Model-based replacement of rounded zeros in compositional data: classical and robust approaches[☆]

J.A. Martín-Fernández[*,a], K. Hron[b], M. Templ[c,d], P. Filzmoser[c], J. Palarea-Albaladejo[e]

[a] *Department of Computer Science and Applied Mathematics, University of Girona, Campus Montilivi, P4, E-17071 Girona, Spain*

[b] *Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic*

[c] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*

[d] *Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria*

[e] *Biomathematics & Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, UK*

Abstract

The log-ratio methodology represents a powerful set of methods and techniques for statistical analysis of compositional data. These techniques may be used for the estimation of rounded zeros or values below the detection limit in cases when the underlying data are compositional in nature. An algorithm based on iterative log-ratio regressions is developed by combining a particular family of isometric log-ratio transformations with censored regression. In the context of classical regression methods, the equivalence of the method based on additive and isometric log-ratio transformations is proven. This equivalence does not hold for robust regression. Based on Monte Carlo methods, simulations are performed to assess the performance of classical and robust methods. To illustrate the method, a case study involving geochemical data is conducted.

Key words: balances, EM algorithm, log-ratio transformations, robust regression, values below detection limit

1. Introduction

Compositional data, or compositions, are vectors of positive values quantitatively describing the contribution of D parts of some whole, which carry exclusively relative information (Aitchison, 1986; Egozcue, 2009). The nature of these data explains why the variables or columns of the data matrix are also known as *parts* in the literature of compositional data analysis (hereafter CODA). This type of data is frequently collected in many applied fields (such as geochemistry, nutrition and behaviour sciences) and is represented as vectors with a constant sum constraint, such as proportions, percentages or ppb. According to Pearson (1897), there is general agreement that taking the usual Euclidean geometry approach to the statistical analysis of compositions may yield misleading results. The log-ratio (logarithm of a ratio) methodology proposed in Aitchison (1986) represents a powerful set of methods and techniques to apply to CODA. Since the 1990s, numerous publications have extended this approach in both theoretical and practical respects. Most of these new ideas and strategies to CODA were presented recently at the four CoDaWork meetings (Thió-Henestrosa and Martín-Fernández, 2003; Mateu-Figueras and Barceló-Vidal, 2005; Daunis-i-Estadella

[☆] Avalaible R file for the simulation study in the electronic version.

[*] Corresponding author. Dept. Computer Science and Applied Mathematics (UdG), Campus Montilivi (P4), E-17071 Girona, Spain. Tel.: +34 972 418426, Fax: +34 972 418792

Email addresses: josepantoni.martin@udg.edu (J.A. Martín-Fernández), hronk@seznam.cz (K. Hron), templ@statistik.tuwien.ac.at (M. Templ), p.filzmoser@tuwien.ac.at (P. Filzmoser), javier@bioss.ac.uk (J. Palarea-Albaladejo)

and Martín-Fernández, 2008; Egozcue et al., 2011) and collected in special publications (Buccianti et al., 2006; Pawlowsky-Glahn and Buccianti, 2011).

The main idea of the log-ratio approach is that compositional data are characterised by a different geometry in their sample space, the simplex. This structure, now known as the Aitchison geometry, confers to the simplex all the properties of a (D-1)-dimensional Euclidean space (Egozcue and Pawlowsky-Glahn, 2006). Because most standard statistical methods are designed for data in a real Euclidean space, compositions need to be expressed as vectors of values that belong to such a space. Two different but related ways make it possible to obtain these new vectors. The first of these ways is used when original compositions are expressed in coordinates with respect to an orthonormal basis in the Aitchison geometry. The second is used when a family of log-ratio transformations (ilr, alr, clr) is applied. From among the transformations in such a family, the isometric log-ratio (ilr) transformations (Egozcue et al., 2003) are preferred because of their advantageous theoretical and practical properties (Egozcue and Pawlowsky-Glahn, 2005; Filzmoser et al., 2009; Hron et al., 2010). In particular, (Egozcue et al., 2003) shows that the ilr-transformed values of a composition are equal to the coordinates of the composition from a particular orthonormal basis. In contrast, additive log-ratio (alr) transformations are related to oblique (not orthonormal) bases, and the centred log-ratio (clr) transformation provides the coefficients in a generating system.

The log-ratio methodology obviously cannot be used for CODA when the data contain zero values. The zeros can be present for various reasons, such as comprehensively described in, e.g., Martín-Fernández et al. (2011). If the parts (compositional variables) are continuous, such as concentrations of chemical elements, often a zero entry in the data matrix represents a consequence of rounding-off error rather than a pure absence of the part in the composition. Another typical situation is one in which detection limits exist and those values below the detection limits (VBDL) are automatically set to be zeros or simply labelled as *less-thans*. Often, each part has a different detection limit, and the detection limit can even change among observations, if the samples (row in data matrix) have been measured by different laboratories. This problem is known in the literature by the term *rounded zeros*, and several proposals for dealing with this type of data have been published in the last decade (e.g., Martín-Fernández et al., 2003; van den Boogaart et al., 2006; Palarea-Albaladejo et al., 2007). Note that when x_{ij} is a rounded zero for a particular observation i and a variable j, it holds that $x_{ij} < e_{ij}$, where e_{ij} is a threshold, i.e., the round-off error or the detection limit. In essence, rounded zeros represent a special case of missing values (Martín-Fernández et al., 2003) what implies a natural way for their treatment. In this paper, we focus on a parametric treatment based on a combination of the modified Expectation-Maximization (EM) alr algorithm introduced in Palarea-Albaladejo and Martín-Fernández (2008) with the ilr model-based technique for imputation of missing values in CODA introduced in Hron et al. (2010). The latter study proposes a procedure based on iterative regressions for the estimation of missing values. The use of the parametric approach in Palarea-Albaladejo and Martín-Fernández (2008) guarantees that the estimated values are below the reported detection limit.

In our approach, it follows that rounded zeros are associated with very small values that usually produce outlier samples in the distribution. The obvious reason for this is the relative scale of compositions as well as the properties of the Aitchison geometry. Unfortunately, the estimation of rounded zeros can itself be influenced by (other) outlying observations in the data set. In Hron et al. (2010), it is shown that compositional outliers need not necessarily be characterised by extreme absolute values in the parts, but can rather be characterised by extreme ratios between parts. To minimise the undesirable influence of deviating samples, robust counterparts to the standard statistical procedures are employed. This is achieved using robust rather than classical regressions in the iterative scheme. Robust regression automatically downweights outlying observations, i.e., observations that are outliers in either the space of the predictors or the residual space (Maronna et al., 2006).

In the following section, principles of compositional data analysis and published treatments for rounded zeros are reviewed. Section 3 presents the new proposed model, which builds on the proper use of ilr transformations. In section 4, we emphasise the advantages of using robust techniques in the context of interest in this study. Section 5 presents practical examples and comparisons of the proposed technique with available tools. Section 6 presents some conclusions and final remarks.

2. Compositional data and rounded zeros

Compositional data are inherently connected with the concept of relative scale and measurements of differences. For example, the differences between two pairs of observations, where the part of interest has concentrations of 1% and 2% for the first pair and 10% and 11% for the second pair, respectively, should not be considered equivalent. For the first pair, the ratio is 2/1=2, while for the second pair, the ratio is 11/10=1.1. In fact, any appropriate CODA should follow two main principles when dealing with this type of multivariate observations, as stated in Aitchison (1986). The first principle, known as *scale invariance*, can be summarised as follows: "Any meaningful characteristic of a composition should be invariant under a change of scale". In other words, no analysis should depend on the particular units in which the composition is expressed because proportional positive vectors represent the same composition. The second principle, known as *subcompositional coherence*, establishes that any analysis obtained from a set of a composition of D parts should not be in contradiction with that obtained from a sub-composition containing d parts, where $d < D$. The subcompositional coherence principle is usually the reason why CODA fails when it is conducted using standard statistical methods (Pearson, 1897; Aitchison, 1986). The log-ratio methodology in CODA guarantees that both principles are always verified. Nevertheless, it implies the following inherent difficulty: the *zeros problem*.

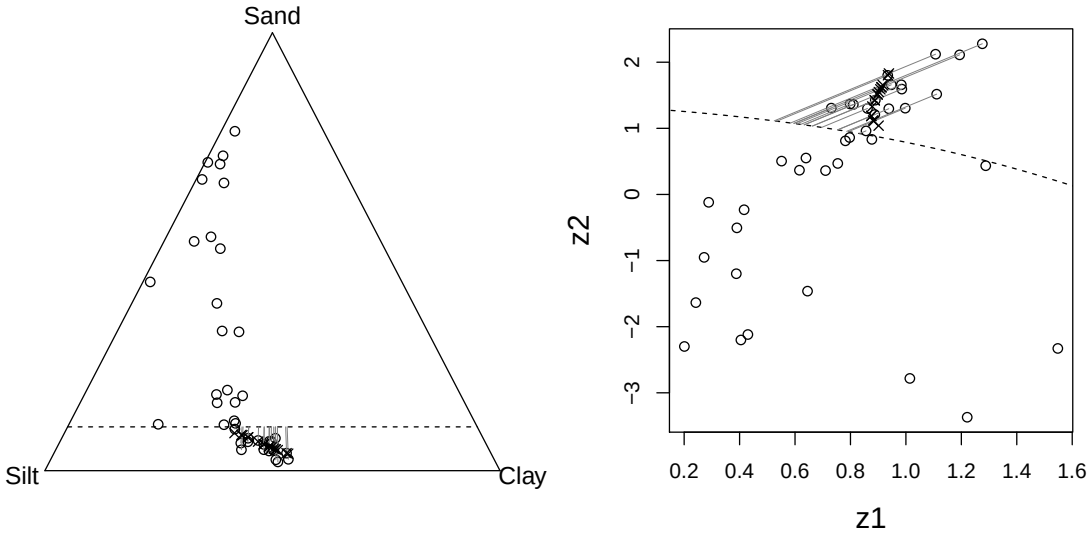
Rounded zeros can be conceptualised as a special type of censored data for which an upper threshold is known. They can be considered as fitting into an NMAR (not missing at random) mechanism of missing values (Little and Rubin, 1987) in the context of when the probability that one value is missing may depend on the missing value itself. Indeed, rounded zeros cannot be observed because their values are below a known threshold (Martín-Fernández et al., 2011). Thus, sound rounded zero replacement procedures are required in CODA for analysts to be able to apply the log-ratio methodology.

Using a non-parametric replacement approach, Martín-Fernández et al. (2003) recommends that values equal to 65% of the threshold be imputed to zeros and that the whole composition be adjusted by the so-called multiplicative modification. This strategy, which minimises the distortion of the covariance matrix, can be useful in cases where the percentage of zeros is not large (for example, less than 10%). Doing otherwise would lead to overestimation of the relationship between compositional parts, expressed in terms of the variation matrix (Aitchison, 1986), or, in general, to production of an undesirable bias in the covariance structure of the compositional data set.

Figure 1 illustrates the presence of a threshold within the Aitchison geometry (Egozcue and Pawlowsky-Glahn, 2006) in a simple case. The data set used for this example contains the proportions of sand, silt and clay in 39 sediment samples obtained at different water depths in an Arctic lake (Aitchison, 1986, p. 359). In Figure 1A, a forced threshold at level 10% in the part Sand is represented by a line. The samples with a value in part Sand below this threshold are considered as VBDL. Figure 1B shows the data set and the threshold line in coordinates, i.e., in the ilr-transformed space. As recommended in Martín-Fernández et al. (2003), to minimise the distortion of the covariance matrix, the ratios between the available data values should be preserved when any replacement of zeros is applied. In both figures, for the samples with VBDL in the part Sand, the trajectories of the ratios Silt/Clay constant are plotted. The positions of the imputed values, represented by the symbol $\times$, need to be present on these trajectories to minimise distortion. The estimation of the imputed values $\times$ is performed using the robust version of the ilr algorithm that will be described in the next section.

