Math Geol (2007) 39: 625–645
DOI 10.1007/s11004-007-9100-1

A Parametric Approach for Dealing with Compositional Rounded Zeros

Javier Palarea-Albaladejo · Josep A. Martín-Fernández · Juan Gómez-García

Received: 12 March 2006 / Accepted: 16 February 2007 / Published online: 14 September 2007

Abstract In this work, a parametric approach for replacing data below the detection limit, also known as rounded zeros, in compositional data sets is proposed. Compositional rounded zeros correspond to small proportions of some whole that cannot be reliably detected by the analytical instruments under given operating conditions. This kind of zeros appear frequently in the data collection process in geosciences. They must be treated in an adequate way before some multivariate analysis can be applied. Our procedure results from a modification of the Expectation-Maximization (EM) algorithm and is based on the additive log-ratio transformation. Its coherence with the nature of compositional data and with basic operations in the simplex sample space is checked. Using real data sets, we find that this approach improves other parametric and non-parametric techniques for compositional rounded zeros.

Keywords Additive log-ratio transformation · Compositional data · EM algorithm · Simplex

J. Palarea-Albaladejo (✉)
Departamento de Informática de Sistemas, Universidad Católica San Antonio,
Campus de Los Jerónimos, 30107 Murcia, Spain
e-mail: jpalarea@pdi.ucam.edu

J.A. Martín-Fernández
Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi,
17071 Girona, Spain
e-mail: josepantoni.martin@udg.es

J. Gómez-García
Departamento de Métodos Cuantitativos para la Economía, Universidad de Murcia,
Campus de Espinardo, 30100 Murcia, Spain
e-mail: jgomezg@um.es

Introduction

Compositional data refer to proportions of a whole. This type of data is present (Thió-Henestrosa and Martín-Fernández 2003; Mateu-Figueras and Barceló-Vidal 2005; Pawlowsky-Glahn 2005) in a great variety of practical problems associated with many disciplines: archeology (the chemical compositions of ceramics), environometrics (pollutant compositions), geology (major oxide compositions of rocks and sediment compositions), medicine (blood, urine and renal calculi compositions), economics (household budget patterns), and others. Formally, compositional data can be represented by a vector $\mathbf{x} = [x_1, \ldots, x_D]$ with D components or parts. Its natural sample space (Aitchison 1986) is the open simplex S^D defined as

$$S^D = \left\{ \mathbf{x} = [x_1, \ldots, x_D]\colon\ x_1 > 0, \ldots, x_D > 0; \sum_j x_j = c \right\}.$$

The value of c is irrelevant from the mathematical point of view, and is usually 1, 100, or 10^6 depending on the units of measurement. This sample space has a vector space structure induced by the operations of perturbation and power transformation. For any two compositions $\mathbf{x}$ and $\mathbf{x}'$, the perturbation operation $\mathbf{x} \oplus \mathbf{x}' = \mathrm{C}(x_1 x_1', \ldots, x_D x_D')$ is defined on $S^D \times S^D$ and the power transformation $\alpha \otimes \mathbf{x} = \mathrm{C}(x_1^\alpha, \ldots, x_D^\alpha)$ is defined on $R \times S^D$. Here, C is the closure operator defined by $\mathrm{C}(\mathbf{w}) = [w_1/\sum_j w_j, \ldots, w_D/\sum_j w_j]$ for $\mathbf{w} \in R_+^D$. Moreover, an inner product can be defined by

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \left[\frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{x_i'}{x_j'} \right]$$

conferring an Euclidean space structure to the simplex. From the statistical point of view, this structure is not irrelevant since the basic measures of central tendency, distance and variability must be coherent with the nature of the data. The above inner product induces a distance in S^D defined by

$$d_{\mathrm{a}}(\mathbf{x}, \mathbf{x}') = \left[\frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i'}{x_j'} \right)^2 \right]^{1/2} = \left[\sum_{i=1}^{D} \left(\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{x_i'}{g(\mathbf{x}')} \right)^2 \right]^{1/2}, \quad (1)$$

where $g(\mathbf{x}) = (x_1 \cdot \cdots \cdot x_D)^{1/D}$ is the geometric mean of the composition $\mathbf{x}$. This distance is known as Aitchison distance and is fully compatible with the vector space structure of the simplex (Aitchison et al. 2000).

Following Aitchison (1986), a proper analysis of compositions should be focused on the study of relative magnitudes, rather than absolute magnitudes, and so every statement about a composition can be stated in terms of ratios of components or parts. It is well known that log-ratios $\ln(x_i/x_j)$ are easier to handle mathematically than ratios. Using log-ratios, one can define transformations which provide a one-to-one correspondence between compositional vectors and log-ratio vectors in real space, opening up all available multivariate techniques for real data. Consequently, the log-ratio methodology has become a standard approach for the statistical analysis

of compositional data. The main log-ratio transformation used to date in parametric studies is the additive log-ratio (alr) transformation defined by

$$\mathrm{alr}(\mathbf{x}) = \left[\ln\frac{x_1}{x_D}, \ldots, \ln\frac{x_{D-1}}{x_D}\right]. \tag{2}$$

The centered log-ratio (clr) transformation is defined by $\mathrm{clr}(\mathbf{x}) = [\ln(x_1/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x}))]$. This transformation is mainly applied in non-parametric contexts since from the clr transformation we can define the log-ratio distance (1) as the Euclidean distance between the clr-transformed vectors.

