochastic, controls, they are spurious.

As summarised in Pawlowsky-Glahn and Burger, (1992), the problems derived from the nature of spatially distributed compositional data, when the raw data are analysed, are

1. The mathematical necessity of at least one non-zero cross-covariance.
2. The bias towards negative cross-covariances.
3. The singularity of the cross-covariance matrix for any lag h.
4. The distorted description and interpretation of the spatial dependence between the compositional variables under study.

   Nowadays we know that the problem of spurious spatial covariance or correlation is generated by the fact that compositional data are analysed as *real data*, with the usual Euclidean geometry. In fact, most statistical methods have been developed for real data without constraints under the implicit assumption that the Euclidean geometry holds. This means that the difference between observations is measured as an absolute difference, that the sum and its opposite make sense. This holds even with constraints, i.e. restricting the support of the sample to a subset of real space without changing the geometry.

3. The beginning — 1986: the additive log-ratio approach

The initial approach (Pawlowsky, 1986; Pawlowsky-Glahn and Olea, 2004) was to use the additive log-ratio (*alr*) transformation (Aitchison, 1982; Aitchison, 1986). The r-composition is transformed into log-ratios as

$$\mathbf{W}(x)=\left(\ln\frac{Z_1}{Z_D},\ln\frac{Z_2}{Z_D},\ldots,\ln\frac{Z_{D-1}}{Z_D}\right),.$$

thus obtaining a regionalized vector of D–1 components which can be treated using cokriging. As we are aware nowadays, this was done under the implicit assumption that the Euclidean geometry holds for *alr* transformed vectors. Under this assumption the *alr*-transformation leads to BLU (Best Linear Unbiased) estimates (Pawlowsky-Glahn and Egozcue, 2002). Nevertheless, soon problems appeared, like the fact that cokriging seamed to lead to worse results than kriging, a fact that stands in contradiction with theoretical results (Pawlowsky-Glahn and Olea, 2004, p. 160–161). The reasons for these problems could not be explained in a consistent way until the algebraic–geometric structure of the sample space of compositional data was recognised (Aitchison et al., 2002; Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) and the *alr* was understood within this framework. Essentially, the problem was the computation of variances and covariances using the *alr* coordinates, which at that moment was not clear.

The covariance structure of compositional data can be described by the so-called variation matrix (Aitchison, 1982; Aitchison, 1986). This matrix contains the variances of each possible log-ratio of pairs of compositional parts. It was shown that the variation matrix completely describes the covariance structure of the composition, independently of which transformation is used to analyse the data. These facts inspired the introduction of the spatial structure of r-compositions, first defined

Please cite this article as: Pawlowsky-Glahn, V., Egozcue, J.J., Spatial analysis of compositional data: A historical review, J. Geochem. Explor. (2015), http://dx.doi.org/10.1016/j.gexplo.2015.12.010

in Pawlowsky, (1986) and summarised in Pawlowsky-Glahn and Burger, (1992) and Pawlowsky-Glahn and Olea, (2004), p. 29:

DEFINITION 3.1. [Spatial covariance structure] The spatial covariance structure of a D-part r-composition is defined as the set of functions of the lag h.

$$\sigma_{ij\cdot k\ell}(h) = \mathrm{Cov}(\ln\tfrac{Z_i(x)}{Z_k(x)},\ \ln\tfrac{Z_j(x+h)}{Z_\ell(x+h)}), \qquad i,j,k,\ell\in 1,2,\ldots,D, x\in\mathcal{D}..$$

At a first glance, the geostatistical analysis of $\mathbf{W}(x)$ can be performed as a cokriging. This means that variograms and cross-variograms have to be fitted to their empirical versions. However, the spatial covariance structure allows the modelling of each component of $\sigma_{ij\cdot k\ell}(h)$ by a simple variogram, thus avoiding modelling of cross-variograms. A matrix transformation can transform the spatial covariance structure into the cross-variograms required for a cokriging of $\mathbf{W}(x)$.

As stated in Pawlowsky-Glahn and Burger, (1992), the most difficult part—compared to a spatial analysis of several variables—is that, in addition to the usual difficulties, problems have to be reformulated in terms of logratios, and interpretation and description of spatial dependencies have to be made in the same terms.