Nevertheless, a much more serious problem arises from the fact that the particular imputed values could significantly influence any statistical method applied to compositions after a log-ratio transformation. As an example, suppose that the true value of a component is barely below the threshold. Due to the relative scale of compositions, an imputation of a value of, for example, 65% of the threshold can turn the observation into a multivariate outlier, while it would suit the main data structure if the imputed value was correct. An outlier detection method (Filzmoser and Hron, 2008) would identify this observation, but it would be incorrectly flagged as an outlier simply because of an incorrect imputation. Such outliers can also have a severe effect on statistical estimates, such as the mean, the covariance or regression coefficients (Maronna et al., 2006).

Because the method of Martín-Fernández et al. (2003) is only appropriate when the percentage of zeros



(A) Simplex $\mathcal{S}^3 \subset \mathbb{R}^3$ (B) ilr-transformed space $\equiv \mathbb{R}^2$

Figure 1: Aitchison geometry of a threshold: (A) Simplex represented as a ternary diagram and (B) ilr-transformed space. Circles (∘) represent original samples; crosses (×) represent imputed samples using the robust ilr algorithm.

is below 10%, Palarea-Albaladejo et al. (2007) proposed a parametric replacement approach in which the alr transformation (Aitchison, 1986) is used in conjunction with an adapted version of the EM algorithm to generate plausible values for rounded zeros. In essence, the modification of the EM algorithm is based on a censored regression model proposed in Amemiya (1984). In Palarea-Albaladejo and Martín-Fernández (2008), it is shown that the modified EM algorithm performs somewhat better than the non-parametric approach, with an increasing relative amount of zeros in a compositional data set. However, this approach has associated with it some possible difficulties that need to be mentioned. Although the alr transformation appears to be convenient from a computational point of view, it is not isometric, as the ilr transformation is. This implies that the censored regression method used in the algorithm can be influenced by outlying observations in the residual alr space that are not outliers in the simplex space. Thus, when the regression method is robustified by downweighting the influence of outliers, it might still result in a strong bias in the resulting imputed values.

3. Isometric log-ratio approach

From Egozcue et al. (2003), it is well known that in CODA, the selection of an orthonormal basis for the simplex fully determines a particular ilr transformation. Following an ilr transformation, it is generally not possible that each of the resulting ilr variables can be assigned to an individual compositional part, in the sense that the ilr variable explains all of the relative information of this particular part. This is problematic in a setting where the estimation of rounded zeros is achieved by regression on the individual parts. However, D different ilr transformations can be constructed, each consisting of D-1 new variables, such that the first ilr variable always contains all of the relative information on the i-th part of the original composition (Fišerová and Hron, 2011). To be specific, each D-part composition $\mathbf{x} = (x_1, \ldots, x_D)^t$ can be associated with another composition $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})^t$ resulting from a permutation of the parts $\mathbf{x} = (x_1, \ldots, x_D)^t$, where the l-th part is moved to the first position, i.e., $\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)^t$.

The ilr transformation

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\,\ln\frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ i=1,\ldots,D-1, \tag{1}$$

transforms compositions $\mathbf{x}^{(l)}$ into $(D-1)$-dimensional real vectors $\mathbf{z}^{(l)} = (z_1^{(l)},\ldots,z_{D-1}^{(l)})^t$, $l=1,\ldots,D$. The inverse transformation of $\mathbf{z}^{(l)}$ to the original (permuted) composition $\mathbf{x}^{(l)}$ is then given by

$$\begin{aligned}
x_1^{(l)} &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}} z_1^{(l)}\right), \\
x_i^{(l)} &= \exp\left(-\sum_{j=1}^{i-1}\frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)} + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i^{(l)}\right), i=2,\ldots,D-1, \text{ and} \\
x_D^{(l)} &= \exp\left(-\sum_{j=1}^{D-1}\frac{1}{\sqrt{(D-j+1)(D-j)}} z_j^{(l)}\right).
\end{aligned} \tag{2}$$

Consequently, the obtained composition can be represented as vectors with a constant sum constraint, such as proportions or percentages. Note that the first variable $z_1^{(l)}$ includes all parts of the original composition $\mathbf{x}$. The coordinates $z_2^{(l)},\ldots,z_{D-1}^{(l)}$ explain the remaining log-ratios in the composition (Fišerová and Hron, 2011). Although $z_1^{(l)}$ *explains all of the information concerning* x_l, we cannot say that $z_1^{(l)}$ *is* the original compositional part x_l, but rather that it *stands for* x_l. The remaining ilr variables can, in principle, be chosen arbitrarily because they do not represent any specific compositional part. Different ilr transformations are orthogonal rotations of each other (Egozcue et al., 2003).

According to Hron et al. (2010), the choice of the above ilr transformation is particularly useful when the goal is to predict x_l from the rest of the composition. Thus, in the following discussion, a new parametric treatment for rounded zeros will be introduced using multiple regression analysis for compositional data. This ilr approach will be combined with the censored regression suggested in the modified EM algorithm Palarea-Albaladejo and Martín-Fernández (2008).

Consider a compositional data set $\mathbf{X} = [x_{ij}]$ with n observations (rows) and D compositional parts (columns). Initially, for each part x_l that includes rounded zeros, the data set must be ilr-transformed into an unconstrained real data set $\mathbf{Z}^{(l)} = [z_{ij}^{(l)}]$ using equation (1). Note that because all parts are involved in equation (1), a complete version of the data matrix with initialised values for the rounded zeros is required as a *starting point* for the iterative EM algorithm. This initialisation can be achieved by multiplicative replacement, 65% of the threshold, as described in Martín-Fernández et al. (2003). In this way, the initial distortion of the covariance structure of the data set is minimised. The algorithm is consistent with the concept of iterative model-based imputation of missing values for compositional data, as described in Hron et al. (2010). Note that here the compositional parts are sorted, in descending order, according to the number of rounded zero values to be imputed (let parts $x_1,\ldots,x_D$ correspond to such a configuration), and then, in each step of the algorithm, a regression of $z_{i1}^{(l)}$ on $z_{i2}^{(l)},\ldots,z_{i,D-1}^{(l)}$ is applied to impute values in $x_1^{(l)}$, $l=1,\ldots,D$. The estimated values from the preceding steps are used to perform the regression.

Imputation of rounded zeros requires a modification to this general algorithm because the imputed values must not exceed the value of the threshold(s). Let $e_{i1}^{(l)} \equiv e_{il}$ be the thresholds in the l-th compositional part of the original data set $\mathbf{X}$; then, the ilr transformation of rounded zeros, when $x_{i1}^{(l)} < e_{i1}^{(l)}$ occurs, results in unknown values $z_{i1}^{(l)}$ with the property $z_{i1}^{(l)} < \psi_{i1}^{(l)}$, where

$$\psi_{i1}^{(l)} = \sqrt{\frac{D-1}{D}}\,\ln\frac{e_{i1}^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{ij}^{(l)}}}. \tag{3}$$

In the following we use the notation $\mathbf{Z}^{(l)} = [\mathbf{z}_1^{(l)}, \mathbf{Z}_{-1}^{(l)}]$, where $\mathbf{z}_1^{(l)}$ is the first column of the matrix $\mathbf{Z}^{(l)}$ and $\mathbf{Z}_{-1}^{(l)}$ contains the remaining columns. According to Palarea-Albaladejo and Martín-Fernández (2008), in the t-th step of the iteration process, our proposed algorithm replaces unknown values in the variable $\mathbf{z}_1^{(l)}$ by its conditional expected value

$$E[\mathbf{z}_1^{(l)} | \mathbf{Z}_{-1}^{(l)}, \mathbf{z}_1^{(l)} < \boldsymbol{\psi}_1^{(l)}], \tag{4}$$

using all other variables as explanatory variables. This procedure assumes that the ilr-transformed compositions follow a normal distribution: the so-called ilr-normal distribution or normal distribution in the simplex (Mateu-Figueras and Pawlowsky-Glahn, 2008). Each unknown value $z_{i1}^{(l)}$ in the variable $\mathbf{z}_1^{(l)}$ is replaced by

$$\hat{z}_{i1}^{(l)} = {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}^{(l)} - \hat{\sigma}^{(l)} \frac{\phi\left(\frac{\psi_{i1}^{(l)} - {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}{\Phi\left(\frac{\psi_{i1}^{(l)} - {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}^{(l)}}{\hat{\sigma}^{(l)}}\right)}, \tag{5}$$

where ϕ and Φ are the density and distribution functions of the standard normal distribution, respectively; $\hat{\sigma}^{(l)}$ is the estimated conditional standard deviation of the variable $\mathbf{z}_1^{(l)}$; and $\hat{\boldsymbol{\beta}}^{(l)}$ denotes the vector of estimated coefficients of the linear regression of $\mathbf{z}_1^{(l)}$ on $\mathbf{Z}_{-1}^{(l)}$ (for such rows of the data matrix, where the values of $\mathbf{z}_1^{(l)}$ are observed). More details on the estimation of the regression parameters will be provided in section 4. Note that the second term in equation (5), the novelty of the censored regression Amemiya (1984)], guarantees that imputed values are below the threshold. After the treatment of the variable $\mathbf{z}_1^{(l)}$, the completed data set is transformed back to the simplex by the inverse ilr transformation (equation 2). The next compositional part is then considered for imputation. In this way, all compositional parts containing rounded zeros are sequentially imputed. The entire procedure is repeated iteratively until convergence is reached, i.e., when the Frobenius matrix norm of the difference between the empirical covariance matrices computed from the ilr-transformed data according to equation (1) from the current and the previous iteration is smaller than a specified limit ($\zeta = 0.0001$). After convergence, the last completed data set produced by the algorithm is back-transformed to the simplex using equation (2). In this way, a compositional data set is obtained where the rounded zeros have been replaced by estimated values below the detection limit(s).

The full algorithm can be summarised in the form of the following pseudo code:

```
Stage 1:  Preliminaries.
 1 Initialise the rounded zeros to a value of 65% of the threshold for the
corresponding compositional part
 2 Sort the compositional parts by decreasing order of the number of their
rounded zeros

Stage 2:  Impute missing values and check for convergence.
 3 FOR l = 1,...,D:
 4   ilr-transform the data and thresholds by equations (1) and (3)
respectively
 5   Replace rounded zeros by conditional expected values using equation (5)
 6   Back-transform the data using equation (2)
 7 END FOR

Stage 3:  Restore original format.
 8 Rearrange the parts in the original order.
```

It is interesting to note that an equivalence exists between an approach based on an ilr transformation and one based on an alr transformation; namely, in the general case, multiple linear regression models based on least-squares estimation applied in the alr-transformed space and in the ilr-transformed space produce

exactly the same estimation of the response values. A proof is given in Appendix A. This fact implies that the term $\mathbf{z}_{i,-1}^{(l)} \cdot \hat{\boldsymbol{\beta}}^{(l)}$ in equation (5) gives exactly the same estimates of the response variable with both log-ratio transformations. A proof that censored regression as given in (5) is also equivalent in both log-ratio transformations is given in Appendix B.