One must proceed with caution when the alr transformation (2) is applied because it is asymmetric as regards the components. We must verify that our statistical technique is invariant in the face of permutations of the components. However, the main drawback of the alr transformation is the fact that it is not an isometric transformation. Consequently, angles and distances in the simplex with the Aitchison metric cannot be associated to angles and distances in the real space with the Euclidean metric. On the other hand, the clr transformation is symmetric and isometric, but it has the weakness that the image of S^D remains constrained to a subspace and the covariance matrix of the clr-transformed data set is singular. To overcome these shortcomings, Egozcue et al. (2003) proposed the isometric log-ratio (ilr) transformation defined by $\mathrm{ilr}(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1\rangle, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1}\rangle]$, where $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}$ is an orthonormal basis in S^D obtained by the Gram–Schmidt procedure in combination with the inner product defined above. This transformation is isometric and the ilr-transformed data are represented in orthogonal axes, as we usually do in real spaces. The main problem here consists of determining which one is the most appropriate orthonormal basis in a specific problem. Furthermore, the ilr-coordinates are usually difficult to interpret. In Egozcue et al. (2003) the authors obtain an explicit expression of the ilr transformation for a particular orthonormal basis

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{y} = [y_1, \ldots, y_{D-1}] \in R^{D-1}, \quad \text{where } y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i}\right).$$

As regards our stated goal, neither transformations nor measures based on log-ratios can be applied if a compositional vector has parts with zero values. The presence of rounded or trace zeros is very common when a sample of compositional data is recorded. Frequently, accuracy limitations of the analytical instruments of measurement or some other physical, chemical or artificial effects prevent us from detecting small concentrations when collecting data about compositions of different elements. These small concentrations are then registered as false zero concentrations. In practice, the threshold value is usually set just below the smallest observed concentration that has been distinguished. Note that a zero value corresponding to a component which is truly zero, called essential or structural zero, must be treated in a different manner (Aitchison and Kay 2004; Bacon-Shone 2003). For rounded zeros, imputation techniques can be used because, in essence, a rounded zero can be considered as a missing value (Martín-Fernández and Thió-Henestrosa 2006; Martín-Fernández et al. 2003a). It is very important to emphasize that the decision to apply specific techniques for missing values when dealing with rounded zeros is quite

independent of the statistical methodology (Euclidean, log-ratio, and so on) selected. In other words, even when the log-ratio option is not selected, the compositional rounded zeros problem exists, because the missing values must be dealt with.

When the number of rounded zeros is not large, a non-parametric method of imputation works quite well (Martín-Fernández and Thió-Henestrosa 2006; Martín-Fernández et al. 2003a). For a data set with a large number of rounded zeros, parametric methods are recommended. In this respect, we propose a parametric method based on the alr transformation (2) which simultaneously takes into account the special nature of compositional data and rounded zeros. The reasons why we select the alr transformation rather than clr or ilr transformations are simple. As it will be seen with the alr transformation, the information about the detection limit can be easily incorporated to the alr-transformed data model. Only one component free of zeros is required. Note that this is not possible using the other transformations since their expressions include a product involving all the components. As regards the lack of isometry, this is not a question of special relevance for the procedure since it only uses the alr transformation as a device to replace the zeros; the data analysis and interpretation per se is not the purpose. Furthermore, following Aitchison (1986), the consistency of results is guaranteed when inference is based on the likelihood of the additive logistic normal distribution.

Our proposal is a modification of the well-known EM algorithm, an iterative procedure for maximum-likelihood estimation from incomplete real data sets. We apply the modified EM to the alr-transformed data set and then transform back the completed data set to the simplex in order to obtain a completed compositional data set without rounded zeros. In this work, we first review some strategies for replacing rounded zero values, both from a parametric and non-parametric perspective. Next, we introduce our approach and present its properties, before proceeding with some real data sets, comparing the performance of our approach with other methods. Finally, the main conclusions are presented and future developments are proposed.

Rounded Zeros Replacement Strategies

In applying an imputation or replacement technique, care must be taken not to distort the general structure of the data. In particular, the covariance structure and the metric properties of the data set should be preserved so that further analysis on subpopulations does not become misleading. Note that this last sentence expresses the key of the replacement techniques for rounded zeros in compositional data. Compositional data are formed by continuous variables, whose scale of measurement is the ratio scale, while their main operations are perturbation, power transformation and subcomposition. Consequently, at least for compositional data, the replacement strategies must be coherent with all these basic aspects, preserving the covariance structure and metrics induced by the log-ratio methodology.

