Noname manuscript No.
(will be inserted by the editor)

Units recovery methods in compositional data analysis

J.A. Martín-Fernández[1,4], J.J. Egozcue[2], R.A. Olea[3] and V. Pawlowsky-Glahn[1]

Received: date / Accepted: date

Compositional data carry relative information. Hence, their statistical analysis has to be performed on coordinates with respect to a log-ratio basis. Frequently, the modeler is required to back transform the estimates obtained with the modeling to have them in the original units such as euros, kg or mg/liter. Approaches for recovering original units need to be formally introduced and its properties explored. Here we formulate and analyze the properties of two procedures: a simple approach consisting of adding a residual part to the composition and an approach based on the use of an auxiliary variable. Both procedures are illustrated using a geochemical data set where the original units are recovered when spatial models are applied.

KEY WORDS: Aitchison geometry, Logratio, Percentages, Simplex, Spatial analysis.

INTRODUCTION: THE PRACTICAL PROBLEM

Compositional data (CoDa) conveys relative information that is meaningful when expressed in the form of ratios between parts. These data are common in environmental and geochemical studies when the constituents and compounds are described in terms of their concentration in air (Jarauta-Bragulat et al. 2016), water (Olea et al. 2018), or in terms of solids and other wastes (Edjabou et al. 2017). When one decides to analyze a data set $\mathbf{X}$ ($n \times D$; rows$\times$columns) using compositional methods, such as weight (kg) of different materials in waste data, one is assuming that any observation $\mathbf{x}$ (a row of $\mathbf{X}$) is a member of an equivalence class (Barceló-Vidal and Martín-Fernández 2016). That is,

[1]Dpto. Informática, Matemática Aplicada y Estadística, Universidad de Girona, Spain
[2]Dpto. Ingeniería Civil y Ambiental, Universidad Politécnica de Cataluña, Barcelona, Spain
[3]U.S. Geological Survey,12201 Sunrise Valley Drive, Mail Stop 956, Reston, VA 20192, USA
[4]To whom correspondence should be addressed; E-mail: josepantoni.martin@udg.edu

the relative information contained in $\mathbf{x}$ is the same as in $k \cdot C(\mathbf{x})$ for any real scalar $k > 0$ and $C(\cdot)$ the closure operation defined by

$$C(\mathbf{x}) = \left(\frac{x_1}{\sum x_j}, \frac{x_2}{\sum x_j}, \ldots, \frac{x_D}{\sum x_j} \right). \tag{1}$$

This property is known as scale invariance (Aitchison 1986). Importantly, CoDa occupy a quotient space (Barceló-Vidal and Martín-Fernández 2016). A representative of the quotient space is the D-part unit simplex $S^D = \{\mathbf{p} \in R^D : p_j > 0,\ j = 1, \ldots, D;\ \sum_{k=1}^{D} p_k = 1\}$, that is, in practice, for convenience, compositions are commonly expressed as a vector of proportions $\mathbf{p} \in S^D$. Following Barceló-Vidal and Martín-Fernández (2016), a logarithmic isomorphism between the quotient spaces S^D, which is governed by Aitchison geometry (Pawlowsky-Glahn et al. 2015c), and the hyperplane $Z^D = \{\mathbf{z} \in R^D : \sum_{j=1}^{D} z_j = 0\}$ can be defined. Accordingly, a composition $\mathbf{x}$ can be expressed in terms of the vector $\mathbf{z} = (\ln(x_1/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x})))$, where $g(\mathbf{x})$ is the geometric mean of $\mathbf{x}$. The vectors $\mathbf{z} \in Z^D$, known as the centered log-ratio (clr) vectors (Aitchison 1986), are in a hyperplane of dimension $D - 1$. The inner product, distance and norm in S^D can be defined via the *clr* variables (Barceló-Vidal and Martín-Fernández 2016). These metric elements are used to construct orthonormal log-ratio bases in S^D. A composition $\mathbf{x}$ can be expressed in terms of its corresponding orthonormal log-ratio (olr) coordinates $\mathbf{y} = \mathrm{olr}(\mathbf{x}) = (y_1, \ldots, y_{D-1})$ (Egozcue and Pawlowsky-Glahn 2019; Martín-Fernández 2019), where, for example

$$y_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}}, j = 1, \ldots, D-1.$$

Ratios and logratios cannot be computed when one of the parts is zero or missing. Methods to deal with this problem have been described in numerous papers. Readers will find a general description in Palarea-Albaladejo and Martín-Fernández (2015). Importantly, a composition $\mathbf{x}$ and any member of its equivalence class have the same log-ratio coordinates (Barceló-Vidal and Martín-Fernández 2016). Conversely, given a vector of coordinates $\mathbf{y} = \mathrm{olr}(\mathbf{x})$ one can easily recover the original composition $\mathbf{x}$ using the procedure

$$\mathbf{x} = (\sum_{j=1}^{D} x_j) \cdot C(\mathrm{olr}^{-1}(\mathbf{y})). \tag{2}$$

The term $C(\mathrm{olr}^{-1}(\mathbf{y}))$ is a vector of proportions $\mathbf{p} \in S^D$. The vector $\mathbf{p}$ takes the same value for all the members in an equivalence class. On the other hand, the term $(\sum x_j)$ determines the particular composition $\mathbf{x}$ recovered using information based on its original units.

It is generally agreed upon that a statistical analysis of CoDa has to be performed on coordinates with respect to a log-ratio basis (Mateu-Figueras et al. 2011). In particular, the Aitchison distance d_a between two compositions $\mathbf{x}_1$

and $\mathbf{x}_2$ can be calculated as the Euclidean distance d_e between their corresponding vectors of olr-coordinates: $d_a(\mathbf{x}_1, \mathbf{x}_2) = d_e(\text{olr}(\mathbf{x}_1), \text{olr}(\mathbf{x}_2))$. Analogous definitions can be provided for the norm and scalar product, and for the log-ratio normal probability distribution (Mateu-Figueras et al. 2013). These basic elements are the basis of most statistical methods. Commonly, researchers apply statistical methods such as, among others, linear regression, time series, or cokriging, to get predictions or estimates. When the *response* variable is a composition, the statistical method provides the estimates expressed in log-ratio coordinates. Frequently, the researcher requires back transforming these estimates to express them in the original units such as euros, kg, mg/liter or percentages. In the latter case, the modeler is dealing with non-closed subcompositions where the values are expressed in percentages or proportions. However, in all these cases, it is not possible to apply Eq. 2 to the estimates because in this case the term based on the original units is unknown. In consequence, other different strategies must be explored. This communication explores the advantages and disadvantages of two solutions to the units recovery problem.

The work is organised as follows. In Section 2, two different approaches for recovering original units are formally introduced. In Section 3, we apply the approaches when the goal is the estimate of the expected value of a random composition. We illustrate the procedures using a geochemical data set. Section 4 introduces how to recover the original units when using spatial models such as cokriging. Lastly, Section 5 concludes with some final remarks.