4. Robust treatment of rounded zeros

When *classical* least-squares regression is used, the model described by equation (5) results in the same treatment of rounded zeros as the modified EM alr algorithm introduced in Palarea-Albaladejo and Martín-Fernández (2008). Because real data frequently contain outliers or deviating data points, a more robust regression approach is preferable (Maronna et al., 2006). Furthermore, rounded zeros are assumed to represent very small values that might potentially have a strong influence if they appear in the divisor of a log-ratio transformation. In this case, the ilr approach as proposed here is preferable over an alr-based method because it can be robustified easily and because the treatment of the outliers in the regression model is consistent with the Aitchison geometry.

Figure 2 illustrates an example in which both classical and robust regressions were applied to the same data set. For this example, the Arctic lake data set shown in Figure 1 was used. In both cases, a linear regression of coordinate $\mathbf{z}_2$ to $\mathbf{z}_1$ was applied. Figure 2 (left) shows the resulting regression lines, from which it is evident that the classical regression (the dashed line) is seriously affected by the outliers. The robust regression (the solid line), conversely, captures the trend in the data much better. Figure 2 (right) shows both lines back-transformed to the ternary diagram where the original data set is plotted. Here as well, it is evident that the robust regression fit of the data structure is much better than the classical regression fit.

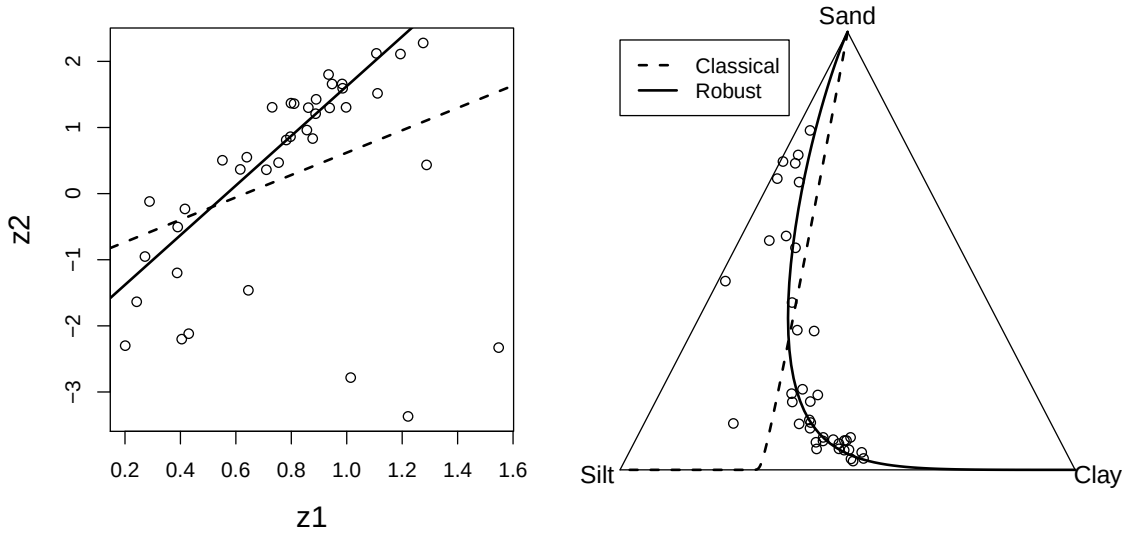


Figure 2: Classical (dashed line) and robust (solid line) regression of variable z_2 against z_1 in the ilr-transformed space (left), and back-transformation of the results to the ternary diagram (right).

Generally, in multiple linear regression, we consider a model of the form

$$\gamma_i = \boldsymbol{\xi}_i^t \boldsymbol{\beta} + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \tag{6}$$

with n observations of the response γ and of the explanatory variables $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})^t$ (an intercept term is included by setting $\xi_{i1} = 1$), the vector of regression coefficients $\boldsymbol{\beta}$ and the error term ε_i. The classical

least-squares (LS) estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{LS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\gamma_i - \boldsymbol{\xi}_i^t \boldsymbol{\beta}\right)^2 . \tag{7}$$

The LS estimator is considered the best linear unbiased estimator (BLUE) if the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with zero mean and the same residual variance σ^2 (e.g., Johnson and Wichern, 2002). If these strict assumptions are violated, the LS estimator loses its desirable properties, and another estimator might be preferable. A robust estimation of the regression parameters can be achieved using a so-called MM estimator for regression, which is defined as

$$\widehat{\boldsymbol{\beta}}_{MM} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{\gamma_i - \boldsymbol{\xi}_i^t \boldsymbol{\beta}}{\hat{\sigma}}\right) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right). \tag{8}$$

The function ρ can be treated as a weighting function applied to the residuals $r_i(\boldsymbol{\beta}) = \gamma_i - \boldsymbol{\xi}_i^t \boldsymbol{\beta}$, which are scaled by a robust scale estimator $\hat{\sigma}$. Obviously, when the scale $\hat{\sigma}$ is omitted, the LS criterion described by equation (7) is a special case of equation (8) with $\rho(r) = r^2$. The idea of robust regression is to appropriately downweight the influence of large (absolute) residuals. Accordingly, the specific choice of the ρ function will affect the properties of the resulting regression estimator (Maronna et al., 2006). Note that the *M estimator* for regression (Huber, 1981) is defined as in equation (8), with the important difference being the choice of algorithm used to solve the minimisation problem. As opposed to the M estimator, the MM estimator achieves the highest possible level of robustness combined with a tunable efficiency. Computation of the MM estimator is achieved using a highly robust but inefficient initial estimator for the regression coefficients (an S estimator). This produces an initial estimation of the residuals, which are used to obtain $\hat{\sigma}$ in equation (8) by a so-called M estimator of scale. Finally, the MM estimator can be computed iteratively, starting from the initial S estimator. For more details on the MM estimator algorithm and statistical properties, refer to Maronna et al. (2006).

Here, robust regression is applied to the censored regression model described by equation (5). The MM estimator is obtained by solving

$$\hat{\boldsymbol{\beta}}_{MM}^{(l)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{z_{i1}^{(l)} - (\mathbf{z}_{i,-1}^{(l)})^t \boldsymbol{\beta}}{\hat{\sigma}^{(l)}}\right). \tag{9}$$

Both the resulting regression coefficients $\hat{\boldsymbol{\beta}}_{MM}^{(l)}$ and the estimated robust residual scale $\hat{\sigma}^{(l)}$ are then used to adjust the prediction according to equation (5):

$$\hat{z}_{i1}^{(l)} = {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}_{MM}^{(l)} - \hat{\sigma}^{(l)} \frac{\phi\left(\frac{\psi_{i1}^{(l)} - {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}_{MM}^{(l)}}{\hat{\sigma}^{(l)}}\right)}{\Phi\left(\frac{\psi_{i1}^{(l)} - {\mathbf{z}^{(l)}}_{i,-1}^{t} \cdot \hat{\boldsymbol{\beta}}_{MM}^{(l)}}{\hat{\sigma}^{(l)}}\right)}. \tag{10}$$

The censored regression is performed iteratively, as outlined in section 3. Note that because we deal with an equivariant regression estimator (Maronna et al., 2006), a change of the ilr basis does not alter the results (see, e.g. Filzmoser and Hron, 2011).

5. Practical examples

To compare the performance of the classical and robust methods used in this study, several different data sets and scenarios were considered. We focused our comparison of the classical and robust methods on the ilr approach outlined in section 3. Hereinafter, we refer to the classical ilr approach as CI and to its robust version as RI. We tested these two possibilities, CI and RI, using real and simulated data sets.

5.1. Compositional statistics and measures of distortion

Initially, we considered a data set $\mathbf{X}$ without rounded zeros. Taking a range of realistic thresholds, we transformed observed values that were smaller than the threshold to rounded zeros. The compositional data set resulting from this procedure was denoted by $\mathbf{X}^*$. The two regression methods were then applied to replace the rounded zeros to obtain completed compositional data sets. For a performance evaluation of the two methods, we used the following three basic descriptive measures (e.g., Egozcue and Pawlowsky-Glahn, 2011) typically obtained from a compositional data set $\mathbf{X}$ formed by n compositions from the simplex S^D:

- Central tendency: *compositional geometric mean* $cen(\mathbf{X}) = (\frac{g_1}{\sum g_i}, \ldots, \frac{g_D}{\sum g_i})$, where g_i is the geometric mean of the part $\mathbf{x}_i$ (ith column of the data matrix $\mathbf{X}$).
- Total variance: $totvar(\mathbf{X}) = \frac{1}{n}\sum_{k=1}^{n} d_a^2(\mathbf{x}_k, cen(\mathbf{X}))$, where $\mathbf{x}_k$ is the k-th row of the data matrix. We denote by d_a the Aitchison distance between two compositions $\mathbf{x}$ and $\mathbf{y}$, defined as $d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D}\sum_{i<j}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2}$.
- Variability: *compositional variation matrix* $\mathbf{T}$, the symmetric matrix containing the log-ratio variances $\tau_{ij} = \mathrm{Var}(\ln(\mathbf{x}_i/\mathbf{x}_j))$ between two parts $\mathbf{x}_i$ and $\mathbf{x}_j$, $i, j = 1, \ldots, D$.

Note that the compositional geometric mean plays the role of the multivariate mean when the sample space is the simplex. In addition, *totvar* evaluates the spread of the data, and the matrix $\mathbf{T}$ completely determines the covariance structure of a compositional data set. In other words, given the variation matrix $\mathbf{T}$, the covariance matrix in any log-ratio transformed space can be obtained (e.g., Tolosana-Delgado et al., 2011).

For our purposes, two measures of distortion with respect to the original samples were also computed:

- *Relative difference in covariance matrix* (RDCM): Let $\boldsymbol{S} = [s_{ij}]$ be the sample covariance matrix of the original ilr-transformed observations z_{ij}, and let $\boldsymbol{S}^* = [s_{ij}^*]$ be the sample covariance matrix computed with the same ilr-transformed observations for which all the rounded zeros have been imputed. The measure of the relative difference between both covariance matrices, based on the Frobenius matrix norm $\|\cdot\|_F$ (e.g., Seber, 2008, p. 68), is

$$\frac{\|\boldsymbol{S} - \boldsymbol{S}^*\|_F}{\|\boldsymbol{S}\|_F} = \frac{\sqrt{\sum_{i,j=1}^{D-1}\left(s_{ij} - s_{ij}^*\right)^2}}{\sqrt{\sum_{i,j=1}^{D-1} s_{ij}^2}}. \tag{11}$$

- *Compositional error deviation* (CED):

$$\frac{\frac{1}{n_M}\sum_{k\in M} d_a(\mathbf{x}_k, \mathbf{x}_k^*)}{\max_{\{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}\}}\{d_a(\mathbf{x}_i, \mathbf{x}_j)\}}, \tag{12}$$

  is a generalisation of the measure applied in Hron et al. (2010). Here, n_M is the number of samples $\mathbf{x}_k$ containing at least one rounded zero and M is the index set referring to such samples. The denominator is the maximum distance in the original data set.