From a non-parametric point of view, Aitchison (1986, p. 269) proposed an additive replacement method for replacing the rounded or trace zeros. Fry et al. (2000) and, independently, Martín-Fernández et al. (2000) showed that the additive replacement does not preserve the ratio of non-zero values and, consequently, dis-

torts the covariance structure of the data set. Martín-Fernández et al. (2003a) proposed and analysed in depth a multiplicative replacement. Consider a composition $\mathbf{x} = [x_1, \ldots, x_D] \in S^D$ containing rounded zeros. Using the multiplicative replacement, $\mathbf{x}$ can be replaced by a new composition $\mathbf{r} = [r_1, \ldots, r_D] \in S^D$ without zeros according to the following replacement rule

$$r_j = \begin{cases} \delta_j & \text{if } x_j = 0, \\ \left(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}\right) x_j & \text{if } x_j > 0, \end{cases} \tag{3}$$

where δ_j is a small value less than a given threshold γ_j for the component x_j. This approach is "natural" in the sense that it recovers the "true" composition if replacement values δ_j are identical to the missing values. In addition, it is coherent with the basic operations on the simplex, and the covariance structure of subcompositions without zeros is preserved. In contrast, since the multiplicative replacement (3) imputes exactly the same value in all the null values of the part, this replacement introduces artificial correlation between parts which have null values in the same rows of the data set. This effect is undesirable and can distort the results of the subsequent multivariate analysis when the number of null values in the data set is large, i.e. more than 10% (Sandford et al. 1993).

When the number of zeros in the data set is quite large, parametric methods of imputation are recommended. These methods take into account the information included in the covariance structure. As a consequence, the imputed values would be different for each row. As stated in Schafer (1997) or Little and Rubin (2002), the EM algorithm (Dempster et al. 1977) and the multiple imputation (MI) method (Rubin 1987) are the most reliable parametric methods for missing data in a real space and represents the current state of the art. For an illustrative simulation study see Gómez-García et al. (2006). Both techniques rely on fully parametric models for multivariate data, usually the normal distribution. When one of these methods is crudely applied to compositional data in the simplex, i.e. without previously log-ratio transforming the data, the structure of the data is seriously distorted. It is easy to show (Martín-Fernández et al. 2003b) how this strategy can replace the zeros by negative values or by values over the detection limit. In addition, a constant-sum constraint is not respected. Following Aitchison (1986), if one applies the alr transformation to the original compositional data, a multivariate normal model can be considered since the data are in a real space. This transformation methodology is applied in Buccianti and Rosso (1999). From an empirical point of view, the authors describe the performance of the EM algorithm and of its extension, the Sandford method (Sandford et al. 1993), applied to the alr-transformed data set.

When the MI method is applied to the alr-transformed data, it generates k values for each zero value. In multivariate contexts, these values can be simulated from the posterior predictive distribution of the missing part of the data set, given the observed part, using Markov Chain Monte Carlo (MCMC) algorithms. Next, the k completed data sets are transformed back to the simplex. Finally, using simple rules set out by Rubin (1987), we obtain a unique global estimation and its variance, combining the k estimated measures from each of the k completed data set. The main feature of the MI method is that it incorporates the uncertainly due to the presence of non-observed

data in the estimates. Following transformation methodology, Martín-Fernández et al. (2003b) apply MI via MCMC simulation to rounded zeros and describe its behaviour.

Another strategy using MI has been applied in the political science field. Honaker et al. (2002) deal with a special type of compositional data, electoral data from a multiparty voting system. These authors also considered the alr transformation (2) for the model. Nevertheless, zero votes in electoral data (when not all parties contest in every district) constitute politically crucial information, and so these zeros must be treated differently from our rounded zeros which are due to the limited measurement capabilities of the instruments. A set of assumptions, uniquely related to political science, are imposed for partly contested districts. In Honaker et al. (2002) a modification of the multiple imputation algorithm presented in King et al. (2001) is applied for the estimation. The authors adopt a rejection sampling/resampling procedure from the imputation model without constraints, checking whether the imputed value fits the constraint and discarding it and drawing another if the constraint is violated. We have implemented this sampling/resampling strategy for rounded zeros in compositional data without success. The results were not satisfactory due to the very low proportion of small values (values under the detection limit) generated by an imputation model which does not include constraints related to the detection limit. From the computational and time consumption point of view, this prevents the necessary imputed data sets from being obtained. Despite the above, several ideas can be extracted for future developments related to the rounded zeros problem.

A Modified EM Algorithm for the Compositional Rounded Zeros Problem

Some Comments about the EM Algorithm and Missing Data

The EM algorithm is an iterative procedure designed to find maximum-likelihood (ML) estimates in the context of parametric models where portions of a data set are not observed. Let $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ be an incomplete multivariate sample, where $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ denote, respectively, the observed and missing parts of $\mathbf{Y}$. Let θ be the unknown parameters of the probability distribution P for the complete data. Let $\ell(\theta \mid \mathbf{Y})$ denote the associated log-likelihood function. In the tth iteration of the EM algorithm, there are two steps:

E-step: given an estimate $\theta^{(t)}$ of θ, compute the conditional expectation $Q(\theta \mid \theta^{(t)})$, with

$$Q(\theta \mid \theta^{(t)}) = \int \ell(\theta \mid \mathbf{Y}) P\left[\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}; \theta^{(t)}\right] d\mathbf{Y}_{\text{mis}};$$

M-step: determine $\theta^{(t+1)}$ so that $\theta^{(t+1)} = \arg\max_{\theta} Q(\theta; \theta^{(t)})$.