All data analyses discussed in this work were done using the R statistical programming environment (R Core-Team 2019).

TWO DIFFERENT APPROACHES FOR RECOVERING ORIGINAL UNITS

Consider n realizations $\mathbf{x}_i, i = 1, 2, \ldots, n$ of a D-part random composition. That is, consider a set of D-part compositions

$$\mathbf{X} = \{x_{ij} : i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, D\},$$

expressed in original units. Importantly, we assume that the units of compositions in $\mathbf{X}$ are homogeneous. For instance, all values are percentages, ppm, mol/L, or μg/m^3.

Available methods for estimation will lead to results in closed form, that is, summing to unity, percent, or the like. Here we explore two different approaches when the data are in some units that do not sum to a constant, like μg/L, or when one is dealing with a non-closed subcomposition.

First approach: adding a residual part

A pragmatic approach consists in imposing a total T to the composition and performing the estimation using an auxiliary part **Res**, namely the residual, where $Res_i = \mathrm{T} - \sum_{j=1}^{D} x_{ij}$. The idea of considering the part **Res** in a compositional analysis was used for the first time for different purposes in Buccianti et al. (2014) and Buccianti (2015). Note that Res_i allows the recovery of the original units because it includes the term $\sum_{j=1}^{D} x_{ij}$ with the information related to the original units (Eq. 2).

Figure 1 shows the complete procedure for recovering the original units of an estimate. The residual **Res** and the total T play a crucial role. Given a set $\mathbf{X}$ of D-part compositions, the procedure consists of the following steps:

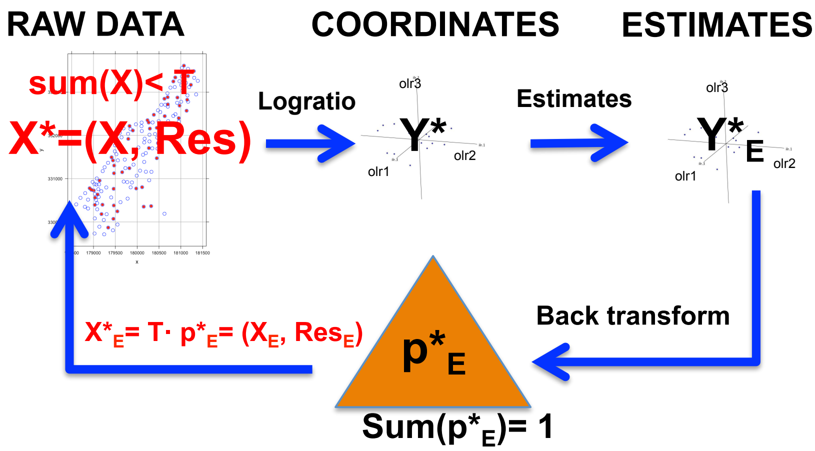

Figure 1: Procedure diagram for recovering original units of a set $\mathbf{X}$ of D-part compositions when adding a residual part **Res**, where $\mathbf{X}^*, \mathbf{p}_E^*, \mathbf{X}_E^* \in S^{D+1}$ and $\mathbf{Y}^*, \mathbf{Y}_E^* \in R^D$. Background pictures represent the corresponding spaces for $D = 3$.

1. Select a total T, $\mathrm{T} > \max_i\{\sum_{j=1}^{D} x_{ij}\}$, where $i = 1, 2, \dots, n$. For each sample compute the residual $Res_i = \mathrm{T} - \sum_{j=1}^{D} x_{ij}, i = 1, 2, \dots, n$. Consider the extended data set adding the residual part. That is, for $i = 1, 2, \dots, n$, consider the $(D+1)$-composition $\mathbf{x}_i^* = (x_{i1}, x_{i2}, \dots, x_{iD}, Res_i)$, where $\sum_{j=1}^{D+1} x_{ij} = \mathrm{T}$.
2. Express the extended compositions using log-ratio coordinates. That is, for $i = 1, 2, \dots, n$, consider the D-vector

$$\mathbf{y}_i^* = (\mathrm{olr}_1(\mathbf{x}_i^*), \mathrm{olr}_2(\mathbf{x}_i^*), \dots, \mathrm{olr}_D(\mathbf{x}_i^*)).$$

3. Apply the statistical method to obtain the corresponding estimate $\mathbf{y_E}^*$.
4. Back transform the estimate to obtain the corresponding vector of proportions $\mathbf{p_E}^* = (p_{E1}^*, p_{E2}^*, \dots, p_{ED}^*, p_{ED+1}^*)$. The part p_{ED+1}^* is the proportion estimated for the residual.

5. Multiply the vector of proportions $\mathbf{p_E}^*$ by T to recover the original total: $\mathbf{x_E}^* = \mathrm{T} \cdot \mathbf{p_E}^*$. The parts $(x_{E1}, x_{E2}, \ldots, x_{ED})$ are the estimated composition expressed in original units.

Importantly, when the statistical method allows to work on the simplex, steps 2 to 4 can be replaced by: *apply a simplicial method using raw data to obtain the estimated vector of proportions* $\mathbf{p_E}^*$. The procedure above describes the process for only one estimate but it can be easily extended by repetition to a number of estimates, for example, when obtaining the estimates in a linear regression model where the composition is the response variable then one has an estimate for each observation (row) of the data set.

Second approach: using an auxiliary real variable

A different approach results when using an auxiliary real variable related to the original units of the composition. Typical examples of these real variables are, among others, variables based on the sum of the composition ($\sum_{j=1}^{D} x_j$), the sum of any subcomposition of the composition (i.e., $x_3 + x_6$), or the geometric mean ($(\prod_{j=1}^{D} x_j)^{1/D}$). In general, let t be an auxiliary real variable based on the original units, that is, one can assume that there exists a function f where $t = f(\mathbf{x})$.

The complete procedure for recovering the original units using an auxiliary variable is shown in Fig. 2. The relation $t = f(\mathbf{x})$ allows to recover the original units. Given a set $\mathbf{X}$ of D-part compositions, the procedure consists of the following steps:

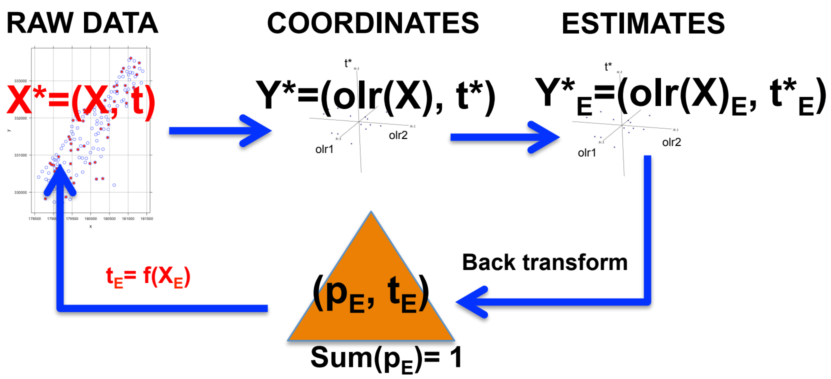

Figure 2: Procedure diagram for recovering original units of a set $\mathbf{X}$ of D-part compositions when using an auxiliary real variable t, where $\mathbf{p}_E \in S^D$ and $\mathbf{Y}^*, \mathbf{Y}_E^* \in R^D$. Background pictures represent the corresponding spaces for $D = 3$.