The basic properties of the Frobenius norm and the Aitchison distance make it possible to establish that both measures are basis invariant. That is, both measures are preserved when a change of the ilr basis is applied. The former criterion, RDCM, measures the influence of the imputation to the covariance structure. The factor $\|\boldsymbol{S}\|_F$ in the denominator allows interpretation of RDCM as a relative error in relation to the original covariance structure. For the case of an imputation method that does not produce distortion, the optimal value RDCM=0 is achieved. If the imputation method produces a data set $\mathbf{X}^*$ composed of

only one multivariate data point, i.e., it has a null covariance structure, $s^*_{ij} = 0,\ i, j = 1, \ldots, D-1$, then RDCM=1. The latter criterion, CED, evaluates the relative distortion in those samples that include at least one rounded zero. The numerator of CED consists of an average of the deviation between the original and imputed samples. To avoid the undesirable effect of the amount of spread in the data set, it is divided by the maximum distance within the data set. In this way, CED becomes a scaled measure.

5.2. Simulation study

For the simulation study, a 6-part random composition $\mathbf{x} \in S^6$ was considered, with centre (in %) $cen(\mathbf{x}) = (2.5, 5.0, 20.0, 15.0, 7.5, 50.0)^t$ and variation matrix

$$\mathbf{T} = \begin{pmatrix} 0 & 2.33 & 0.48 & 1.93 & 0.74 & 2.00 \\ & 0 & 2.98 & 0.08 & 1.11 & 0.20 \\ & & 0 & 2.31 & 0.97 & 2.50 \\ & & & 0 & 0.79 & 0.11 \\ & & & & 0 & 1.00 \\ & & & & & 0 \end{pmatrix}.$$

From the compositional geometric mean $cen(\mathbf{x})$, it can be seen that smaller values were concentrated primarily in the parts x_1, x_2 and x_5. According to (Aitchison, 1986), a log-ratio variance close to zero suggested a meaningful association among the involved parts. The values in the variation matrix $\mathbf{T} = (\tau_{ij})$ indicated that the strongest related parts were x_2, x_4 and x_6 because their highest value of τ_{ij} is 0.20. The parts x_1 and x_3 had a weak association ($\tau_{13} = 0.48$) which was nonetheless stronger than the association between parts x_4 and x_3 with x_5, where $\tau_{45} = 0.79$ and $\tau_{35} = 0.97$. The rest of the associations were weaker because the values in the matrix $\mathbf{T}$ were in the range of 1.00 to 2.98. Note that D clr coefficients form a hyperplane of transformed compositions in $\mathbf{R}^D$. Thus, once the matrix $\mathbf{T}$ was fixed, the covariance matrix of the ilr-transformed random vector, $\mathbf{\Sigma}$, was obtained easily using the equality $\mathbf{\Sigma} = -\frac{1}{2}\mathbf{V}^t\mathbf{T}\mathbf{V}$, where the columns of the $D \times (D-1)$ matrix $\mathbf{V}$ are vectors of the orthonormal basis in the clr-transformed space (e.g., Tolosana-Delgado et al., 2011).

Using the above centre and variation matrix, a simulated data set $\mathbf{X}$ consisting of $n = 300$ compositions was produced with Monte Carlo simulations under a multivariate ilr-normal model (Mateu-Figueras and Pawlowsky-Glahn, 2008). The resulting compositional geometric mean $cen(\mathbf{X})$ (Table 1) was very close (Aitchison distance 0.085) to the *true* centre. The 25th and 75th percentiles showed that most of the small values were concentrated in parts $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_5$. In addition, the RDCM, described by equation (11), equal to 0.04, was obtained between the sample ilr-covariance matrix $\mathbf{S}$ and the *true* values in $\mathbf{\Sigma}$.

Table 1: Univariate descriptive statistics of the data set $\mathbf{X}$ (concentrations in %).

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$
min	0.14	0.15	0.43	0.84	0.81	1.89
p25	1.15	2.83	7.48	10.06	3.85	30.84
$cen(\mathbf{X})$	2.66	4.92	21.14	14.73	7.51	49.04
p75	4.23	6.75	38.82	16.92	9.42	60.03
max	23.00	13.91	90.41	27.81	26.60	79.28

The reproduced associations among the parts were observed in the clr-biplot (i.e., the biplot in the clr-transformed space) in Figure 3. The rays of parts $\mathbf{x}_2$, $\mathbf{x}_4$ and $\mathbf{x}_6$ were the closest, revealing that they were the parts with strongest relations. Note that the quality of the clr-biplot was reasonably high because the axes explained 91% of the total variance. In addition, the fact that the rays of the parts $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_5$ approximately formed an equilateral triangle, combined with the fact that smaller values were mostly concentrated in these parts (Table 1), explained the spherical shape of the cloud of points in the clr-biplot.

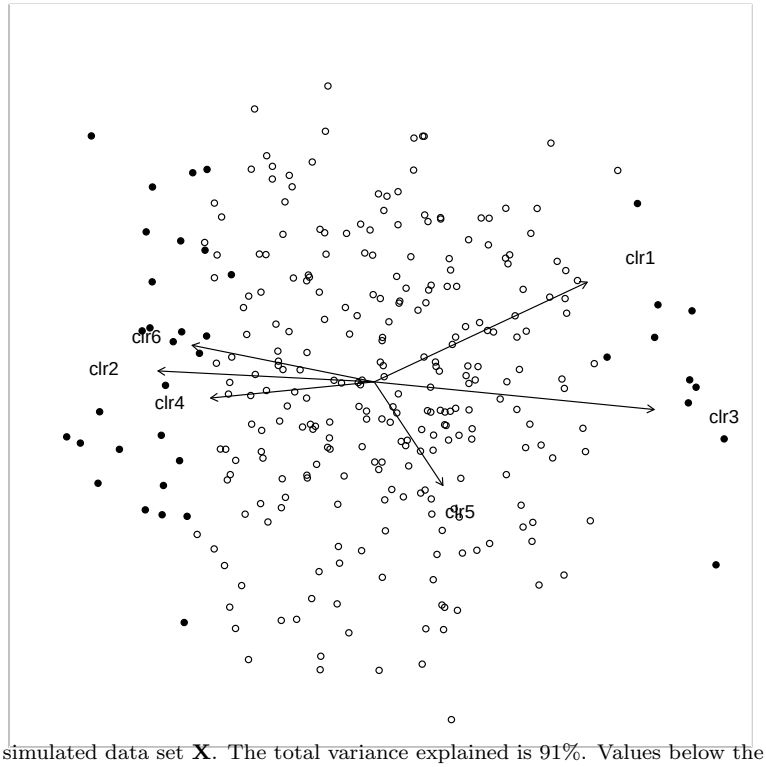


Figure 3: Clr-biplot of the simulated data set $\mathbf{X}$. The total variance explained is 91%. Values below the detection limit of the samples plotted as filled circles, are forced to zero in the scenario $\mathbf{X}_2^*$. See the text for more details.

Note that some samples were plotted using filled circles in Figure 3. These samples corresponded to the samples with values below the detection limit (VBDL). These values were forced, i.e., rounded, to zero in the simulation study. Observe that most such samples were far away from the centre of the distribution, suggesting that they were candidates to be classified as outliers. This example illustrates why robust statistics is an appropriate tool to utilise with rounded zeros in CODA.

From the original data set $\mathbf{X}$, a class of artificial data sets $\mathbf{X}^*$ with a different distribution of rounded zeros was created. For this data set, once the initial threshold, the detection limit (DL), for each element was fixed, a set of different DLs was considered (Table 2) in order to define twelve different scenarios. The DLs were accordingly established for each part at each level, as shown in Table 2. As a result, twelve different scenarios, i.e., twelve different sets, $\{\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_{12}^*\}$, were established. Note that an increase in the number of rounded zeros was obtained across the scenarios and that, simultaneously, there were different patterns in the number of zeros in the parts. For each scenario, we simulated 1000 data sets with similar percentage distributions of rounded zeros in each part, as shown by the percentages in parentheses in Table 2. Note that the forced rounded zeros were mainly concentrated in parts $\mathbf{x}_1^*$, $\mathbf{x}_2^*$ and $\mathbf{x}_5^*$, with approximately around 40% in the last scenario. Part $\mathbf{x}_3^*$ had some zeros in all scenarios but not many, with a maximum of 16% in scenario $\mathbf{X}_{12}^*$. Parts $\mathbf{x}_4^*$ and $\mathbf{x}_6^*$ were those with the lowest number of rounded zeros, below 9% in all cases. For the more extreme scenario of $\mathbf{X}_{12}^*$, the 1000 simulated data matrices had approximately 27% of their entries equal to zero.

After simulating the data sets with forced zeros, $\mathbf{X}_i^*$, $i = 1, \ldots, 12$, both imputation methods, CI and RI, were applied and the corresponding completed data sets were obtained. The performance of the two methods was compared using the above statistics. Figure 4 (left) shows that the two methods, CI and RI, produced very similar results in the *worst-case* scenario ($\mathbf{X}_{12}^*$) when the number of VBDL was the greatest and all of the parts contained zeros. Even in this worst-case scenario, both methods estimated the centre of the distribution reasonably well. The plot shows the interval percentiles (p5, p95) for each part of the compositional geometric mean (i.e., the sample centre) $cen(\mathbf{X}_{12}^*)$ across all of the simulations and both imputation methods. The true values of the $cen(\mathbf{x}) = [2.5, 5.0, 20.0, 15.0, 7.5, 50.0]$ (in %). Note that the intervals had different lengths, suggesting the relative nature of the scale of proportions. When these intervals were plotted in the ilr-transformed space (not shown here), their lengths were comparable. Figure 4 (right) shows the results of CI and RI for the estimated total variance across all of the 1000 simulations in the 12 scenarios, where the true value $totvar(\mathbf{X}) = 3.254$ is indicated by a horizontal line.

Table 2: Twelve synthetic rounded zeros scenarios. The associated percentage of rounded zeros in each part are shown in parentheses. The column at the right shows the total percentage of rounded zeros in the set.

	Detection limits in % (percentage of rounded zeros in the part)						
Scenario	$\mathbf{x}_1^*$	$\mathbf{x}_2^*$	$\mathbf{x}_3^*$	$\mathbf{x}_4^*$	$\mathbf{x}_5^*$	$\mathbf{x}_6^*$	Total
$\mathbf{X}_1^*$	0.15 (0.33)	0.15 (0.00)	2.50 (7.67)	2.50 (3.33)	0.15 (0.00)	5.00 (2.00)	2.22
$\mathbf{X}_2^*$	0.30 (3.33)	0.50 (3.00)	2.75 (8.67)	2.75 (3.33)	0.65 (0.00)	6.00 (2.67)	3.50
$\mathbf{X}_3^*$	0.45 (7.33)	0.85 (4.67)	3.00 (9.00)	3.00 (4.00)	1.15 (1.00)	7.00 (3.33)	4.89
$\mathbf{X}_4^*$	0.60 (12.33)	1.20 (7.33)	3.25 (9.33)	3.25 (4.00)	1.65 (4.00)	8.00 (3.67)	6.78
$\mathbf{X}_5^*$	0.75 (16.67)	1.55 (10.33)	3.50 (10.00)	3.50 (4.00)	2.15 (8.33)	9.00 (3.67)	8.83
$\mathbf{X}_6^*$	0.90 (20.67)	1.90 (14.00)	3.75 (10.33)	3.75 (4.00)	2.65 (12.67)	10.00 (5.00)	11.11
$\mathbf{X}_7^*$	1.05 (24.00)	2.25 (18.67)	4.00 (11.33)	4.00 (5.33)	3.15 (18.00)	11.00 (5.33)	13.78
$\mathbf{X}_8^*$	1.20 (27.00)	2.60 (21.67)	4.25 (12.33)	4.25 (5.33)	3.65 (21.33)	12.00 (6.00)	15.61
$\mathbf{X}_9^*$	1.35 (30.67)	2.95 (26.33)	4.50 (13.00)	4.50 (6.33)	4.15 (29.33)	13.00 (6.67)	18.72
$\mathbf{X}_{10}^*$	1.50 (32.33)	3.30 (33.67)	4.75 (14.67)	4.75 (6.67)	4.65 (34.67)	14.00 (7.33)	21.56
$\mathbf{X}_{11}^*$	1.65 (35.67)	3.65 (40.00)	5.00 (15.00)	5.00 (7.67)	5.15 (40.33)	15.00 (8.00)	24.44
$\mathbf{X}_{12}^*$	1.80 (38.33)	4.00 (46.67)	5.25 (15.67)	5.25 (8.33)	5.65 (46.00)	16.00 (8.00)	27.17