Hence, the EM algorithm replaces direct maximization of the unknown complete data log-likelihood $\ell(\theta \mid \mathbf{Y})$ by successive maximizations of the conditional expectation $Q(\theta \mid \theta^{(t)})$, given $\mathbf{Y}_{\text{obs}}$ and the current estimate $\theta^{(t)}$ for θ. From an initial estimation $\theta^{(0)}$, the E- and M-steps are repeated alternatively, generating a sequence $\{\theta^{(t)}\}$ of

estimators. The EM has the basic property that each iteration increases $\ell(\theta \mid \mathbf{Y}_{\text{obs}})$, the observed data log-likelihood, i.e.

$$\ell\big(\theta^{(t+1)} \mid \mathbf{Y}_{\text{obs}}\big) \geq \ell\big(\theta^{(t)} \mid \mathbf{Y}_{\text{obs}}\big),$$

with equality if $Q(\theta^{(t+1)} \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)})$. A detailed account of convergence properties of the EM algorithm can be found in Dempster et al. (1977) and Wu (1983). Under suitable regularity conditions, $\{\theta^{(t)}\}$ converges to the ML estimator of θ. A frequently used practical criterion is to stop the iterations when the difference between two consecutive estimations of θ is sufficiently small. For more theoretical and practical aspects of the EM algorithm and its extensions, we refer the reader to the monograph of McLachlan and Krishnan (1997), and Little and Rubin (2002).

Almost all of the missing data methods used in statistical practice (both ad hoc procedures and principled ones) rely, at least implicitly, on an assumption called ignorability, that is, the analyst can ignore the mechanism which generates the missing data and consider the observed data likelihood as the relevant likelihood. Two ignorability situations are typically distinguished: MAR (missing at random), when the probability that one value is missing depends on the $\mathbf{Y}_{\text{obs}}$ part of the vector but not on $\mathbf{Y}_{\text{mis}}$; and MCAR (missing completely at random), when the missing data are a simple random sample of all data values, that is, the missingness does not depend on the data values. MCAR is a more restrictive, unrealistic and infrequent case of MAR. Only under MCAR can some ad hoc missing data methods provide proper inferences (Gómez-García et al. 2006). The common EM algorithm, the Sandford method and the MI method assume that the missingness mechanism is MAR and their development is based on observed data likelihood.

On the other hand, a missingness mechanism is called NMAR (not missing at random) when the probability that one value is missing depends on $\mathbf{Y}_{\text{mis}}$ part of the vector. This case represents the nonignorable situation and special models and methods are usually required. For continuous data, one group of nonignorable methods is based on models known as stochastic censoring or selection models (Heckman 1976; Amemiya 1984). Even so, in some empirical NMAR situations with rich information in the observed data, likelihood-based methods under the MAR hypothesis have been found to yield a more accurate inference than nonignorable modeling. In a compositional data setting, the rounded zeros problem is a particular NMAR case: data cannot be observed because their value is below the detection limit. This fact could explain why the common EM algorithm, the Sandford method and the MI method do not produce satisfactory results.

Modifying the E-Step to Take into Account the Detection Limit

Suppose that $\mathbf{y} = [y_1, \ldots, y_{D-1}]$ are the alr-transformed parts of the composition $\mathbf{x} = [x_1, \ldots, x_D]$. Assume that $\mathbf{y}$ is distributed according to a $(D-1)$-dimensional normal distribution with mean vector μ and covariance matrix Σ, i.e. $\mathbf{x}$ is distributed according to an additive logistic normal (aln) distribution (Aitchison 1986). As is well known, the complete-data log-likelihood for μ and Σ based on a sample $\mathbf{Y}$ of

size n is given by

$$\ell(\mu, \Sigma \mid \mathbf{Y}) = -n \ln(2\pi) - \frac{1}{2} n \ln|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{y}_i - \mu).$$

Since the normal model belongs to the regular exponential family, the EM algorithm is particularly simple to implement. Namely, the complete data log-likelihood is linear in the non-observed data. Hence, on the tth iteration of the EM algorithm, the computation of the conditional expected complete data likelihood in the E-step reduces to the computation of the conditional expected value of the missing part, $\mathrm{E}[\mathbf{Y}_{\mathrm{mis}} \mid \mathbf{Y}_{\mathrm{obs}}; \theta^{(t)}]$, given a parameter estimate $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$.