1. For $i = 1, 2, \ldots, n$ compute $t_i = f(\mathbf{x}_i)$ and create the vector $\mathbf{x}_i^* = (x_{i1}, x_{i2}, \ldots, x_{iD}, t_i)$.
2. Express the extended vector using appropriate coordinates $\mathbf{y}_i^*$, for $i = 1, 2, \ldots, n$. The composition $\mathbf{x}_i$ is expressed using olr coordinates. The coordinates of the auxiliary variable are the most appropriate for its sample space. For example, when $t_i = \sum_{j=1}^{D} x_{ij}$ a simple logarithm is used $t_i^* = \ln(\sum_{j=1}^{D} x_{ij})$.
3. Apply the statistical method to obtain the corresponding estimate $\mathbf{y_E}^*$.
4. Back transform the estimate to obtain the corresponding vector of proportions $\mathbf{p_E}$ and the value of the auxiliary variable t_E. The value t_E informs about the original units of the composition.
5. Finally, use the function $t = f(\mathbf{x})$ to recover the original units vector $(\mathbf{x_E}_1, \mathbf{x_E}_2, \ldots, \mathbf{x_E}_D)$ for the composition estimated. For example, when $t = x_3 + x_6$, using $\mathbf{p_E}$, t_E, and the relation $p_3 + p_6 = (x_3 + x_6)/(\sum x_j)$, one calculates the estimate of the total $\sum x_{Ej}$ to obtain the composition in original units $\mathbf{x_E} = (\sum x_{Ej}) \cdot \mathbf{p_E}$.

Pawlowsky-Glahn et al. (2015b) present a practical example of this procedure. Importantly, steps 2 to 4 can be removed from the procedure when the statistical method is able to provide the estimates $\mathbf{p_E}$ working on raw data, that is, when it is not required to work on log-ratio coordinates. One example of this situation is the center of a CoDa set.

ESTIMATING THE EXPECTED VALUE OF A RANDOM COMPOSITION: THE CENTER

The simplest case where one is dealing with estimates is when the expected value of a random composition is analyzed. In this analysis, the expected value is estimated by calculating the center of the data set available (Pawlowsky-Glahn et al. 2015c). Let $\mathbf{X} = \{x_{ij} : i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, D\}$ be a data set, the sample center is defined as

$$\mathbf{g} = cen\,(\mathbf{X}) = C\left(\left(\prod_{i=1}^{n} x_{i1}\right)^{1/n}, \left(\prod_{i=1}^{n} x_{i2}\right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD}\right)^{1/n}\right),$$

the closure of the columnwise geometric mean. Remarkably, the sample center $\mathbf{g}$ can be obtained by back transforming the columnwise arithmetic mean of the log-ratio coordinates (Pawlowsky-Glahn et al. 2015c). The question that automatically arises is the advantages and disadvantages of the two approaches introduced above when expressing the estimate of the expected value in original units.

How to obtain the center by adding a residual part?

According the definition of center given above, it is not necessary to work on coordinates to obtain the estimate of the expected value of a random composition. In consequence, the procedure consists of the following steps:

1. Select a total T, $\mathrm{T} > \max_i\{\sum_{j=1}^{D} x_{ij}\}$. Compute the residual $Res_i = \mathrm{T} - \sum_{j=1}^{D} x_{ij}$, $i = 1, 2, \ldots, n$ and create the composition $(x_{i1}, x_{i2}, \ldots, x_{iD}, Res_i)$.
2. Compute the non-closed geometric mean of each part of the extended data set, that is, the geometric mean columnwise:

$$\begin{aligned}\mathbf{G}^* &= (Gx_1, Gx_2, \ldots, Gx_D, Gr)\\ &= \left(\left(\prod_{i=1}^{n} x_{i1}\right)^{1/n}, \left(\prod_{i=1}^{n} x_{i2}\right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD}\right)^{1/n}, \left(\prod_{i=1}^{n} Res_i\right)^{1/n}\right).\end{aligned}$$

3. Apply the closure operation to $\mathbf{G}^*$:

$$\mathbf{g}^* = \left(\frac{Gx_1}{SG}, \frac{Gx_2}{SG}, \ldots, \frac{Gx_D}{SG}, \frac{Gr}{SG}\right) \cdot \mathrm{T},$$

where $SG = \left(\sum_{j=1}^{D} Gx_j\right) + Gr$, T is the closure constant, and, in this case, $\mathbf{p_E}^* = \mathbf{g}^*/\mathrm{T}$.
4. Consider only the parts corresponding to the original composition to make the estimate of the center of $\mathbf{x}$ on the original units:

$$\mathbf{x_E} = cen\,(\mathbf{X}) = \left(\frac{Gx_1}{SG}, \frac{Gx_2}{SG}, \ldots, \frac{Gx_D}{SG}\right) \cdot \mathrm{T}.$$

Steps 4 and 5 can be replaced by: first, consider the log-ratio coordinates of the samples (including the residual) and compute the arithmetic mean of these coordinates; second, back transform the vector of arithmetic means and apply the closure operation with the closure constant T.

Four important remarks follow:

I. Consider two different totals T_1 and T_2. Following the procedure above the two estimates of the center for a part x_k in $\mathbf{x}$ using each of the two totals are, respectively,

$$\frac{Gx_k}{SG_1} \cdot \mathrm{T}_1 \quad \text{and} \quad \frac{Gx_k}{SG_2} \cdot \mathrm{T}_2.$$

Note that the factors $\frac{\mathrm{T}_1}{SG_1}$ and $\frac{\mathrm{T}_2}{SG_2}$ are different (see Appendix A for a detailed proof). Consequently, the two expressions in the original units of the estimate are different, and the procedure is not invariant under a change of the total.

II. The factor $\frac{\mathrm{T}}{SG}$ tends to 1 when the total T tends towards infinity (Appendix A). In that case, the estimate of the center $\mathbf{g}^*$ approaches the non-closed geometric mean $\mathbf{G}^*$.