4. The breakthrough 2000

Around the year 2000, compositional data analysis attains a further maturity level. The achievements can be summarised in two main points: (1) the simplex, as sample space of compositional data, is endowed with a Euclidean space structure, called Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001); and (2) compositional data are no longer conceived as vectors constrained to a constant sum but as equivalence classes of proportional vectors with positive components (Barceló-Vidal et al., 2001). These new points of view influenced the way of identifying and analysing r-compositions and they are briefly described in the following sections.

Subsequent developments (Tolosana-Delgado, 2006), based on the sample space approach and the *Principle of Working in Coordinates* (Mateu-Figueras et al., 2011; Pawlowsky-Glahn, 2003), proved the potential for the log-ratio approach within the Aitchison geometry of the simplex, setting the foundations for a rigorous theory. Based on the principles of scale invariance, subcompositional dominance, and permutation invariance, the operations of perturbation, powering, and the inner product associated to the distance introduced by (Aitchison, 1982; Aitchison, 1986; Aitchison, 1997), provide, as mentioned, the simplex with a Euclidean space structure (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001), different, but nonetheless isometric to the Euclidean space structure of real space. The Euclidean space structure of the simplex was termed *Aitchison geometry* in Pawlowsky-Glahn and Egozcue (2001). It opened up the door to a deeper understanding of the nature of compositional data, of the available methods to analyse them, and of the problems linked to different approaches. In particular, the advantage of using isometric log-ratio transformations was recognised. Within this family of transformations, those known as balances (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005) have shown a high potential based on their interpretability, and can be used for spatial analysis of compositional data.

4.1. Compositions are representatives of equivalence classes

In Aitchison, (1986), the so called *principle of scale invariance of compositions* was formulated. It states that the analysis of a composition must remain invariant when the composition is multiplied by any positive constant. This was the motivation for preconising the use of log-contrasts as the main tool in the analysis. Log-contrasts are combinations of the logarithms of the parts such that, when all parts of the composition are multiplied by a positive constant, the value of the combination remains unaltered. Also, vectors of positive components are reduced to constant sum by using the closure operation. These concepts were clear from the beginning of compositional data analysis, but there was a lack of mathematical formulation reflected in the wording of compositional data analysis. For instance, when referring to the closure problem as the only origin of pitfalls in compositional data analysis.

The progress consists in thinking that all vectors having proportional positive components are equivalent and convey the same compositional information. A composition is then an equivalence class which can be represented by choosing an arbitrary element of the class. Equivalence classes can be represented in many ways and each choice defines a potential sample space, whether constraint to a constant sum or not (see explanations in Pawlowsky-Glahn et al. (2015c), ch. 2). When compositional data are represented as data subject to a constant sum constraint, their sample space is a simplex, and the simplex is nothing else but a choice of one out of all the possible sample spaces of compositions. This choice is not only convenient because it is the usual choice in practice, but also because it is mathematically easy to define a meaningful and interpretable Euclidean vector space structure in the simplex (Pawlowsky-Glahn and Egozcue, 2001).

Other representations of compositions are possible. For instance, when compositions of air pollutants are expressed in $\mu g/m^3$ or solutes are given in Mol per litre, concentrations do not add to a constant and they are not represented in the simplex. Simply, the representative of the equivalence class has been taken in another way, but still the ratios of the parts are the relevant information. In these kind of representations, perturbation is also easily interpretable. Other possibilities are less intuitive, for instance, when compositions are represented in an orthant of a hypersphere (e.g. Wang et al., 2007).

It is remarkable that this interpretation of compositions as equivalence classes only arose in 2000. This may be the reason why, in the decade from 1980 to 1990, concentrations in units like mol/l, concentration of a single element, or just removing a large component, were considered to be non-compositional and, consequently, free of the difficulties of analysing compositional data.

The influence of these new concepts in geostatistics is reflected in the identification of what is an r-composition, independently of whether the collected data are closed to a constant or not.

4.2. Aitchison geometry of the simplex and consequences

The simplex endowed with perturbation (the compositional sum), powering (compositional multiplication by real numbers) and Aitchison distance, constitute a (D–1)-dimensional Euclidean vector space (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001). The Euclidean space structure of the simplex was termed *Aitchison geometry* in Pawlowsky-Glahn and Egozcue, (2001). The value of this mathematical result is supported by the fact that perturbation is an interpretable operation in most compositional scenarios. In fact, perturbation can be interpreted as filtering in geochemistry or particle size analysis; or as the Bayes formula for probabilities (for details, see Pawlowsky-Glahn et al. (2015c), ch. 2).