For our purpose, we consider a matrix $\Psi = (\psi_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, D-1$, where $\psi_{ij} = \ln(\gamma_j / x_{iD})$, being γ_j the detection limit for the component x_j and x_D a component without zeros. Note that Ψ contains the information about the detection limit in relation to the alr-transformed data. We modify the E-step to incorporate this information considering the conditional expectation $\mathrm{E}[\mathbf{Y}_{\mathrm{mis}} \mid \mathbf{Y}_{\mathrm{obs}}, \mathbf{Y}_{\mathrm{mis}} < \Psi; \theta^{(t)}]$ for the missing part. Then, on the tth iteration, the modified E-step replaces the values in $\mathbf{Y}$ by

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{if } y_{ij} \geq \psi_{ij}, \\ \mathrm{E}\big[y_{ij} \mid \mathbf{y}_{i,-j}, y_{ij} < \psi_{ij}; \theta^{(t)}\big] & \text{if } y_{ij} < \psi_{ij}, \end{cases} \tag{4}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, D-1$, where $\mathbf{y}_{i,-j}$ denotes the set of observed variables for the row i of the data matrix. By analogy with the models studied in Amemiya (1984), under normality, the expectation in (4) is obtained by

$$\begin{aligned} &\mathrm{E}[y_j \mid \mathbf{y}_{-j}, y_j < \psi_j] \\ &\quad = \frac{1}{P[y_j < \psi_j]} \int_{-\infty}^{\psi_j} y_j \big(2\pi\sigma_j^2\big)^{-1/2} \exp\left[-\frac{1}{2\sigma_j^2}\big(y_j - \mathbf{y}_{-j}^{\mathrm{T}}\beta\big)^2\right] \mathrm{d}y_j, \end{aligned}$$

where σ_j^2 denotes the variance of y_j, and β is the vector of coefficients of the linear regression of y_j on $\mathbf{y}_{-j}$. Some straightforward calculations (Palarea-Albaladejo and Martín-Fernández 2007) produce the expression

$$\mathrm{E}[y_j \mid \mathbf{y}_{-j}, y_j < \psi_j] = \mathbf{y}_{-j}^{\mathrm{T}}\beta - \sigma_j \frac{\phi((\psi_j - \mathbf{y}_{-j}^{\mathrm{T}}\beta)/\sigma_j)}{\Phi((\psi_j - \mathbf{y}_{-j}^{\mathrm{T}}\beta)/\sigma_j)}, \tag{5}$$

where ϕ and Φ are, respectively, the density and the distribution function of the standard normal random variable. From a computational point of view, since there are a high number of regression equations, we use a mathematical device to estimate them efficiently. The sweep operator is commonly used in this context. See Palarea-Albaladejo and Martín-Fernández (2007) for the computational details of this procedure.

Given the completed data set from the E-step, the modified M-step fully coincides with the common M-step. Thus, ordinary complete data ML estimates for $\theta^{(t+1)} = (\mu^{(t+1)}, \Sigma^{(t+1)})$ are obtained for the next iteration.

The E- and M-steps are iteratively repeated until convergence. In our implementation the algorithm stops when $\max\{|\mu^{(t+1)} - \mu^{(t)}|, |\Sigma^{(t+1)} - \Sigma^{(t)}|\}$ is lower than a fixed tolerance level $\varepsilon = 0.0001$. Once the convergence has been reached, we consider the last completed data set stored in the E-step and transform it back to the simplex by the inverse alr transformation, obtaining a completed compositional data set without zeros.

Properties of the Modified EM Algorithm

As we emphasized in the introduction, a zero replacement strategy must be coherent with the nature of the compositional data and with the basic operations on the simplex. We have verified (Palarea-Albaladejo and Martín-Fernández 2007) the main features of the modified EM algorithm, as follows.

1. The results are the same independently of the selected divisor in the alr transformation. That is, the algorithm is invariant under the group of permutations of the parts of a composition.
2. It replaces the rounded zero values by values below the detection limit.
3. It is coherent with the basic operations and the vector space structure defined on S^D. Let $\mathbf{x}_{-Z}$ be the vector of non-zero parts of a composition $\mathbf{x}$, and consider $\mathbf{x}_S$ the associate subcomposition $\mathbf{x}_S = \mathrm{C}(\mathbf{x}_{-Z})$. Analogously, as regards the replacement vector $\mathbf{r}$, we consider $\mathbf{r}_{-Z}$ ($\mathbf{r}$ without the originally zero parts) and $\mathbf{r}_S = \mathrm{C}(\mathbf{r}_{-Z})$. Then, the following properties are satisfied:
   (a) subcomposition invariance: $\mathbf{r}_S = \mathbf{x}_S$;
   (b) perturbation invariance: $(\mathbf{p} \oplus \mathbf{r})_S = (\mathbf{p} \oplus \mathbf{x})_S \ \forall \mathbf{p} \in S^D$;
   (c) power transformation invariance: $(\alpha \otimes \mathbf{r})_S = (\alpha \otimes \mathbf{x})_S \ \forall \alpha \in R$.
4. It is verified that $r_j/r_k = x_j/x_k$ for all non-zero parts x_j and x_k. The maintenance of the ratios implies that the covariance structure of parts without zeros is preserved. Hence, any subcompositional analysis based on the non-zero parts of the original or completed compositional data set gives the same outcomes.

Empirical Results

The Halimba Data Set

The data set used herein corresponds to the composition [Al_2O_3, SiO_2, Fe_2O_3, TiO_2, H_2O, Res_6] of 332 samples from 34 core-boreholes in the Halimba bauxite deposit (Hungary). The sixth part, Res_6, consists of a residual part of the composition, i.e. it is equal to $(100 - (Al_2O_3 + \cdots + H_2O))\%$. Since this paper focuses on the modified EM algorithm, we put no special emphasis on geological aspects in this case study. Let us call this data matrix $\mathbf{X}$. Initially, the data set does not include zero values and percentages are recorded using one decimal. For our calculations, we take the constant sum $c = 1$, i.e. compositions $\mathbf{x}$ satisfy $x_1 + \cdots + x_D = 1$. We summarize here the descriptive analysis made in Martín-Fernández et al. (2003a).