III. Let $\mathbf{x}_N$ be the random composition obtained by adding a new part to $\mathbf{x}$. That is, $\mathbf{x}_N = (x_1, x_2, \ldots, x_D, x_{D+1})$ and $\mathbf{x} = (x_1, x_2, \ldots, x_D)$. Consider a total T, $\mathrm{T} > \max_i\{\sum_{j=1}^{D+1} x_{ij}\}$. Let Res be the residual part for $\mathbf{x}$, and Res_N the residual part for $\mathbf{x}_N$; thus $Res_N = Res - x_{D+1}$. Let x_k be a single common part in $\mathbf{x}_N$ and $\mathbf{x}$, that is, $k = 1, 2, \ldots, D$. Following the procedure above the two estimates of the center for the part x_k in $\mathbf{x}$ and $\mathbf{x}_N$ are, respectively,

$$\frac{Gx_k}{SG} \cdot \mathrm{T} \quad \text{and} \quad \frac{Gx_k}{\left(\sum_{j=1}^{D+1} Gx_j\right) + Gr^*} \cdot \mathrm{T}\,.$$

Note that the factors $\frac{1}{SG}$ and $\frac{1}{(\sum_{j=1}^{D+1} Gx_j)+Gr^*}$ are different because the geometric mean is a non-linear operator, that is, despite $Res_N = Res - x_{D+1}$, its geometric mean Gr^* is not equal to the subtraction of geometric means $Gr - Gx_{D+1}$. Consequently, the two expressions in the original units of the estimate are different, and the procedure is not invariant under the change of the subcomposition where the part is included.

IV. Although the procedure gives expressions in the original units of the center which are not invariant under change of total or change of subcomposition, all of them are in the same equivalence class. Thus, the common subcomposition $\mathbf{x_E}$–the composition $\mathbf{x_E}^*$ without the residual part–expressed in any log-ratio coordinates (i.e., alr, clr and any olr) are invariant and equal to the log-ratio coordinates of the corresponding subcomposition of $\mathbf{G}^*$. That is, the closed subcomposition is exactly the same regardless of the total selected.

Estimating the expected value using the multiplicative total

A different approach results when using the concept of *multiplicative total* of a composition. Let $\mathbf{x}$ be a D-part composition and $m = (\prod_{j=1}^{D} x_j)^{1/D}$ its geometric mean. The value m^D informs about the *multiplicative total* of the composition. Importantly, the additive total of a vector is equal to its arithmetic mean multiplied by D, the number of parts. That is, the arithmetic mean gives information about the additive total. Analogously, the geometric mean gives information about its multiplicative total. Note that the sum of the vector $\mathbf{x}/(\sum_{j=1}^{D} x_j)$ equals to one, while the vector $\mathbf{x}/m$ has a multiplicative total equal to one. Therefore, given two arbitrary positive values T and M, the vectors $(\mathbf{x}/\sum_{j=1}^{D} x_j) \cdot \mathrm{T}$ and $(\mathbf{x}/m) \cdot \mathrm{M}$ have additive total T and geometric mean M, respectively.

The procedure to estimate the center in original units consists of the following steps:

1. Compute the geometric mean of each part of the data set:

$$\mathbf{G} = \left(\left(\prod_{i=1}^{n} x_{i1} \right)^{1/n}, \left(\prod_{i=1}^{n} x_{i2} \right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD} \right)^{1/n} \right).$$

Consider the notation $\mathbf{G} = (Gx_1, Gx_2, \ldots, Gx_D)$.

2. Apply the closure operation to $\mathbf{G}$ to obtain a closed geometric mean or center of the data set:
$$\mathbf{g} = \left(\frac{Gx_1}{\sum_{j=1}^{D} Gx_j}, \frac{Gx_2}{\sum_{j=1}^{D} Gx_j}, \ldots, \frac{Gx_D}{\sum_{j=1}^{D} Gx_j} \right) .$$
Here $\mathbf{p_E} = \mathbf{g}$. That is, it plays the role of the estimate.
3. For each sample compute the row-wise geometric mean:
$$m_i = \left(\prod_{j=1}^{D} x_{ij} \right)^{1/D} , i = 1, 2, \ldots, n .$$
Here $t_i = m_i$. That is, it plays the role of the auxiliary variable.
4. Compute the geometric mean of these geometric means:
$$\mathrm{M} = \left(\prod_{i=1}^{n} m_i \right)^{1/n} .$$
This value informs about the average of the row-wise geometric mean in the data set, and it plays the role of the estimate $t_E = \mathrm{M}$.
5. Let m_g be the geometric mean of the closed geometric mean $\mathbf{g}$. Scale accordingly the closed geometric mean $\mathbf{g}$ to obtain the estimate of the center in the original units:
$$cen\left(\mathbf{X}\right) = \left(\frac{Gx_1}{\sum_{j=1}^{D} Gx_j}, \frac{Gx_2}{\sum_{j=1}^{D} Gx_j}, \ldots, \frac{Gx_D}{\sum_{j=1}^{D} Gx_j} \right) \cdot \frac{\mathrm{M}}{m_g}.$$
Note that the geometric mean of $cen\left(\mathbf{X}\right)$ is equal to M.

Three important remarks follow:

I. Note that
$$\begin{aligned} cen\left(\mathbf{X}\right) &= \left(Gx_1, Gx_2, \ldots, Gx_D\right) \cdot \frac{\mathrm{M}}{m_g \cdot \left(\sum_{j=1}^{D} Gx_j\right)} \\ &= \left(Gx_1, Gx_2, \ldots, Gx_D\right) \cdot \frac{\mathrm{M}}{\left(\prod_{j=1}^{D} \frac{Gx_j}{\sum_{k=1}^{D} Gx_k}\right)^{1/D} \cdot \left(\sum_{j=1}^{D} Gx_j\right)} \\ &= \left(Gx_1, Gx_2, \ldots, Gx_D\right) \cdot \frac{\mathrm{M}}{\left(\prod_{j=1}^{D} Gx_j\right)^{1/D}} \\ &= \left(Gx_1, Gx_2, \ldots, Gx_D\right) \cdot \frac{\left(\prod_{i=1}^{n} \left(\prod_{j=1}^{D} x_{ij}\right)^{1/D}\right)^{1/n}}{\left(\prod_{j=1}^{D} Gx_j\right)^{1/D}} \\ &= \left(Gx_1, Gx_2, \ldots, Gx_D\right) . \end{aligned}$$
That is, the estimate of the center expressed in original units is equal to the non-closed geometric mean vector.

II. Pawlowsky-Glahn et al. (2015a) studied the properties of the T-space defined by a composition and a *total*. They state that D-1 olr coordinates $\mathbf{y}_i$ together with the coordinates of a multiplicative total m_i^D lead to the same distances among individuals as in the space of the logarithms of absolute values. Following Coenders et al. (2017), the coordinates of the value m_i is associated to the projection of vector $\ln(\mathbf{x}_i)$ to the unit normalized vector $(1/\sqrt{D})\mathbf{1}_D$, where $\mathbf{1}_D$ is the D-vector $(1,1,\ldots,1)$. This vector is orthogonal to the space of log-ratio coordinates, forming an orthonormal basis of the complete real space R^D. Let $\mathbf{U}$ be a $D\times D$ matrix where the first D-1 columns are the vectors of an olr basis and the last column is the vector $(1/\sqrt{D})\mathbf{1}_D$. It holds

$$(\mathbf{y}_i\, m_i^*) = \ln(\mathbf{x}_i)\cdot\mathbf{U} \quad \text{and} \quad \ln(\mathbf{x}_i) = (\mathbf{y}_i\, m_i^*)\cdot\mathbf{U}^{-1}, \tag{3}$$

where $\mathbf{U}^T\cdot\mathbf{U} = \mathbf{I}_D$ and $\mathbf{U}^T = \mathbf{U}^{-1}$, being $\mathbf{I}_D$ the $D\times D$ identity matrix, and m_i^* the coordinates of m_i ($m_i^* = \sqrt{D}\ln(m_i)$). That is, the $\mathbf{U}$ matrix is an orthonormal change of basis from $\ln(\mathbf{x}_i)$ into the R^D space (Coenders et al. 2017).