The Aitchison geometry points out that orthonormal basis of the space exist, and that the corresponding (Cartesian) coordinates can efficiently represent compositions; orthogonal projections are possible; the concepts of linear combination, linear dependence, Euclidean distances, and all the typical geometrical elements are available. All these tools are readily used once compositions are represented by their coordinates with respect to a basis of the space, as perturbation is the sum in coordinates, powering is scaling in coordinates, and the Aitchison distance is the standard Euclidean distance between coordinates. This constitutes the core of the *Principle of Working in Coordinates* (Mateu-Figueras et al., 2011).

An important step ahead is the construction of orthonormal (Cartesian) coordinates in the simplex. The function assigning orthonormal (Cartesian) coordinates to a composition has been named isometric log-ratio transformation (*ilr*) (Egozcue et al., 2003). The *ilr*-

Please cite this article as: Pawlowsky-Glahn, V., Egozcue, J.J., Spatial analysis of compositional data: A historical review, J. Geochem. Explor. (2015), http://dx.doi.org/10.1016/j.gexplo.2015.12.010

transformation is not unique, as there are infinitely many basis of the space. As a consequence, it was clear that the *alr*-transformation is an assignation of coordinates with respect to an oblique basis (Egozcue et al., 2003), while the centred log-ratio transformation (*clr*) (Aitchison, 1986)

$$\mathrm{clr}(\mathbf{Z}(x)) = \left(\ln \frac{Z_1(x)}{g(\mathbf{Z}(x))}, \ln \frac{Z_2(x)}{g(\mathbf{Z}(x))}, \ldots, \ln \frac{Z_D(x)}{g(\mathbf{Z}(x))} \right),$$

where $g(\mathbf{Z}(x))$ is the geometric mean of the $\mathbf{Z}(x)$ components, gives coordinates with respect to a generating system of the simplex. The *clr*-transformation was not used for geostatistical analysis, as its covariance matrix is always singular. It is, nevertheless, extremely useful for computation in compositional data analysis. For example, the *ilr*-coordinates are readily obtained through a *clr*-transformation as

$$\mathrm{ilr}(\mathbf{Z}(x)) = V\ \mathrm{clr}(\mathbf{Z}(x)),$$

where V, called contrast matrix (Egozcue et al., 2011; Pawlowsky-Glahn et al., 2015c), is a (D,D−1)-matrix satisfying the property that $V'V$ is the identity matrix. An easy way of building coordinates, called balances, was introduced in Egozcue et al. (2003); Egozcue and Pawlowsky-Glahn, (2005). The procedure, called sequential binary partition (SBP), provides such contrast matrices, and the resulting *ilr*-coordinates are called *balances* (Egozcue and Pawlowsky-Glahn, 2005).

It is remarkable that *ilr*, *alr*, and *clr* transformations are different assignations of coordinates to a composition more than different transformations leading to different approaches. An important point is that, within the Aitchison geometry of the simplex, the predictors used in all classes of kriging are linear, as they are linear combinations of coordinates. However, in the case of *alr*-coordinates distances and covariances should be handled very carefully, paying special attention to the fact that they are representations in an oblique coordinate system. This explains the problems detected when using cokriging on *alr*-coordinates (Pawlowsky-Glahn and Olea, 2004), p. 108, where this fact was not taken into account.

4.3. Cokriging of regionalized compositions

Initially, the problems for cokriging of r-compositions appeared to be centred on the modelling of cross-variograms of log-ratio transformed data, although it was known that a simple matrix transformation leads from the matrix-valued variation-variogram, the matrix of variograms of all possible simple log-ratios, to any log-ratio representation (Pawlowsky, 1986; Pawlowsky-Glahn and Burger, 1992; Pawlowsky-Glahn and Olea, 2004; Tolosana-Delgado, 2006; Tolosana-Delgado et al., 2011). Later, Tolosana-Delgado and Boogaart, (2013) recognised the potential of the matrix-valued variation-variogram, specially to model cross-variograms using first a Linear Model of Coregionalisation for the matrix-valued variation-variogram, followed by a matrix transformation to obtain the corresponding variograms and cross-variograms for the coordinates chosen by the scientist to represent the available data. Note that the matrix-valued variation-variogram is a matrix with all its entries simple variograms and no cross-variogram. Standard cokriging can then be applied to obtain the desired predictions. In summary, spatial compositional data analysis consists in the following steps (Tolosana-Delgado and Boogaart, 2013):

1. transform the D-part compositional vectors into (D−1)-dimensional real vectors by means of a convenient isometric log-ratio (*ilr*) transformation;
2. apply any standard geostatistical technique to the vectors obtained;
3. back-transform interpolated and/or simulated scores back using the *ilr* inverse.