The compositional variation array (Aitchison 1986) is a very useful tool since it provides a descriptive summary of the relative variability of components. This

Table 1 Descriptive measures of Halimba data set

$g(\mathbf{X})$: (0.5644, 0.0246, 0.2421, 0.0282, 0.1242, 0.0166); totvar($\mathbf{X}$): 0.9718

j \ k	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.8946	**0.1288**	0.1793	**0.0885**	0.6105
SiO_2	3.1314	0	0.9095	0.9703	0.8515	0.9321
Fe_2O_3	**0.8464**	−2.2850	0	0.1915	**0.1519**	0.6194
TiO_2	2.9981	−0.1333	2.1516	0	0.2214	0.6603
H_2O	**1.5140**	−1.6174	**0.6676**	−1.4841	0	0.5566
Res_6	3.5284	0.3970	2.6819	0.5303	2.0144	0

array (Table 1) sets out the log-ratio variance $\mathrm{var}(\ln(x_j/x_k))$, $j = 1, \ldots, 5$ and $k = j+1, \ldots, 6$ as an upper triangular array. The lower triangle is exploited to display in position (j, k) an estimate of the log-ratio expectation $\mathrm{E}(\ln(x_k/x_j))$, $k = 1, \ldots, 5$ and $j = k+1, \ldots, 6$. The sign of these log-ratio sample means in Table 1 (lower triangular array) indicates that the parts SiO_2, TiO_2 and Res_6 take the smallest values. In this table we represent in bold the values corresponding to the parts without values smaller than 0.01 (1%). Table 1 also shows the compositional descriptive measures: geometric mean $g(\mathbf{X})$ and the total variability (totvar) of the data set $\mathbf{X}$. These measures are defined by

$$g(\mathbf{X}) = \mathrm{C}(g_1, \ldots, g_D) \quad \text{and} \quad \mathrm{totvar}\,(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{a}}^2\big(\mathbf{x}_i, g(\mathbf{X})\big),$$

where $g_j = (x_{1j} \cdot \cdots \cdot x_{nj})^{1/n}$ is the geometric mean of the component x_j and d_{a} is the Aitchison distance (1).

Figure 1 illustrates these descriptive measures in a compositional biplot (Aitchison and Greenacre 2002) which retains a high percentage (96.71%) of total variability. Note that, in essence, a compositional biplot is equivalent to a common biplot of the clr-transformed data set. For example, in Fig. 1(A) we observe that the longest rays correspond to the variables $\mathrm{clr}(SiO_2)$ and $\mathrm{clr}(Res_6)$ since they have the greatest log-ratio variability. We also see (Fig. 1(B)) that the observations with values below 0.01 (1%)—empty circles—mainly appear in the opposite direction of rays $\mathrm{clr}(SiO_2)$ and $\mathrm{clr}(Res_6)$.

In our case study, we transform every observed value smaller than $\gamma_j = 0.01$ to a rounded zero value (i.e. we interpret that our measurement device cannot detect proportions of chemical compounds smaller than 1%). We denote by $\mathbf{X}^*$ the compositional data set resulting from this procedure. Table 2 shows the distribution of the fictitious zeros in the six components. Within the data set, 104 observations have some zeros, but their total number is reduced since they only represent a 6.33% of the recorded values in the matrix. Table 2 indicates that components Al_2O_3, Fe_2O_3 and H_2O have no zeros. Except for one zero, the null values are concentrated in SiO_2 and Res_6.

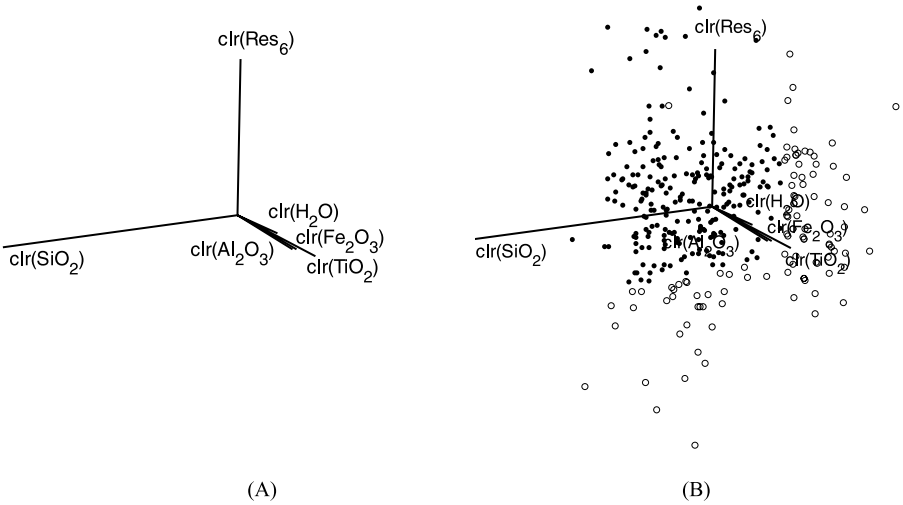


Fig. 1 Biplot in clr-transformed space: (**A**) clr-variables; (**B**) clr-transformed samples. *Empty circles* correspond to samples with some recorded value smaller than 0.01 (1%)