According this approach, the procedure above to estimate the center in original units is equivalent to: first, olr-transform the compositional data set; next, extend the data set of log-ratio coordinates by adding the log-score of the geometric mean $\sqrt{D}\cdot\ln(m_i), i = 1,2,\ldots,n$, to each vector of olr-coordinates; compute the arithmetic mean of the extended data set; back transform the vector of arithmetic means. Where the back transformation simply consists of a change of basis and the exponential function (Eq. 3):

$$\mathbf{x_E} = \exp((\mathbf{y_E}\, m_E^*)\cdot\mathbf{U}^{-1}).$$

III. The procedure provides an estimate of the center that is invariant under change of subcompositions, where the part is included.

Example: center of a compositional data set

The Meuse data set (Rikken and Rijn 1993) is included in the "gstat" R-package (Graler et al. 2016). The data set gives locations (in meters) and topsoil heavy metal concentrations (in ppm), along with a number of soil and landscape variables at $n = 155$ observation locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations of (Cd, Cu, Pb, Zn) have been measured in composite samples of an area of approximately 15 m $\times$ 15 m. Table 1 shows the values in original units (ppm) of the estimates for the center of the Meuse data set. Because the maximum value of the sum for the samples in the 4-part subcomposition (Cd, Cu, Pb, Zn) is 2630, the sequence of 10 different totals considered is

$$\mathrm{T} = \{2650, 3150, 3650, 4150, 4650, 5150, 5650, 10^4, 10^5, 10^6\}\,.$$

For this sequence the procedure consisting of adding a residual part is applied to the 3-part subcomposition (Cu, Pb, Zn) as well as to the 4-part subcomposition (Cd, Cu, Pb, Zn). Also, the procedure based on the multiplicative total is applied to this subcomposition. Table 1 shows that, as expected, the expression in original units of the 3-part subcomposition (Cu, Pb, Zn) is different when the total changes, and also when one uses the 4-part subcomposition. In addition, when the residual part increases, the other four parts diminish, approaching the results obtained using the multiplicative total.

Table 1: Values in original units (ppm) of the estimates for the center of Meuse data set. (3-sub = 3-part subcomposition; 4-sub = 4-part subcomposition)

Total	Subcomp.	Cd	Cu	Pb	Zn	$Resid$
2650	3-sub		38.90	135.81	399.40	2075.89
	4-sub	1.95	39.02	136.23	400.64	2072.16
3150	3-sub		37.61	131.31	386.18	2594.90
	4-sub	1.88	37.65	131.44	386.54	2592.49
3650	3-sub		37.06	129.40	380.55	3102.99
	4-sub	1.85	37.09	129.49	380.81	3100.77
4150	3-sub		36.72	128.20	377.01	3608.07
	4-sub	1.84	36.74	128.27	377.22	3605.94
4650	3-sub		36.48	127.36	374.55	4111.62
	4-sub	1.82	36.49	127.42	374.72	4109.55
5150	3-sub		36.30	126.73	372.72	4614.25
	4-sub	1.82	36.31	126.78	372.86	4612.22
5650	3-sub		36.16	126.25	371.30	5116.29
	4-sub	1.81	36.17	126.30	371.43	5114.29
10^4	3-sub		35.62	124.36	365.73	9474.29
	4-sub	1.78	35.62	124.38	365.79	9472.42
10^5	3-sub		35.10	122.55	360.41	99481.94
	4-sub	1.76	35.10	122.55	360.42	99480.17
10^6	3-sub		35.05	122.39	359.93	999482.62
	4-sub	1.75	35.05	122.39	359.94	999480.87
Mult. total	4-sub	1.75	35.05	122.37	359.88	

Residual versus multiplicative total

The results shown in the previous sections suggest an analysis of the relation between the residual part and the multiplicative total of a composition. Let $\mathbf{x}$ be a D-part composition and $m^D = \prod_{j=1}^{D} x_j$ its multiplicative total. Let T be a fixed total and $Res = \mathrm{T} - \sum_{j=1}^{D} x_j$ the corresponding residual part. Once one has defined a particular orthonormal basis to get the olr coordinates of $\mathbf{x}$, one can assume that the variable added by the residual part in the first procedure is the last olr variable $\sqrt{\frac{D}{D+1}} \cdot \ln \frac{Res}{m}$, whereas in the second procedure the variable added by the multiplicative total is $\sqrt{D} \cdot \ln m$. At this point, one can assume that T is fixed but as large as we need, that is, the residual Res is as large as we need. In consequence, for each sample, it holds that (Appendix B):

$$\sqrt{\frac{D}{D+1}} \cdot \ln \frac{Res}{m} \approx \sqrt{\frac{D}{D+1}} \cdot \ln \mathrm{T} - \sqrt{\frac{D}{D+1}} \cdot \ln m.$$

Let $a = \sqrt{\frac{D}{D+1}} \cdot \ln \mathrm{T}$ and $b = -\sqrt{\frac{1}{D+1}}$ be two constants. For a large T it holds that the log-ratio coordinate of the residual part is approximately a linear transformation of the log-score of the multiplicative total: $a + b \cdot \sqrt{D} \cdot \ln m$. This linear relationship suggests that, when the total is "large", the results provided by any equivariant method applied to the set of log-ratio coordinates including a residual part will be related to the results obtained using the multiplicative total. Note that the results shown in Table 1 confirm this idea. Because the estimation of the mean using log-ratio coordinates is an equivariant method, the center obtained for a large T approaches the center using the multiplicative total.

