Journal of Geochemical Exploration 164 (2016) 28–32

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 30 December 2015

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 $(Z_j(x) - Z_j(x+h))$,

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

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

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

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