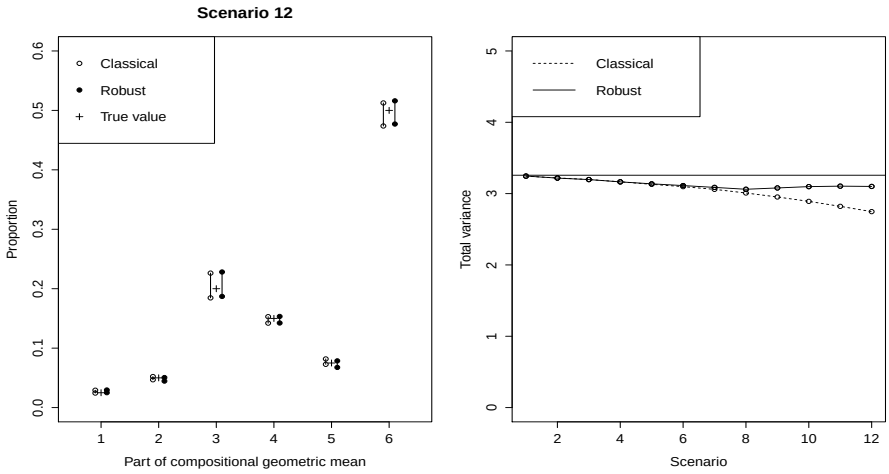


Figure 4: Basic compositional statistics: Interval percentile (p5, p95) for each part of the compositional geometric mean in Scenario 12 (left); total variance for each scenario. The horizontal line represents the true value $totvar(\mathbf{X}) = 3.254$ (right).

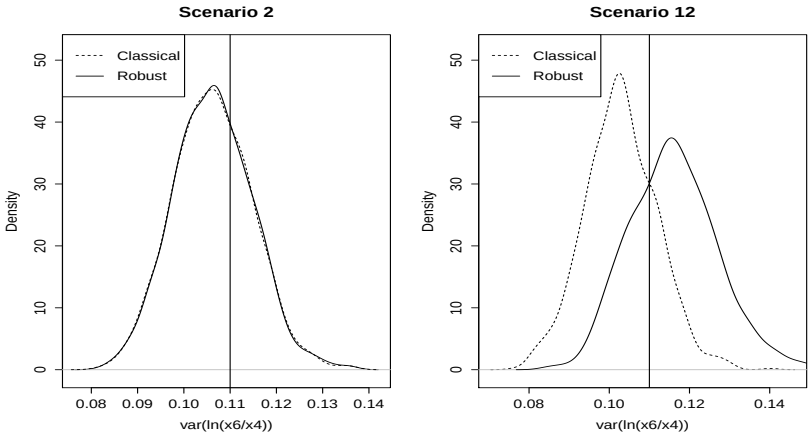
Both methods tended to underestimate the true value, suggesting that the imputed values led to a narrower distribution. For scenarios $\mathbf{X}_7^*$ to $\mathbf{X}_{12}^*$ (with increasing numbers of VBDL), the robust method RI exhibited better performance than CI. Surprisingly, RI approached the true value in the last scenario. This is related to the pattern in the estimation of the elements from the variation matrix $\mathbf{T} = (\tau_{ij})$. Total variability is, by definition, a measure that accumulates the variability from log-ratios of all the parts. Consequently, it is possible for a given method to underestimate the log-ratio variance of some parts and overestimate others, with the total variance being ultimately well estimated. To investigate this in more detail, the estimates of the elements of the variation matrix were computed for the two imputation methods. A similar general pattern was observed for all of the elements $\tau_{ij} = var(\ln(\mathbf{x}_i/\mathbf{x}_j))$, $i, j = 1, \ldots, 6$ of the matrix. For scenarios in which the number of VBDL is small, the estimation was approximately unbiased; as the number of zeros increases, the CI method tended to underestimate the true value more. RI sometimes underestimated and sometimes overestimated the true value, but it usually yielded larger values than CI did. To illustrate this tendency, in Figure 5, we plotted the distribution of the following estimates of three entries in the variation matrix: the smallest, $\tau_{64} = var(\ln(\mathbf{x}_6/\mathbf{x}_4)) = 0.11$, an intermediate value $\tau_{64} = var(\ln(\mathbf{x}_6/\mathbf{x}_4)) = 1.11$, and the largest $\tau_{32} = var(\ln(\mathbf{x}_3/\mathbf{x}_2)) = 2.98$. The corresponding true values were plotted as vertical lines in the figures. For each element, we plotted the scenario $\mathbf{X}_2^*$ at the left and the worst-case scenario $\mathbf{X}_{12}^*$ at the right. Note that CI always tended to yield a lower variance value, resulting in an underestimation of the total variance (Figure 4). However, when the number of VBDL was increased, RI overestimated the variance for some log-ratios and underestimated it for others, resulting in a better final estimation of the total variance (Figure 4).

Figure 6 shows the results for the measures of distortion. The results referred to the averages of the measures, computed over all 1000 simulations for each of the 12 scenarios. Figure 6 (left) compares the resulting RDCM for RI and CI. As the percentage of zeros was increased, the performance of both methods worsened. However, in the scenarios with a high number of VBDL, the robust method performed better than the classical one, suggesting that RI better estimates the covariance structure of the data set. Figure 6 (right) presents the results for the CED. As the percentage of zeros was increased, the CI method performed better than the RI method. In other words, the distance between the original and the imputed values was smaller for CI. The differences between CI and RI, however, were still small and would increase if *artificial* outliers were added to the data. A scenario with just such a *contamination* of outliers is analysed in the next section.

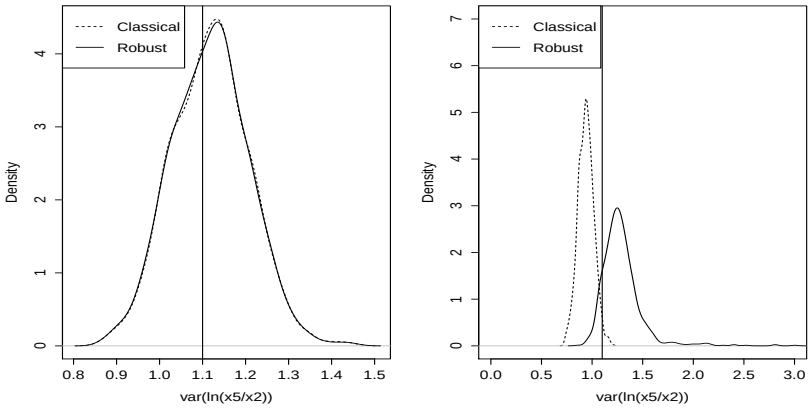
5.3. Real data set

We used a data set from the so-called *Kola project* (see http://www.ngu.no/Kola), a geochemical mapping project, covering an area of 188,000 km^2 north of the Arctic Circle. This project was conducted from 1992—1998 by the Geological Surveys of Finland (GTK) and Norway (NGU) and the Central Kola Expedition (CKE) in Russia. Approximately 600 soil samples in different layers were collected and analysed for the concentration of more than 50 chemical elements (Reimann et al., 1998). The complete data set is available in the R package StatDA (R development core team, 2008). To test the two imputation methods described in this paper, the data from the *moss* layer were used. Our data set, denoted by $\mathbf{X}$, was composed of 26 chemical elements measured in 594 samples, without rounded zeros. In contrast to the approach taken in the simulation study described above, we focused on only one part $\mathbf{x}_j$ of the compositional data set $\mathbf{X}$, and we set some values in this part to zero. We considered a sequence of 16 sample quantiles q_i corresponding to selected percentiles, from 5% to 95%, and transformed every observed value of this part smaller than q_i to a zero value (i.e., a VBDL smaller than q_i). We denoted by $\mathbf{X}_i^*$, $i = 1, \ldots, 16$, the compositional data sets resulting from this procedure. Next, the two imputation methods, CI and RI, were applied to each data set $\mathbf{X}_i^*$, and the final results were compared.

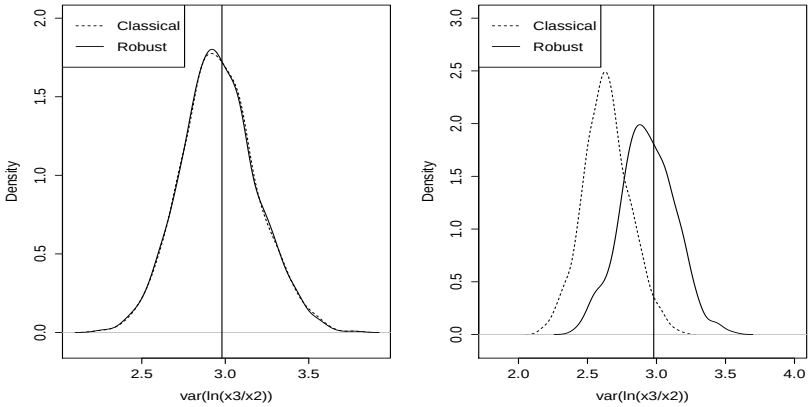
In our case study, we concentrated on the element vanadium (V). In the Kola project area, the trace element V is mostly generated by emissions from the industrial centres in Murmansk, Monchegorsk, Kirovsk, Zapoljarnij, Kovdor and Nikel. Oil combustion can be another significant source of V emissions into the atmosphere. Additionally, in the uncontaminated Finnish project area, a general increase of V levels in the moss layer from south to north is evident, which could be caused by the increasing population and traffic density. Low levels of V could thus be an indication of uncontaminated areas.



(A) Log-ratio variance τ_{64}



(B) Log-ratio variance τ_{52}



(C) Log-ratio variance τ_{32}

Figure 5: Kernel density estimation of the distribution of the elements of the variation matrix: (A) $\tau_{64} = var(\ln(x_6/x_4))$; (B) $\tau_{52} = var(\ln(x_5/x_2))$; (C) $\tau_{32} = var(\ln(x_3/x_2))$. Results for scenario $\mathbf{X}_2^*$ (left) and for scenario $\mathbf{X}_{12}^*$ (right). Vertical lines represent the true values.

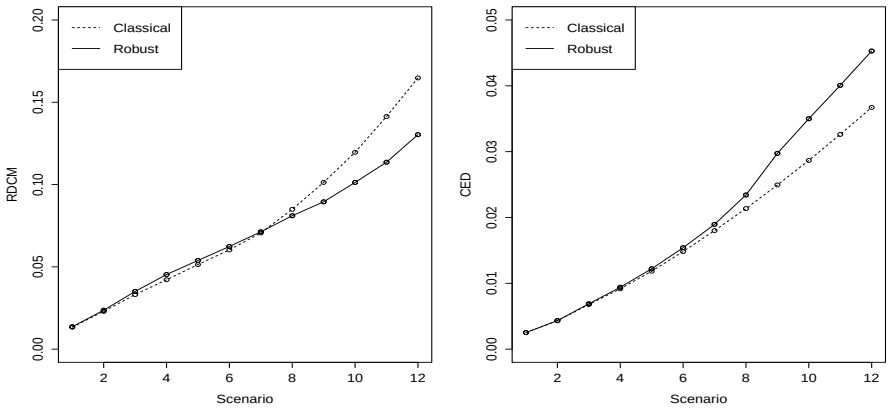


Figure 6: Measures of distortion: RDCM-relative difference of the covariance matrix (left); CED-compositional error deviation (right).