SPATIAL ANALYSIS

Method for regionalised compositions

The general case of spatial interpolation can be summarized by the expression

$$\mathbf{Y}_E(s) = \sum_{i=1}^{n} \lambda(s_i) \cdot \mathbf{Y}(s_i), \quad \text{with} \sum_{i=1}^{n} \lambda(s_i) = 1, \tag{4}$$

where $\mathbf{Y}_E(s)$ is a vector of estimates, $\lambda(s_i)$ is a scalar "weight", and $\mathbf{Y}(s_i)$ is a vector of observations at location s_i in a spatial domain $\mathcal{D}$, $i = 1, 2, \ldots, n$. Observe that the estimate $\mathbf{Y}_E(s)$ is a weighted arithmetic mean. For the CoDa case, consider a D-part regionalised composition $\mathbf{X}(s_i)$ observed at locations s_i in the spatial domain, $i = 1, 2, \ldots, n$ and the compositional geostatistics workflow summarized as follows (Tolosana-Delgado et al. 2019):

1. Express the compositions $\mathbf{X}(s_i)$, $i = 1, 2, \ldots, n$ in one of the log-ratio scores $\mathbf{Y}(s_i)$.
2. Compute variation-variograms of $\mathbf{Y}$.
3. Fit a valid model, such as the linear model of coregionalization.
4. Apply cokriging to the log-ratio scores at the nodes of a suitable chosen grid.
5. Back transform the predicted values.

These available methods for estimation/interpolation of regionalised compositions like $\mathbf{X}$ will lead to results in closed form, that is, summing to unity, percent, or the like, which can be a problem or not desired result when the data are in some units, like μg/liter, or data are a non-closed subcomposition expressed in proportions, percentages or ppm. That is, the compositions $\mathbf{X}(s_i)$, $i = 1, 2, \ldots, n$, in their original units do not sum to a constant. To solve this problem one can use one of the two approaches proposed: add a residual part or use an auxiliary variable. Importantly, in both approaches the log-ratio scores vector $\mathbf{Y}(s_i)$ will be replaced by the extended vector of coordinates $\mathbf{Y}^*(s_i)$ to compute the variograms and fit a model. In order to avoid an influence of the variogram modeling in our study about the two approaches, we replace these points (steps 2 and 3) by an interpolation method based on the geographical distance. That is, we consider the scalar weight $\lambda(s_i)$ is proportional to the inverse of the geographical distance between locations s and s_i, $i = 1, 2, \ldots, n$. In consequence, in both approaches the expression to calculate the estimates is

$$\mathbf{Y}_E^*(s) = \sum_{i=1}^{n} \lambda(s_i) \cdot \mathbf{Y}^*(s_i), \quad \text{where} \quad \lambda(s_i) = \frac{\frac{1}{d_e(s,s_i)}}{\sum_{k=1}^{n} \frac{1}{d_e(s,s_k)}}, \tag{5}$$

and $\mathbf{Y}^*(s_i)$ is respectively obtained by adding the residual part or using the auxiliary variable selected. The estimated value in original units $\mathbf{X}_E(s)$ is respectively obtained following the schemes in Fig. 1 and Fig. 2.

Example: maps using original units

Table 2 shows the main quantiles and the geometric mean of the Meuse data set. In this table, the geometric mean is the non-closed geometric mean vector (Table 1).

Table 2: Basic statistics of Meuse data set (in ppm)

	Cd	Cu	Pb	Zn
Minimum	0.2	14.0	37.0	113.0
Q1	0.8	23.0	72.5	198.0
Median	2.1	31.0	123.0	326.0
Geomean	1.8	35.1	122.4	359.9
Q3	3.9	49.5	207.0	674.5
Maximum	18.1	128.0	654.0	1839.0

The values of the statistics suggest, for the four parts, a right skewed distribution, which is common for geochemical elements. Note that all the values in Table 2 are expressed in ppm but the ranges of the parts are very different, with Cd being the part taking the smallest values and Zn the largest values. In particular, the minimum of Zn is approximately 500 hundred times the minimum of Cd. For this 4-part composition (Cd, Cu, Pb, Zn) we will show the results for the element Cu using a kriging method based on the inverse of the geographical distance (in meters) between locations with both approaches proposed and comparing the results for the 3-part subcomposition (Cu, Pb, Zn). Analogous results were obtained when the part analyzed was Cd, Pb, and Zn.

Figure 3 shows the maps with the concentration estimated for element Cu using the 4-part composition. Following the same procedure as in previous section, the approach consisting of adding a residual part was applied for the sequence of totals $\mathrm{T} = \{2650, 3150, 3650, 4150, 4650, 5150, 5650, 10^4, 10^5, 10^6\}$. Figures 3(a-c) show the results for $\mathrm{T} = 2650, 4650, 10^6$. The maps state differences between the estimates, where the values decrease when the total T increases. No relevant differences are detected when the map for $\mathrm{T} = 10^6$ (Fig. 3c) is compared to the map using the multiplicative total (Fig. 3d). This similarity agrees with the performance analyzed for the estimation of the centre of a random composition.

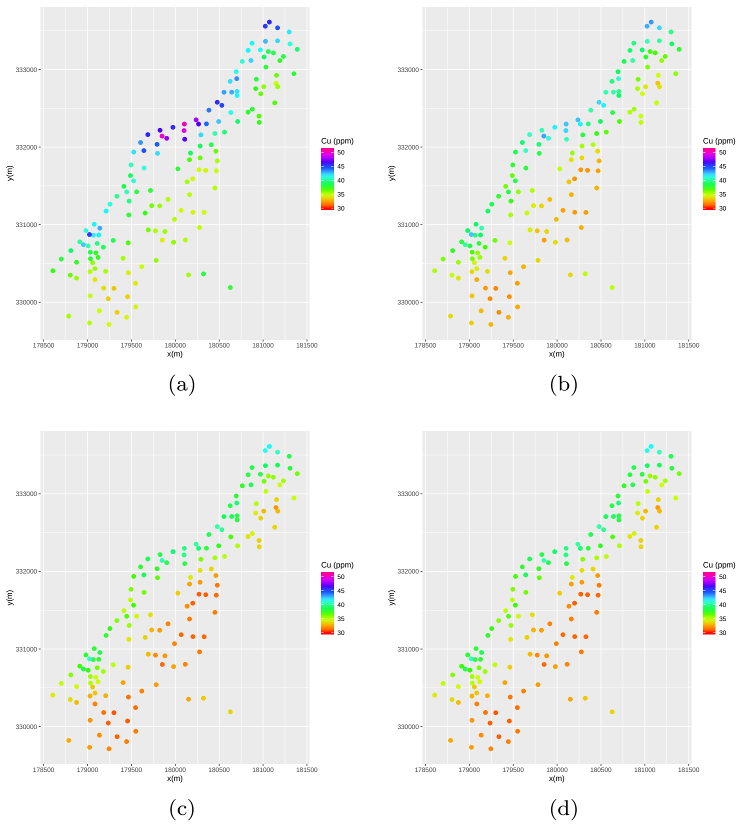

Figure 3: Map of Meuse data set: concentration estimated in element Cu for the 4-part composition (Cd, Cu, Pb, Zn). The approach used is: (a) residual part with T= 2650; (b) residual part with T= 4650; (c) residual part with T= 10^6; (d) multiplicative total. The data set gives locations (in meters) and topsoil heavy metal concentrations (in ppm).