To model necessary variograms and cross-variograms

- compute the matrix-valued variation matrix and adjust a Linear Model of Coregionalisation;
- apply the corresponding matrix transformation to obtain the desired matrix-valued variogram (containing variograms in the diagonal and cross-variograms off-diagonal) of the *ilr* transformation used before.

Details of the procedure can be found in Tolosana-Delgado and Boogaart, (2013).

Note that, as stated in Tolosana-Delgado et al., (2008a), the proposed procedure leads to BLU estimators when performing cokriging.

4.4. Simplicial Indicator Kriging

The recognition of the Euclidean vector space structure of the sample space of compositional data and the understanding that probabilities can be considered to be a composition allowed to solve the problems intrinsic to Indicator Kriging (Pawlowsky-Glahn et al., 2006; Tolosana-Delgado, 2006; Tolosana-Delgado et al., 2008c; Tolosana-Delgado et al., 2008b). By construction, Simplicial Indicator Kriging avoids all the known problems associated with usual Indicator Kriging (Journel, 1983), namely negative predictions, order relation violations, or predictions larger than one.

4.5. Further developments — 2015: cokriging r-compositions with a total

As mentioned before, compositional data are multivariate positive real data that carry only relative information, and can be represented simply taking closure, i.e. taking proportions or concentrations. In this case, the information about their total sum is lost. In some cases, in addition to the composition, the sum of some of the positive variables, called total, can be informative or of interest. Consequently, the need of a joint analysis of composition and total arises. Some possibilities were studied in Pawlowsky-Glahn et al., (2015a) which concluded that the chosen total can be included as an additional coordinate to those coming from the composition. This applies to r-compositions where some regionalized total is of interest. The geostatistical analysis can be conducted by cokriging of compositional *ilr*-coordinates, jointly with the coordinate of the total.

A first application of this procedure was performed in Pawlowsky-Glahn et al., (2015b) although the main goal was dimension reduction of a geochemical data set. The problem appears when applying compositional techniques of dimension reduction since, after orthogonal projections, the original units of the composition are lost. In order to recover original units, cokriging of *ilr*-coordinates of the composition is carried out jointly with the sum of initial concentrations. This joint cokriging of *ilr*-coordinates with supplementary real variables appears to be a promising technique in compositional geostatistics.

5. Other approaches

Not many attempts have been made to find spatial interpolation methods for regionalized compositional data. Methods that comply with nonnegativity and the representation as data constraint to a constant sum include nearest neighbour interpolation, triangulation, local sample (arithmetic) mean, and inverse distance interpolation, which are described in (Isaaks and Srivastava, 1989). Another approach, called *compositional kriging*, was introduced by (Walwoort and de Gruijter, 2001). All of them are implicitly based on the assumption that the sample space of compositional data is the simplex as a constraint subset of real space, and that they obey the induced geometry, i.e. the standard Euclidean geometry. This fact implies the assumption

Please cite this article as: Pawlowsky-Glahn, V., Egozcue, J.J., Spatial analysis of compositional data: A historical review, J. Geochem. Explor. (2015), http://dx.doi.org/10.1016/j.gexplo.2015.12.010

that compositional data carry absolute and not relative information, a decision that lies with the researcher analysing the data. Furthermore, as stated by Walwoort and de Gruijter, (2001), the former methods do not take the spatial structure into account, but neither does *compositional kriging* completely, as it does not take into account cross-correlations, and thus cross-variograms, to avoid problems with spurious correlation. As shown by Pawlowsky-Glahn and Egozcue, (2002), even using the *alr* representation of compositional data leads to BLU estimators within the Aitchison geometry of the simplex (Pawlowsky-Glahn and Egozcue, 2001), and numerical comparisons of results based on different assumptions for the structure of the sample space make no sense. Whichever is the assumption made by the scientist, spatial interpolation using cokriging will be optimal within the assumed geometry.