Table 2 The pattern of null values in the data set $\mathbf{X}^*$. 0 symbolizes that the component contains null values

Number of obs.	Pattern of zeros						[a]
	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6	
228							228
1				0			229
34						0	262
22		0				0	331
47		0					275

[a]Number of observations without zeros if the corresponding variables with zero values are not considered

Zero Replacement Strategies

Our aim is to compare the performance of our modified EM algorithm with other methods: non-parametric multiplicative replacement, common EM algorithm, Sandford method and MI. To make this comparison, our strategy starts by replacing the zero values in $\mathbf{X}^*$ using each method to obtain a completed compositional data set. The comparison is then made between the descriptive measures of the initial data set $\mathbf{X}$ and the completed data set. Let $\mathbf{r}_i$ be the completed composition obtained from $\mathbf{x}_i^* \in \mathbf{X}^*$. Since initial observations $\mathbf{x}_i \in \mathbf{X}$ are known, besides the usual descriptive measures, we can consider two measures of distortion (Martín-Fernández et al. 2001), the MSD (mean squared distances) and the STRESS (standardized residual sum of squares) given by

$$\text{MSD} = \frac{\sum_i d_{\text{a}}^2(\mathbf{x}_i, \mathbf{r}_i)}{332} \quad \text{and} \quad \text{STRESS} = \frac{\sum_{i<j}(d_{\text{a}}(\mathbf{x}_i, \mathbf{x}_j) - d_{\text{a}}(\mathbf{r}_i, \mathbf{r}_j))^2}{\sum_{i<j} d_{\text{a}}^2(\mathbf{x}_i, \mathbf{x}_j)},$$

where d_a represents the Aitchison distance (1). The STRESS measures the distortion due to compositions where both have zero values and the distortion due to compositions where only one of them has zero values. Note that both measures of distortion incorporate the Aitchison distance. Recall that the distance d_a between two compositions is equivalent to the Euclidean distance between their clr-transformed vectors or their ilr-transformed vectors. The alr transformation is not isometric, consequently these measures could show a different pattern if they are computed on the alr-transformed space using the Euclidean distance (Aitchison et al. 2000).

Multiplicative Replacement

Using multiplicative replacement (3), the imputed value δ_j must be selected in advance. Following Martín-Fernández et al. (2003a), we replace the rounded zeros by the small value $\delta_j = 0.0065$ for $j = 1, \ldots, 6$ (65% of 0.01, the detection limit). Table 3 shows descriptive measures of the completed data set resulting from this replacement. The values of MSD and STRESS are reasonably close to zero, which suggests that the distortion is minimal. The same conclusion is reached when we compare the descriptive measures in Table 3 with the true values (Table 1). Note that the relative structure of the parts containing non-zero values is preserved (bold values in Tables 1 and 3).

For a better description of the distortion, we can analyze the percentiles of the subtraction between the true values of the data in $\mathbf{X}$ and the corresponding values in the completed data set. Note that the subtraction is not a linear operation in the simplex. The difference between two compositional vectors would be more correctly expressed in terms of the inverse perturbation operation. Nevertheless, for the purpose at this point, we think subtractions represent distortion in a similar way and are somewhat more illustrative for the reader. In each boxplot diagram of these percentiles one must examine the symmetry and the position of the boxplot in relation to the zero value in the vertical axis. Figure 2(A) shows these percentiles in the boxplot

Table 3 Descriptive measures of completed data set resulting from multiplicative replacement with $\delta_j = 0.0065$

$g(\mathbf{X})$: (0.5645, 0.0247, 0.2421, 0.0281, 0.1242, 0.0164); totvar($\mathbf{X}$): 0.9568

Variation array

j k	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.8812	**0.1288**	0.1864	**0.0885**	0.6150
SiO_2	3.1305	0	0.8966	0.9581	0.8380	0.9150
Fe_2O_3	**0.8464**	−2.2840	0	0.1979	**0.1519**	0.6231
TiO_2	2.9990	−0.1314	2.1526	0	0.2273	0.6713
H_2O	**1.5140**	−1.6165	**0.6676**	−1.4850	0	0.5595
Res_6	3.5398	0.4093	2.6934	0.5408	2.0258	0