To analyze the error of the estimates in original units one can consider the absolute error $|Cu_{obs} - Cu_{est}|$, where $|Cu_*|$ are respectively the values observed and estimated for the element Cu in one location. When one wants to calculate the accumulated error for all the data, it is preferable to consider the relative error $|\frac{Cu_{obs}-Cu_{est}}{Cu_{obs}}|$. Moreover, when the estimated value approaches the observed, the relative error approaches the logratio $|\ln\frac{Cu_{est}}{Cu_{obs}}|$. This term can be interpreted as a measure of the contribution of element Cu to the perturbation difference vector between the observed and estimated composition (Martín-Fernández et al. 2015, 2019). Table 3 shows the log-ratio accumulated error. This error is calculated using the log-ratio expression for the cases of the 3-part and 4-part composition for some selected totals T and the multiplicative total. The accumulated error is very large suggesting that the cokriging method applied provides poor results in this case. The difference between errors is very small when the results for the 3-part and 4-part are compared. Moreover, in both cases, the error diminishes when the total T increases and tends towards the error provided by the multiplicative total approach. This behavior is shown in Fig. 4. The difference $Cu_4 - Cu_3$ (in ppm) between estimates for part Cu using the 4-part and the 3-part composition decreases and approaches to zero when total T tends towards 10^6. The values of the differences are positive indicating that the estimates provided by the 4-part composition are greater than the estimates for the 3-part composition. This effect is related to the third remark in previous section for the estimation of the center of a data set where the subcompositional coherence for the residual approach is explored and it is also stated in Table 1.

Table 3: Log-ratio accumulated error for estimates of Cu using different totals T and the multiplicative total with the 3-part or the 4-part composition. (3-sub = 3-part subcomposition; 4-sub = 4-part subcomposition.)

	Total T					
Composition	2650	3150	4650	10^5	10^6	Multiplic.
3-sub	63.37	62.27	61.37	60.37	60.33	60.33
4-sub	63.49	62.30	61.38	60.37	60.33	60.33

CONCLUSIONS AND FINAL REMARKS

When the purpose is to recover the original units in a compositional analysis it is necessary to add more information to the relative information provided by the olr coordinates of the original D-part composition. Two different approaches for units recovery in compositional analysis have been explored: adding a residual part and using an auxiliary variable. The approach to add a residual part is the simplest technique and it can be considered the most intuitive approach for CoDa originally expressed in proportions, percentages or ppm. However, we have found that adding a residual part presents undesirable properties that a modeler should be take into account. In particular, the estimates in original units obtained using the residual approach:

– depend on the total T considered;

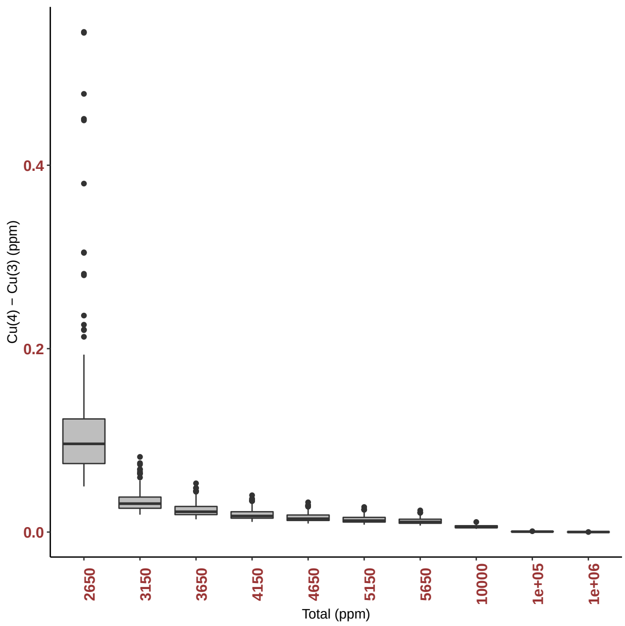

Figure 4: Box plots for the difference in ppm for the estimates of part Cu using 4-part and 3-part compositions when the residual approach is applied for 10 totals, from T=2650 to T=10^6.

– depend on the number of parts forming the composition; and
– approach the estimates obtained using the multiplicative total when the total T tends towards infinity.

On the other side, exploring the approach using an auxiliary variable, we found that the most sensible option is the information provided by the geometric mean of the composition, that is, the variable called multiplicative total. We have stated that the set formed by the olr coordinates of the composition and the log-score of the multiplicative total are appropriate to obtain estimates in original units, being invariant regardless the number of parts forming the composition. This approach, being equivalent to work with the log-transformed data, has the advantage of providing knowledge about the relative (olr coordinates) and the absolute information (multiplicative total), information which remains hidden otherwise. However, when only one part has been measured (D = 1) this approach is inapplicable because there are no multiple proportions to generate auxiliary variables.

Importantly, both approaches provide the same estimates expressed in olr coordinates. In other words, the relative information in the estimates is the same regardless the approach used, the total T considered and the number of parts forming the composition. Only the absolute information of the estimates depends on the approach used. In this sense we recommend to use the approach based on the multiplicative total because it splits the estimates into the olr coordinates of the composition (relative information) and the score of the geometric mean (absolute information).

ACKNOWLEDGEMENTS This research has been funded by the project "CODAMET" (Ministerio de Ciencia, Innovación y Universidades; Ref: RTI2018-095518-B-C21).

REFERENCES

Aitchison, J. (1986). *The statistical analysis of compositional data.* London (UK): Chapman & Hall, reprinted in 2003 by Blackburn Press.

Barceló-Vidal, C., & Martín-Fernández, J. A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45(4), 57–71.

Buccianti, A. (2015). The FOREGS repository: Modelling variability in stream water on a continental scale revising classical diagrams from coda (compositional data analysis) perspective. *Journal of Geochemical Exploration*, 154, 94–104.

Buccianti, A., Egozcue, J. J., & Pawlowsky-Glahn, V. (2014). Variation diagrams to statistically model the behavior of geochemical variables: Theory and applications. *Journal of Hydrology*, 519, 988–998.

Coenders, G., Martín-Fernández, J. A., & Ferrer-Rosell, B. (2017). When relative and absolute information matter. Compositional predictor with a total in generalized linear models. *Statistical Modelling*, 17(6), 494–512.

Edjabou, M. E., Martín-Fernández, J. A., Scheutz, C., & Astrup, T. F. (2017). Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. *Waste Management*, 69, 13–23.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *TEST*, 28(3), 599–638.

Graler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8(1), 204–218.

Jarauta-Bragulat, E., Hervada-Sala, C., & Egozcue, J. J., (2016). Air quality index revisited from a compositional point of view. *Mathematical Geosciences*, 48(5), 581–593.

Martín-Fernández, J. A. (2019). Comments on: Compositional data: the sample space and its structure. *TEST*, 28(3), 653–657.

Martín-Fernández, J. A., Daunis-Estadella, J., & Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *SORT*, 39, 231–252.

Martín-Fernández, J. A., Engle, M. A., Ruppert, L., & Olea, R. A. (2019). Advances in self-organizing maps for their application to compositional data. *SERRA*, 33, 817–826.