6. Conclusions and comments

Reviewing the early developments in the spatial analysis of compositional data, and in the analysis of compositional data in general, one can see the evolution of the way of thinking on that type of data. One typical example is the statement in Pawlowsky, (1984) that $\mathbf{Z}(x) - \mathbf{Z}(x+h)$ is an r-composition for any $x \in \mathcal{R}$ and any lag h. This is clearly not true, as it always yields at least some non-positive numbers.

Another hurdle were the problems related to the *alr* transformation. After understanding that the *alr* represents the data in an oblique basis of the simplex, one can recognise two ways of proceeding: (1) to avoid the *alr* and use only isometric log-ratio transformations, or (2) to take into account the oblique nature and use appropriate matrix transformations to obtain consistent results. The first approach is straightforward and safe, the second requires more care. It is up to the researcher to choose which transformation is better suited for the case he or she is dealing with.

One of the characteristics of cokriging *ilr*-coordinates is that the modelling of cross-variograms can be afforded modelling the variation variograms, thus avoiding the always difficult cross-variogram modelling.

The main conclusion is that analysing compositional data, regionalized or not, is nowadays summarised by the *principle of working on coordinates*; it transforms the compositional analysis into a standard geostatistical problem where well known procedures can be applied without additional difficulties.

Acknowledgements

The authors thank two anonymous reviewers for their suggestions and comments. This research has been supported by the Spanish Ministry of Education and Science under project 'METRICS' (Ref. MTM2012-33236); and from the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B Stat Methodol. 44 (2), 139–177.
Aitchison, J., 1986. The statistical analysis of compositional data. Monographs on statistics and applied probability. London (UK). Chapman & Hall Ltd., London (UK) (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.
Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, V. (Ed.), Proceedings of IAMG'97 - The III Annual Conference of the International Association for Mathematical Geology. Barcelona (E). International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), pp. 3–35 (1100 p ; volume I, II and addendum).
Aitchison, J., Shen, S.M., 1980. Logistic-normal distributions. Some properties and uses. Biometrika 67 (2), 261–272.
Aitchison, J., Barceló-Vidal, C., Egozcue, J.J., Pawlowsky-Glahn, V., 2002. A concise guide for the algebraic–geometric structure of the simplex, the sample space for compositional data analysis. In: Bayer, U., Burger, H., Skala, W. (Eds.), Proceedings of IAMG'02 – The VIII Annual Conference of the International Association for Mathematical Geology. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, pp. 387–392 (1106 p ; volume I and II).
Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2001. Mathematical foundations of compositional data analysis. In: Ross, G. (Ed.), Proceedings of IAMG'01. The VII Annual Conference of the International Association for Mathematical Geology, Cancun (Mex), p. 20.
Billheimer, D., Guttorp, P., Fagan, W., 2001. Statistical interpretation of species composition. J. Am. Stat. Assoc. 96 (456), 1205–1214.
Chayes, F., 1960. On correlation between variables of constant sum. J. Geophys. Res. 65 (12), 4185–4193.
Chilès, J.P., Delfiner, P., 2012. Geostatistics — modeling spatial uncertainty. Probability and Statistics. United States of America, second ed. Wiley.
Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37 (7), 795–828.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35 (3), 279–300.
Egozcue, J.J., Barceló-Vidal, C., Martín-Fernández, J.A., Jarauta-Bragulat, E., Díaz-Barrero, J.L., Mateu-Figueras, G., 2011. Elements of simplicial linear algebra and geometry. Pawlowsky-Glahn and Buccianti, pp. 141–157 378 p.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY (USA) (592 p).