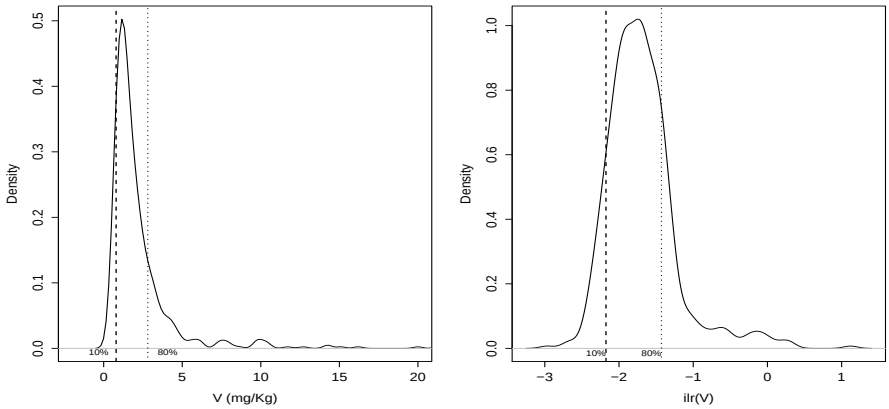


Figure 7: Density function of vanadium in the original space (left) and its corresponding variable in the ilr-transformed space (right). Vertical lines respectively represent the 10th and 80th percentiles of the distribution.

Figure 7 shows the distribution of the element vanadium in the original space, in parts per million, as well as its corresponding ilr-transformed variable as obtained from equation (1). For each data set, the $\mathbf{X}_i^*$ values below the corresponding quantile, from 5% to 95%, were forced to zero. As a result, in each case, the observed values of V used in the log-ratio regression models were those values that were above the corresponding percentiles. For example, in Figure 7, we illustrated the detection limits for the 10th and 80th quantiles of the distribution.

After applying the methods CI and RI to each data set $\mathbf{X}_i^*$, $i = 1, \ldots, 16$, the RDCM and CED were calculated. Figure 8 (left) shows that for RDCM, the robust method exhibited better performance for percentiles below 0.5, which corresponds to 1.55mg/kg of vanadium. When the percentile was higher than the 50th percentile, the classical method produced slightly smaller values of RDCM. This was also due to the robustness properties of RI, for which a high number of VBDL resulted in a nearly exact fit (Maronna et al., 2006). However, for very large percentiles where almost no information on vanadium remains, both methods exhibited poor performance. The results for the CED in Figure 8 (right) revealed similar results. RI outperformed CI for percentiles up to approximately 0.5.

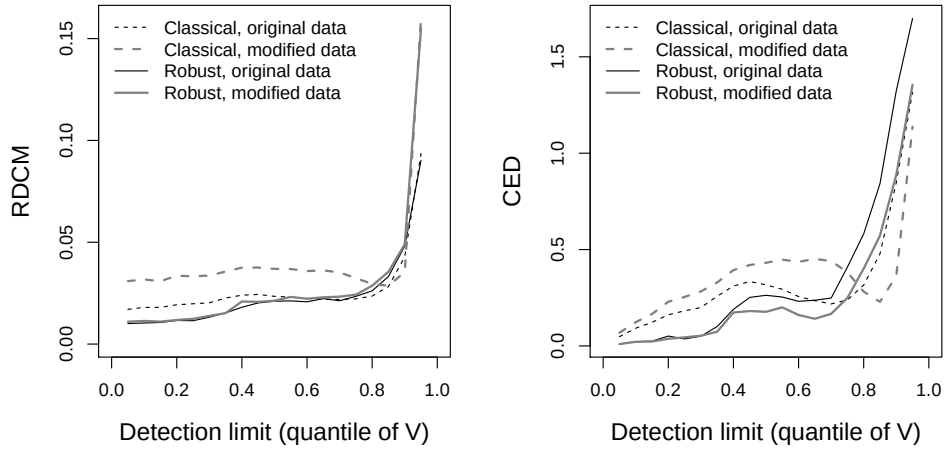
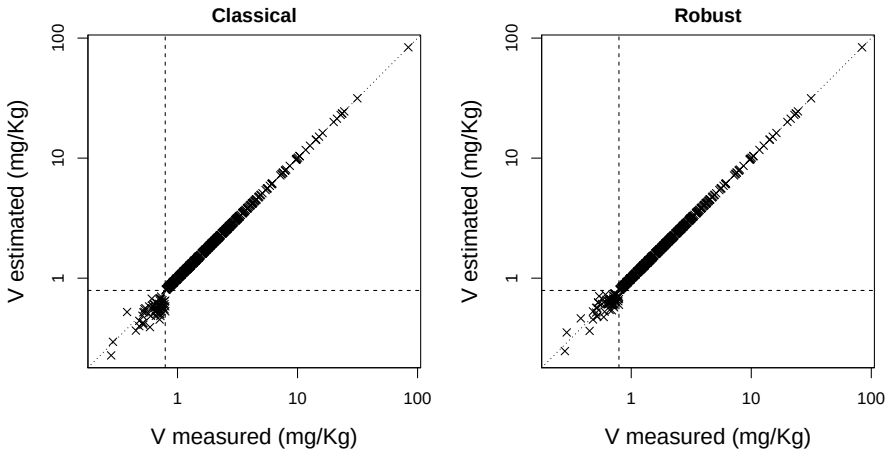


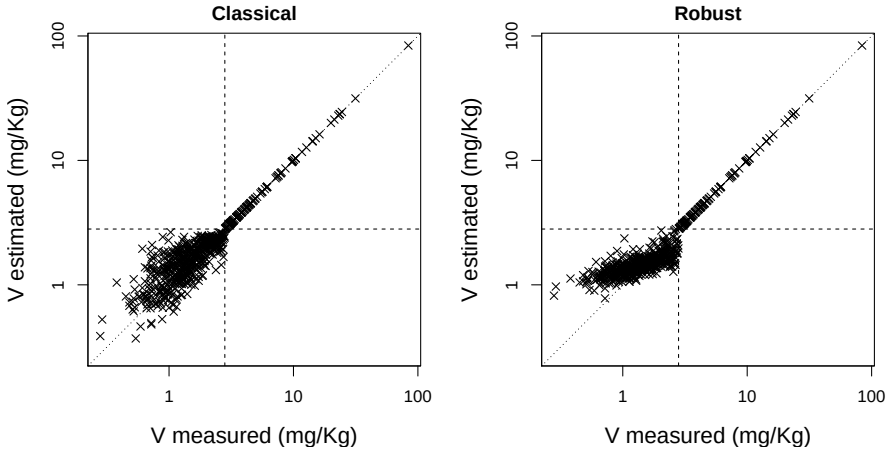
Figure 8: Relative difference of the covariance matrix (left) and compositional error deviation (right) for the classical (dashed lines) and the robust method (solid lines) applied to each detection limit (sample percentile) for the chemical element vanadium. The lines for "modified data" correspond to the outlier-effect experiment.

To gain more insight into the quality of the estimation, Figure 9 shows the original values of vanadium (in mg/kg) versus the estimated values from both imputation methods. We selected the 10th and 80th percentiles to compare the behaviour of the two methods above and below the 50th percentile. Figure 9 (A) shows the results for the 10th percentile. The RI method (right) yielded better estimates because the CI method (left) tended to underestimate the original values: note that most of the points were below the diagonal. This fact was consistent with a smaller CED (Figure 8) and with the idea that the extreme values, i.e., the smallest percentiles, did not affect the robust method but led to underestimation with the classical method. However, when a large sample percentile was considered, e.g., the 80th percentile, as shown in Figure 9(B), the robust method was applied to sparse information and produced values that overestimate the original data, although the corresponding data cloud was narrower than that obtained with the classical method. This effect explained the behaviour illustrated in Figure 6 (right): when the scenario included a large number of VBDL, the distance between the original samples and the imputed samples tended to be larger.

The data set from the *moss* layer does not include extreme outliers (Reimann et al., 2008). As a consequence, the different results obtained from the CI and RI methods were mostly due to the properties of the different regression techniques applied in equations (5) and (10), respectively. To illustrate the superior



(A) Sample percentile 10%



(B) Sample percentile 80%

Figure 9: Plot of measured versus estimated values for vanadium when the DL is equal to: (A) 10th sample percentile and (B) 80th sample percentile. Plots at the left correspond to the classical method; plots at the right correspond to the robust method.

performance of the robust method, artificial outliers were included in a *modified* data set by multiplying the upper 10th percentile of vanadium values 10% of vanadium by a factor of 2. Note that this was still a very mild contamination of the data set, making the ilr-transformed values of vanadium a skewed slightly to the right. Once again, both the CI and RI methods were applied to the modified data set, where rounded zeros were forced using the same procedure as before. In Figure 8 (left), the effect of the presence of outliers was clearly visible for the classical method. Robust imputation was not affected by the outliers. The CED in Figure 8 (right) illustrated an effect for the RI method as well, but the CI method was more sensitive to contamination. We obtained very similar results (not shown here) when only the maximum value of vanadium was multiplied by a factor of 100, i.e., when just one outlier was included. Of course, as the percentage of contamination was increased, the classical imputation results were more strongly affected.

6. Conclusions and remarks

When applying multivariate statistical methods, such as cluster analysis, multidimensional scaling, discriminant analysis, or regression analysis, to real data sets, it is necessary to have a complete data set available. In practice, however, many data sets are reported with values below a detection limit or with zeros. This happens particularly with compositional data, consisting of, for example, concentrations of chemical elements. In this paper, an approach is suggested to estimating rounded zeros in compositional data. We have proven that imputation techniques based on classical censored regression are equivalent for both ilr and alr transformations. Even in those situations where compositional data sets do not include outliers, robust techniques are recommended to minimise the effect of small imputed values in the treatment of rounded zeros. Robust methods are applied following an ilr transformation of the data because the alr transformation does not provide invariance of the distances under permutations (changes in the divisor of the alr transformation) and because it is not an isometric mapping.

Because the applied robust methods are affine equivariant, the use of the ilr transformation becomes appropriate in the rounded zeros problem. Combining this transformation with the censored regression model and robust techniques, we have introduced a procedure that, in general, improves the estimation of the variability and produces minor distortion in the covariance structure of the data. Only in those situations where the data set has a large number of rounded zeros does the classical approach seem to have an advantage. In such scenarios, robust methods are applied to data from the core of the distribution, and the classical approach, which works with all of the available information, could provide better estimations. However, for those more usual scenarios where the percentage of rounded zeros is lower than 30%, robust techniques are helpful because they are more appropriate when the censored values are in the extreme zone of the tail of the distribution.

The R code (R development core team, 2008) of our proposed procedure is available in the package *robCompositions* at the Comprehensive R Archive Network (see http://cran.r-project.org/).