Mateu-Figueras, G., Pawlowsky-Glahn, V., & Egozcue, J. J. (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional data analysis: theory and applications* (pp. 31–42). Chichester (UK): John Wiley & Sons, Ltd.

Mateu-Figueras, G., Pawlowsky-Glahn, V., & Egozcue, J. J., (2013). The normal distribution in some constrained sample spaces. *SORT*, 37(1), 29–56.

Olea, R. A., Raju, N. J., Egozcue, J. J., Pawlowsky-Glahn, V., & Shubhra S (2018). Advancements in hydrochemistry mapping: application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. *SERRA*, 32(1), 241–259.

Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, 143, 85–96.

Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015a). Tools for compositional data with a total. *Statistical Modelling*, 15, 175–190.

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 S Afr Inst Min Metall* 115(1), 59–72.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R (2015c). *Modeling and analysis of compositional data*. Chichester (UK): John Wiley & Sons.

R Core-Team (2019). R: A language and environment for statistical computing. URL http://www.R-project.org. Accessed on 1 December 2019.

Rikken, M. G. J., & Rijn, R. P. G. V. (1993). Soil pollution with heavy metals - an inquiry into spatial variation, cost of mapping and the risk evaluation of copper, cadmium, lead and zinc in the floodplains of the Meuse west of Stein, the Netherlands. PhD thesis, Doctoraalveldwerkverslag, Dept. of Physical Geography, Utrecht University.

Tolosana-Delgado, R., Mueller, U., & van den Boogaart, K. G. (2019). Geostatistics for compositional data: An overview. *Math Geosci*, 51, 485–526.

APPENDIX A

Let $f(T)$ be the function

$$f(T) = \frac{T}{\sum_{j=1}^{D} Gx_j + Gr},$$

then it holds

- $f(\mathrm{T}) > 1$, for any total T. To prove this property one can use the well-known inequality between the geometric and arithmetic means

$$\sum_{j=1}^{D} Gx_j + Gr \leq \sum_{j=1}^{D} \left(\frac{1}{n} \sum_{i=1}^{n} x_{ij} \right) + \frac{1}{n} \sum_{i=1}^{n} Res_i,$$

where the equality holds only for a constant series, which is not the case in our context. Therefore,

$$\sum_{j=1}^{D} Gx_j + Gr < \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{D} x_{ij} + Res_i\right) = \mathrm{T}.$$

– $\lim_{\mathrm{T}->+\infty} f(\mathrm{T}) = 1$. For any $\mathrm{T} > 0$, the expression

$$f(\mathrm{T}) = \frac{\mathrm{T}}{\sum_{j=1}^{D} Gx_j + Gr} = \frac{\mathrm{T}}{\sum_{j=1}^{D} Gx_j + \left(\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)\right)^{1/n}},$$

is equal to

$$f(\mathrm{T}) = \frac{1}{\frac{\sum_{j=1}^{D} Gx_j}{\mathrm{T}} + \left(\prod_{i=1}^{n}\left(1 - \frac{\sum_{j=1}^{D} x_{ij}}{\mathrm{T}}\right)\right)^{1/n}},$$

where $\lim_{\mathrm{T}->+\infty} \frac{\sum_{j=1}^{D} Gx_j}{\mathrm{T}} = 0$ and $\lim_{\mathrm{T}->+\infty} \frac{\sum_{j=1}^{D} x_{ij}}{\mathrm{T}} = 0$.

– $f(\mathrm{T})$ is a monotonically decreasing function. To prove this behaviour one can prove that the function $g(\mathrm{T}) = 1/f(\mathrm{T})$ is a monotonically increasing function. The derivative function $g'(\mathrm{T})$ is equal to

$$g'(\mathrm{T}) = \frac{\frac{1}{n}\left(\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)\right)^{1/n}\left(\sum_{i=1}^{n}\frac{1}{\mathrm{T}-\sum_{j=1}^{D} x_{ij}}\right)\mathrm{T}}{\mathrm{T}^2} - \frac{\sum_{j=1}^{D} Gx_j + \left(\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)\right)^{1/n}}{\mathrm{T}^2},$$

where using the inequality between the geometric and arithmetic mean it holds

$$g'(\mathrm{T}) > \frac{\frac{1}{n}\left(\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)\right)^{1/n}\left(\sum_{i=1}^{n}\frac{1}{\mathrm{T}-\sum_{j=1}^{D} x_{ij}}\right)\mathrm{T} - \mathrm{T}}{\mathrm{T}^2} =$$

$$= \frac{\frac{1}{n}\left(\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)\right)^{1/n}\left(\sum_{i=1}^{n}\frac{1}{\mathrm{T}-\sum_{j=1}^{D} x_{ij}}\right) - 1}{\mathrm{T}}$$

Because the term $\prod_{i=1}^{n}\left(\mathrm{T} - \sum_{j=1}^{D} x_{ij}\right)^{1/n}$ is the geometric mean of the residuals and the term $\frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{\mathrm{T}-\sum_{j=1}^{D} x_{ij}}\right)$ is the inverse of the harmonic mean of the residuals, the sign of the numerator is positive. Therefore $g'(\mathrm{T}) > 0$.

APPENDIX B

Let T be a total fixed but as large as we need, like a “big T”. In consequence, for $i = 1, 2, \cdots, n$, the residual Res_i is as large as we need, that is $\mathrm{T} >> \sum_{j=1}^{D} x_{ij}$. For $i = 1, 2, \cdots, n$, it holds that

$$\sqrt{\frac{D}{D+1}} \cdot \ln \frac{Res_i}{m_i} = \sqrt{\frac{D}{D+1}} \cdot \ln Res_i - \sqrt{\frac{D}{D+1}} \cdot \ln m_i =$$

$$= \sqrt{\frac{D}{D+1}} \cdot \ln \left(\mathrm{T} - \sum_{j=1}^{D} x_j \right) - \sqrt{\frac{D}{D+1}} \cdot \ln m_i =$$

$$= \sqrt{\frac{D}{D+1}} \cdot \ln \left(\mathrm{T} \cdot \left(1 - \frac{\sum_{j=1}^{D} x_{ij}}{\mathrm{T}} \right) \right) - \sqrt{\frac{D}{D+1}} \cdot \ln m_i =$$

$$\sqrt{\frac{D}{D+1}} \cdot \ln \mathrm{T} + \sqrt{\frac{D}{D+1}} \cdot \ln \left(1 - \frac{\sum_{j=1}^{D} x_{ij}}{\mathrm{T}} \right) - \sqrt{\frac{D}{D+1}} \cdot \ln m_i.$$

In consequence, because $\mathrm{T} >> \sum_{j=1}^{D} x_{ij}$, it holds that

$$\sqrt{\frac{D}{D+1}} \cdot \ln \frac{Res_i}{m_i} \approx \sqrt{\frac{D}{D+1}} \cdot \ln \mathrm{T} - \sqrt{\frac{D}{D+1}} \cdot \ln m_i.$$