Journel, A.G., 1983. Nonparametric estimation of spatial distributions. Math. Geol. 15 (3), 445–468.
Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The principle of working on coordinates. Pawlowsky-Glahn and Buccianti, pp. 31–42 378 p.
Matheron G. Traité de Géostatistique Appliquée. Volume I of Mémoires du Bureau de Recherches Géologiques et Minières (France), 14 . Technip, Paris (F) , 1962. 333 p.
Matheron, G., 1965. Les Variables Régionalisées et Leur Estimation–une Application de la Théorie des Fonctions Aléatoires aux Sciences de la Nature. Masson et Cie., Paris (F) 305 p.
Pawlowsky, V., 1984. On spurious spatial covariance between variables of constant sum. Sci. Terre, Sér Informatique 21, 107–113.
Pawlowsky, V., 1986. Räumliche Strukturanalyse und Schätzung ortsabhängiger Kompositionen mit Anwendungsbeispielen aus der Geologie (Ph.D. thesis) Fachbereich Geowissenschaften. Freie Universität Berlin, Berlin (D) (170 p).
Pawlowsky-Glahn V. Statistical modelling on coordinates. In: Thió-Henestrosa S., Martín-Fernández J.A. Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop. Girona (E): Universitat de Girona, ISBN 84–8458-111-X, (http://ima.udg.es/Activitats/CoDaWork2003/) ; 2003.
Pawlowsky-Glahn, V., Burger, H., 1992. Spatial structure analysis of regionalized compositions. Math. Geol. 24 (6), 675–691.
Pawlowsky-Glahn, V., Egozcue, J.J., 2001. Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. 15 (5), 384–398.
Pawlowsky-Glahn, V., Egozcue, J.J., 2002. BLU Estimators and Compositional Data. Math. Geol. 34 (3), 259–274.
Pawlowsky-Glahn, V., Olea, R.A., 2004. In: DeGraffenreid, J.A. (Ed.), Geostatistical analysis of compositional data. Number 7 in studies in mathematical geology. Oxford University Press.
Pawlowsky-Glahn, V., Tolosana-Delgado, R., Egozcue, J.J., 2006. Simplicial Indicator Kriging (4 p).
Pawlowsky-Glahn, V., Egozcue, J.J., Lovell, D., 2015a. Tools for compositional data with a total. Stat. Model. 15 (2), 175–190. http://dx.doi.org/10.1177/1471082X14535526, Nov. 25, 2014.
Pawlowsky-Glahn, V., Egozcue, J.J., Olea, R.A., Pardo-Igúzquiza, E., 2015b. Cokriging of compositional balances including a dimension reduction and retrieval of original units. J. South. Afr. Inst. Min. Metall. 115, 59–72.
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015c. Modeling and analysis of compositional data. Statistics in practice. John Wiley & Sons, Chichester UK 272 pp.
Tolosana-Delgado, R., 2006. Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring (URL: http://www.tesisenxarxa.net/TDX-0123106-122444/index_an.html). (198 p).
Tolosana-Delgado, R., Boogaart, K.G.v.d., 2013. Joint consistent mapping of high-dimensional geochemical surveys. Math. Geosci. 45, 983–1004.
Tolosana-Delgado, R., Egozcue, J.J., Pawlowsky-Glahn, V., 2008a. Cokriging of compositions: log-ratios and unbiasedness. In: Ortiz, J.M., Emery, X. (Eds.), Geostatistics Chile 2008. Gecamin Ltd., Santiago, Chile, pp. 299–308 (2 vols, 1188 p).
Tolosana-Delgado, R., Pawlowsky-Glahn, V., Egozcue, J.J., 2008b. Indicator kriging without order relation violations. Math. Geosci. 40, 327–347.
Tolosana-Delgado, R., Pawlowsky-Glahn, V., Egozcue, J.J., 2008c. Simplicial indicator kriging. J. China Univ. Geosci. 19, 65–71.
Tolosana-Delgado, R., Boogaart, K.G.v.d., Pawlowsky-Glahn, V., 2011. Geostatistics for compositions. Pawlowsky-Glahn and Buccianti, pp. 73–86 378 p.
Walwoort, D.J.J., de Gruijter, J.J., 2001. Compositional kriging: a spatial interpolation method for compositional data. Math. Geol. 33 (8), 951–966.
Wang, H., Liu, Q., Mok, H.M.K., Fu, L., Tse, W.M., 2007. A hyperspherical transformation forecasting model for compositional data. Eur. J. Oper. Res. 179, 459–468.

Further Reading

Pawlowsky-Glahn, V., Buccianti, A. (Eds.), 2011. Compositional data analysis: theory and applications. John Wiley & Sons (378 p).

Please cite this article as: Pawlowsky-Glahn, V., Egozcue, J.J., Spatial analysis of compositional data: A historical review, J. Geochem. Explor. (2015), http://dx.doi.org/10.1016/j.gexplo.2015.12.010