Acknowledgments

This research was supported by the Ministerio de Ciencia e Innovación under the project "CODA-RSS" Ref. MTM2009-13272; by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424; by the Secretaría General de Universidades del Ministerio de Educación Programa "Salvador de Madariaga" Ref: PR2010-0177; and by the Scottish Government.

Appendix A

Let $\mathbf{Y}$ be an $n \times 1$ vector of observed values of a response variable y. Let $\mathbf{X}$ be an $n \times p$ data matrix with the first column equal to a vector $\mathbf{1}_n$ of ones, and let the rest of the columns correspond to the values of $p-1$ variables $x_1, x_2, \ldots, x_{p-1}$. The multiple linear regression (MLR) model is $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + error$, where $\boldsymbol{\beta}$ is the column vector of coefficients and $error$ is the residual term, which is usually assumed to be normally distributed. It is well know that, in any Euclidean space, the estimated regression coefficients

according to the least-squares model are $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^t \cdot \mathbf{Y}$. The estimated values of the response variable are $\widehat{\mathbf{Y}} = \mathbf{X} \cdot \widehat{\boldsymbol{\beta}}$, which can also be written as $\widehat{\mathbf{Y}} = \mathbf{H} \cdot \mathbf{Y}$, where $\mathbf{H}$ is the well-known projection matrix $\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^t$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ be a composition from the simplex S^D. Let $\mathbf{s}_1 = (x_1, x_D)$ and $\mathbf{s}_2 = (x_2, x_3, \ldots, x_D)$ be two subcompositions from $\mathbf{x}$. Let $alr(\mathbf{s}_1) = \ln \frac{x_1}{x_D}$ and $alr(\mathbf{s}_2) = (\ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D})$ be the *alr*-transformed vectors of $\mathbf{s}_1$ and $\mathbf{s}_2$. Consider $\mathbf{C}_{2D}$ to be the matrix $(D-2) \times (D-2)$ such that $ilr(\mathbf{s}_2) = alr(\mathbf{s}_2) \cdot \mathbf{C}_{2D}$ (e.g., Barceló-Vidal et al., 2011). The $D-2$ columns of the matrix $\mathbf{C}_{2D}$ are linear combinations of the $D-2$ vectors that form an orthonormal basis in the *ilr*-transformed space. We can assume that the *ilr* basis is such that each log-ratio is one part x_k over the geometric mean of the following parts $x_{(k+1)}, \ldots, x_D$ in the composition.

Hereafter, the subindex (i) and (a), respectively, stand for ilr and alr transformations. The MLR equations $\mathbf{Y}_{(i)} = \mathbf{X}_{(i)} \cdot \boldsymbol{\beta}_{(i)} + error_{(i)}$ and $\mathbf{Y}_{(a)} = \mathbf{X}_{(a)} \cdot \boldsymbol{\beta}_{(a)} + error_{(a)}$ thus denote the MLR model in each respective transformed space. Note that in our case, we set $\mathbf{Y}_{(a)} = alr(\mathbf{s}_1)$ and $\mathbf{X}_{(a)} = [\mathbf{1}_n, alr(\mathbf{s}_2)]$. We also set $\mathbf{X}_{(i)} = [\mathbf{1}_n, ilr(\mathbf{s}_2)]$. We can consider $\mathbf{Y}_{(i)}$ based on the log-ratio between the first part x_1 and the geometric mean of the parts $x_2, x_3, \ldots, x_D$. Let $\mathbf{M}_{2D}$ be the matrix $(D-1) \times (D-1)$ formed by the first row and column of an identity matrix, and let the other entries be equal to the matrix $\mathbf{C}_{2D}$:

$$\mathbf{M}_{2D} = \begin{pmatrix} 1 & 0 \ldots 0 \\ 0 & \\ \vdots & [\mathbf{C}_{2D}] \\ 0 & \end{pmatrix}.$$

It follows that $\mathbf{X}_{(i)} = \mathbf{X}_{(a)} \cdot \mathbf{M}_{2D}$ and that

$$\mathbf{M}_{2D}^{-1} = \begin{pmatrix} 1 & 0 \ldots 0 \\ 0 & \\ \vdots & [\mathbf{C}_{2D}]^{-1} \\ 0 & \end{pmatrix}.$$

The inverse relation from the log-ratio matrices can be easily obtained.

In addition, it holds that

$$\mathbf{H}_{(i)} = \mathbf{X}_{(i)} \cdot (\mathbf{X}_{(i)}^t \cdot \mathbf{X}_{(i)})^{-1} \cdot \mathbf{X}_{(i)}^t \rightarrow \mathbf{H}_{(i)} = \mathbf{X}_{(a)} \cdot \mathbf{M}_{2D} \cdot (\mathbf{M}_{2D}^t \cdot \mathbf{X}_{(a)}^t \cdot \mathbf{X}_{(a)} \cdot \mathbf{M}_{2D})^{-1} \cdot \mathbf{M}_{2D}^t \cdot \mathbf{X}_{(a)}^t$$

$$\rightarrow \mathbf{H}_{(i)} = \mathbf{X}_{(a)} \cdot (\mathbf{X}_{(a)}^t \cdot \mathbf{X}_{(a)})^{-1} \cdot \mathbf{X}_{(a)}^t = \mathbf{H}_{(a)}.$$

In other words, both projection matrices are equal; therefore, $\mathbf{H}_{(i)} = \mathbf{H}_{(a)} = \mathbf{H}$; $\widehat{\mathbf{Y}}_{(a)} = \mathbf{H} \cdot \mathbf{Y}_{(a)}$ and $\widehat{\mathbf{Y}}_{(i)} = \mathbf{H} \cdot \mathbf{Y}_{(i)}$.

Note the following three important properties of this matrix H:

- $\mathbf{H}^t = \mathbf{H}$,
- $\mathbf{H}^2 = \mathbf{H}$ and
- $\mathbf{H} \cdot \mathbf{X}_{(a)} = \mathbf{X}_{(a)}$ and $\mathbf{H} \cdot \mathbf{X}_{(i)} = \mathbf{X}_{(i)}$.

For our proof, the most important property is the last of these three. The above results are true in the general case of a change of basis in a Euclidean space. In other words, the above properties are not exclusive for our case of log-ratio regressions.

Once the alr-MLR and ilr-MLR models have been obtained, it should be determined whether the estimates from the two models are the same. Let $\mathbf{C}$ be the matrix that verifies $ilr(\mathbf{x}) = alr(\mathbf{x}) \cdot \mathbf{C}$, i.e., for the full composition. Let $\mathbf{M}$ be the matrix resulting from including in the second column and row of the matrix $\mathbf{C}$ a second column and row from an identity matrix as follows:

$$\mathbf{M} = \begin{pmatrix} c_{11} & 0 & c_{12} & \dots & c_{1(D-1)} \\ 0 & 1 & 0 & \dots & 0 \\ c_{21} & 0 & c_{22} & \dots & c_{2(D-1)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ c_{(D-1)1} & 0 & c_{(D-1)2} & \dots & c_{(D-1)(D-1)} \end{pmatrix}.$$

It is easily shown that $(\mathbf{Y}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{Y}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M}$.

Let $(\widehat{\mathbf{Y}}_{(a)}, \mathbf{X}_{(a)})$ and $(\widehat{\mathbf{Y}}_{(i)}, \mathbf{X}_{(i)})$ be the results in the transformed space produced by the two log-ratio MLR models. Therefore, it must be demonstrated that the equality $(\widehat{\mathbf{Y}}_{(i)}, \mathbf{X}_{(i)}) = (\widehat{\mathbf{Y}}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M}$ holds as follows:

$$(\widehat{\mathbf{Y}}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M} = (\mathbf{H} \cdot \mathbf{Y}_{(a)}, \mathbf{H} \cdot \mathbf{X}_{(a)}) \cdot \mathbf{M} = \mathbf{H} \cdot (\mathbf{Y}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M} =$$
$$= \mathbf{H} \cdot (\mathbf{Y}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{H} \cdot \mathbf{Y}_{(i)}, \mathbf{H} \cdot \mathbf{X}_{(i)}) = (\widehat{\mathbf{Y}}_{(i)}, \mathbf{X}_{(i)}),$$

where the properties $\mathbf{H}_{(i)} = \mathbf{H}_{(a)} = \mathbf{H}$, $\mathbf{H} \cdot \mathbf{X} = \mathbf{X}$ and $\widehat{\mathbf{Y}} = \mathbf{H} \cdot \mathbf{Y}$ have been applied.

Appendix B

As in Appendix A, the subindices (i) and (a) stand for ilr and alr transformations, respectively. We denote by $\mathbf{Y}$ the variable that includes unknown values and by $\mathbf{X}$ the observed variables. Without loss of generality, we can assume that the unknown values are in the first part x_1 and the observed variables form the rest $\mathbf{x}_{-1} = (x_2, \dots, x_D)$. In other words, in our case, we consider $\mathbf{Y}_{(a)} = \ln \frac{x_1}{x_D}$ and $\mathbf{X}_{(a)} = [\mathbf{1}_n, alr(\mathbf{x}_{-1})]$; we take the corresponding $\mathbf{Y}_{(i)}$ and $\mathbf{X}_{(i)}$ by applying the ilr-transformation from equation (1). In these terms, the conditional expected value-with a log-ratio-transformed detection limit $\boldsymbol{\psi}$-in our proposed algorithm is as follows:

$$E[\mathbf{Y}|\mathbf{X}, \mathbf{Y} < \boldsymbol{\psi}], \tag{13}$$

where each unknown value y_k in the variable $\mathbf{Y}$ is calculated using a censored regression in both log-ratio transformed spaces, applying the following expressions:

$$\hat{y}_{(a)k} = \mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)} - \hat{\sigma}_{(a)} \frac{\phi\left(\frac{\psi_{(a)k} - \mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}\right)}{\Phi\left(\frac{\psi_{(a)k} - \mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}\right)} \quad \text{and} \quad \hat{y}_{(i)k} = \mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)} - \hat{\sigma}_{(i)} \frac{\phi\left(\frac{\psi_{(i)k} - \mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}}\right)}{\Phi\left(\frac{\psi_{(i)k} - \mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}}\right)}. \tag{14}$$

The proof analyses whether the estimates obtained, $(\widehat{\mathbf{Y}}_{(a)}, \mathbf{X}_{(a)})$ and $(\widehat{\mathbf{Y}}_{(i)}, \mathbf{X}_{(i)})$ are equivalent. In other words, it must be demonstrated that $(\hat{y}_{(i)k}, \mathbf{X}_{(i)k}) = (\hat{y}_{(a)k}, \mathbf{X}_{(a)k}) \cdot \mathbf{M}$ holds, where the matrix $\mathbf{M}$ is the matrix introduced in Appendix A. To accomplish this proof, it is necessary to describe some properties of $\mathbf{M}$. This matrix is composed of the matrix $\mathbf{C}$ that verifies $ilr(\mathbf{x}) = alr(\mathbf{x}) \cdot \mathbf{C}$ for any composition $\mathbf{x} \in S^D$ (e.g., Barceló-Vidal et al., 2011). This matrix is the product of two specific matrices $\mathbf{F}$ and $\mathbf{V}$: $\mathbf{C} = (\mathbf{F} \cdot \mathbf{V})^t$. Here, $\mathbf{F} = [\mathbf{I}_{D-1} : -\mathbf{1}_{D-1}]$, where $\mathbf{I}$ is the identity matrix of order $D-1$. The columns of the matrix $\mathbf{V}$, with order $D \times (D-1)$, are the vectors of the ilr-orthonormal basis. For the basis described by equation (1), applied in our algorithm, the element (v_{ij}) in matrix $\mathbf{V}$ is equal to

$$v_{ij} = \begin{cases} 0 & \text{for } i < j; \\ \sqrt{\frac{D-j}{D-j+1}} & \text{for } i = j; \\ -\frac{1}{\sqrt{(D-j)(D-j+1)}} & \text{for } i > j. \end{cases}$$

Using the above expressions, it is easy to state that for any matrix $\mathbf{M}$, it holds that $m_{ij} = 0$ for $i < j$. This property will be crucial for a proof of the relation between the two estimates. For example, for $D = 6$, one obtains

$$\mathbf{M} = \begin{pmatrix} 1.095 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.224 & 0 & 1.118 & 0 & 0 & 0 \\ 0.289 & 0 & 0.289 & 1.155 & 0 & 0 \\ 0.408 & 0 & 0.408 & 0.408 & 1.225 & 0 \\ 0.707 & 0 & 0.707 & 0.707 & 0.707 & 1.414 \end{pmatrix}.$$

Upon examining the expressions in (14), we conclude that it is only necessary to analyse the second term in the subtraction because the first one, $\mathbf{X}_k \cdot \hat{\boldsymbol{\beta}}$, is the MLR estimate that is analysed in Appendix A. To find the relation between the conditional variances $\hat{\sigma}_{(a)}$ and $\hat{\sigma}_{(i)}$, it is necessary to consider their corresponding expressions from the regression models

$$(\mathbf{Y}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)})^t \cdot (\mathbf{Y}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}) \quad \text{and} \quad (\mathbf{Y}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)})^t \cdot (\mathbf{Y}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}), \tag{15}$$

where $\mathbf{Y}$ stands for the vector $n_o \times 1$ composed by the n_o observed values in the variable. From Appendix A, the observed part in the data set and the estimates from MLR verify that

$$(\mathbf{Y}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{Y}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M} \quad \text{and} \quad (\mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{X}_{(i)}) = (\mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M}. \tag{16}$$

Subtracting both expressions in (16) one obtains

$$(\mathbf{Y}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) = (\mathbf{Y}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0}) \cdot \mathbf{M},$$

where "$\mathbf{0}$" is a matrix of order $n_o \times (D-1)$ in which all entries are equal to zero. Therefore, it holds that

$$(\mathbf{Y}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0})^t \cdot (\mathbf{Y}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) = \mathbf{M}^t \cdot (\mathbf{Y}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0})^t \cdot (\mathbf{Y}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0}) \cdot \mathbf{M}.$$

From the above identity and the special pattern of the matrix $\mathbf{M}$, it is easy to prove that $\hat{\sigma}_{(i)} = m_{11}\hat{\sigma}_{(a)}$, where m_{11} is the first element in the diagonal of $\mathbf{M}$. This expression relates the conditional variances of the two censored regressions. Finally, it is necessary to analyse the relation between the elements

$$\boldsymbol{\psi}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)} \quad \text{and} \quad \boldsymbol{\psi}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \tag{17}$$

that are inside the normal density and distribution functions. Here, $\boldsymbol{\psi}$ stands for the vector $n_u \times 1$ of the log-ratio transformed detection limits of the n_u unknown values. The matrix $\mathbf{X}$ is the observed part of the data set. From the construction of this matrix, it is simple to show that

$$(\boldsymbol{\psi}_{(i)}, \mathbf{X}_{(i)}) = (\boldsymbol{\psi}_{(a)}, \mathbf{X}_{(a)}) \cdot \mathbf{M}. \tag{18}$$

Subtracting the above expression and the expression on the right in (16), one obtains

$$(\boldsymbol{\psi}_{(i)} - \mathbf{X}_{(i)} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{0}) = (\boldsymbol{\psi}_{(a)} - \mathbf{X}_{(a)} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{0}) \cdot \mathbf{M},$$

and therefore, $(\boldsymbol{\psi}_{(i)k} - \mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)}) = m_{11}(\boldsymbol{\psi}_{(a)k} - \mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)})$ for any $k = 1, \ldots, n_u$. Consequently, if the relation between the conditional variances is considered, it holds that

$$\frac{\boldsymbol{\psi}_{(i)k} - \mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)}}{\hat{\sigma}_{(i)}} = \frac{\boldsymbol{\psi}_{(a)k} - \mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}}{\hat{\sigma}_{(a)}}, \qquad k = 1, \ldots, n_u,$$

which implies an equal value in the normal density and distribution functions in the censored regressions.

Once the relations between all the elements in the censored regressions have been analysed, one can state the relation between the estimates. To simplify the following expressions, the ratio between both normal functions has been replaced by the expression $[\phi(\cdot)/\Phi(\cdot)]$ that takes the same value for both log-ratio censored regressions. Let $(\hat{y}_{(i)k}, \mathbf{X}_{(i)k})$ be the k-th estimates from the ilr-censored regression. It holds that

$$\begin{aligned}
(\hat{y}_{(i)k}, \mathbf{X}_{(i)k}) = & \left(\mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)} - \hat{\sigma}_{(i)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{X}_{(i)k}\right) = \\
= & \left(\mathbf{X}_{(i)k} \cdot \hat{\boldsymbol{\beta}}_{(i)}, \mathbf{X}_{(i)k}\right) \quad - \quad \left(\hat{\sigma}_{(i)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{0}\right) \\
= & \left(\mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{X}_{(a)k}\right) \cdot \mathbf{M} \quad - \quad \left(m_{11}\hat{\sigma}_{(a)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{0}\right) \\
= & \left(\mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{X}_{(a)k}\right) \cdot \mathbf{M} \quad - \quad \left(\hat{\sigma}_{(a)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{0}\right) \cdot m_{11} \\
= & \left(\mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)}, \mathbf{X}_{(a)k}\right) \cdot \mathbf{M} \quad - \quad \left(\hat{\sigma}_{(a)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{0}\right) \cdot \mathbf{M} \\
= & \left(\mathbf{X}_{(a)k} \cdot \hat{\boldsymbol{\beta}}_{(a)} - \hat{\sigma}_{(a)}[\phi(\cdot)/\Phi(\cdot)], \mathbf{X}_{(a)k}\right) \cdot \mathbf{M} \\
= & (\hat{y}_{(a)k}, \mathbf{X}_{(a)k}) \cdot \mathbf{M}.
\end{aligned}$$

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London (UK). 416 p.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V., 2000. Logratio analysis and compositional distance. Mathematical Geology 32 (3), 271-275.

Amemiya, T., 1984. Tobit models: a survey. Journal of Econometrics 24, 3-61.

Barceló-Vidal, C., Aguilar, L. and Martín-Fernández, J.A., 2011. Compositional VARIMA Time Series, Ch. 7. In: Pawlowsky-Glahn and Buccianti (2011), pp. 87-103.

Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (2006) (eds). Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, Special Publications 264.

Daunis-i-Estadella, J., Barceló-Vidal, C., Buccianti, A., 2006. Exploratory compositional data analysis. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 161-174.

Daunis-i-Estadella, J., Martín-Fernández, J.A. (eds), 2008. Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop. Universitat de Girona, ISBN 84-8458-272-4, http://ima.udg.es/Activitats/CoDaWork08/. May 27-30, CD-ROM.

Egozcue, J.J., 2009. Reply to "On the Harker Variation Diagrams; " by J.A. Cortés. Mathematical Geosciences 41 (7), 829-834.

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

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 145-160.

Egozcue, J. J. and Pawlowsky-Glahn, V., 2011. Basic concepts and procedures, Ch. 2. In: Pawlowsky-Glahn and Buccianti (2011), pp.12-28.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35 (3), 279-300.

Egozcue, J.J., Tolosana-Delgado, R., Ortego, M.I. (eds), 2011. Proceedings of CODAWORK'11, The 4th Compositional Data Analysis Workshop. Sant Feliu De Guxols, ISBN: 978-84-87867-76-7 (electronic publication). May 10-13.

Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Mathematical Geosciences 40 (3), 233-248.

Filzmoser, P. and Hron, K., 2011. Robust statistical analysis, Ch. 5. In: Pawlowsky-Glahn and Buccianti (2011), pp. 59-72.

Filzmoser, P., Hron, K., Reimann, C., 2009. Principal component analysis for compositional data with outliers. Environmetrics 20 (6), 621-632.

Fišerová, E., Hron, K., 2011. On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43 (4), 455-468.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis 54 (12), 3095-3107.

Huber, P.J., 1981. Robust Statistics. John Wiley, New York.

Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, London, fifth edition.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New Jersey, second edition.

Maronna R., Martin R.D., Yohai V.J., 2006. Robust Statistics: Theory and Methods. John Wiley, New York (USA). 436 p.

Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology 35 (3), 253-278.

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with zeros, Ch. 4. In Pawlowsky-Glahn and Buccianti (2011), pp. 47-62.

Martín-Fernández, J.A., Thió-Henestrosa, S., 2006. Rounded zeros: some practical aspects for compositional data. In Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264, 191-201.

Mateu-Figueras, G., Barceló-Vidal, C. (eds), 2005. Proceedings of CODAWORK'05, The 2nd Compositional Data Analysis Workshop. Universitat de Girona, ISBN 84-8458-222-1, http://ima.udg.es/Activitats/CoDaWork05/. October 19-21, CD-ROM.

Mateu-Figueras, G., Pawlowsky-Glahn, V., 2008. A critical approach to probability laws in geochemistry. Mathematical Geosciences 40 (5), 489-502.

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences 34 (8), 902-917.

Palarea-Albaladejo, J., Martín-Fernández, J.A., Gómez-García, J., 2007. A parametric approach for dealing with compositional rounded zeros. Mathematical Geology 39 (7), 625-645.

Pawlowsky-Glahn, V., Buccianti, A. (Eds.), 2011. Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd., Chichester (UK). 378 p.

Pawlowsky-Glahn, V., Egozcue, J.J., 2002. BLU estimators and compositional data. Mathematical Geology 34 (3), 259–274.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60, 489–502.

Reimann, C., Äyräs, M., Chekushin, V.A., Bogatyrev, I., Boyd, R., de Caritat, P., Dutter, R., Finne, T.E., Halleraker, J.H., Jæger, Ø., Kashulina, G., Niskavaara, H., Lehto, O., Pavlov, V., Räisänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. NGU-GTK-CKE Special Publication, Trondheim (Norway). 745p.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley, Chichester.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J., Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Mining Knowl. Disc. 12, 29-45.

R development core team, 2008, R: A language and environment for statistical computing: Vienna, `http://www.r-project.org`.

Salibian-Barrera, M., Yohai, V.J., 2006. A fast algorithm for S-regression estimates. Journal of Computational and Graphical Statistics, 15, 414–427.

Seber, G. A. F., 2008. A Matrix Handbook for Statisticians. John Wiley & Sons, Inc., Hoboken, New Jersey (USA). 559 p.

Thió-Henestrosa, S., Martín-Fernández, J.A. (eds), 2003. Proceedings of CODAWORK'03, The 1st Compositional Data Analysis Workshop. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/. October 15-17, CD-ROM.

Tolosana-Delgado, R., van den Boogaart, K.G., Pawlowsky-Glahn, V., 2011. Geostatistics for compositions, Ch. 6. In Pawlowsky-Glahn and Buccianti (2011), pp. 73-86.

van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2006. Concepts for handling zeros and missing values in compositional data. In: Proceedings of IAMG'06 - The XI annual conference of the International Association for Mathematical Geology. University of Liege, Belgium. CD-ROM.