MSD: 0.0319; STRESS: 0.0203

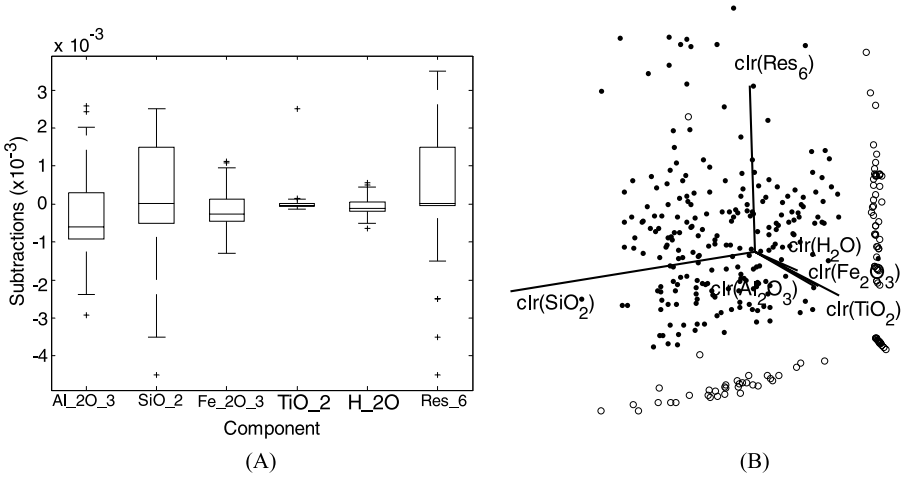


Fig. 2 Plots from multiplicative replacement with $\delta_j = 0.0065$: (**A**) Boxplot of the subtractions between observations of the data set **X** and observations of the completed data set; (**B**) Compositional biplot of completed data set. *Empty circles* correspond to completed samples

diagram of the subtractions for each component. Figure 2(B) shows the compositional biplot for the completed data set. Remember that the fictitious zeros are concentrated in the parts SiO_2, TiO_2 and Res_6.

In Fig. 2(A) we observe that the distortion is not large (10^{-3} on the Y axis) and is symmetric, i.e. the rounded zeros have been replaced by larger or smaller values than the true values in approximately the same proportion. Nevertheless, when we observe the replaced values (empty circles in Fig. 2(B)) we can see that the similarity between the completed observations is grossly exaggerated (empty circles in Fig. 1(B)). This effect is present in all the non-parametric methods of imputation and is not specific for the multiplicative replacement (3).

In addition, after the application of a non-parametric method, some kind of sensitivity analysis (Aitchison 1986, p. 270) is needed. It is necessary to analyze how the statistical results depend on the value δ_j introduced in the procedure (3). This sensitivity analysis was made in Martín-Fernández et al. (2003a), but is not reproduced here because our aim is to analyse the behaviour of the modified EM algorithm.

Previously Studied Parametric Methods

In this section we describe the results obtained applying the common EM algorithm, the Sandford method and the MI method. Following transformation methodology, these methods are applied to the alr-transformed data set and, subsequently, transformed them back to the simplex by the inverse alr transformation to obtain a completed compositional data set. Note that the replaced values are derived from expectations in the alr-transformed space, but for the completed compositional data to be still considered, a measure different from the Lebesgue measure is implicitly used in the simplex. In Mateu-Figueras and Pawlowsky-Glahn (2007) the first theoretical justification for this approach is introduced.

Table 4 Descriptive measures of completed data set resulting from the common EM algorithm

$g(\mathbf{X})$: (0.5581, 0.0328, 0.2394, 0.0279, 0.1228, 0.0189); totvar($\mathbf{X}$): 0.4540

Variation array

j k	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5590	**0.1288**	0.1678	**0.0885**	0.4599
SiO_2	2.8337	0	0.5723	0.6286	0.5222	0.6272
Fe_2O_3	**0.8464**	−1.9872	0	0.1819	**0.1519**	0.4762
TiO_2	2.9948	0.1611	2.1484	0	0.2119	0.5222
H_2O	**1.5140**	−1.3197	**0.6676**	−1.4808	0	0.3973
Res_6	3.3832	0.5496	2.5368	0.3884	1.8692	0

MSD: 0.4808; STRESS: 0.2322

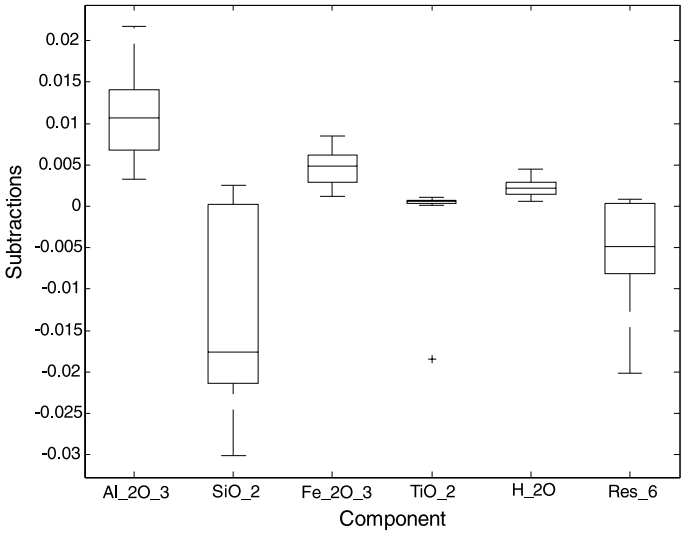


Fig. 3 Boxplot of the subtraction between observations of the data set $\mathbf{X}$ and observations of the completed data set resulting from the common EM algorithm

The described strategy needs one part free of zero values in the data set $\mathbf{X}^*$ in order to use it as divisor of the alr transformation (2). In our study we choose one of the parts Al_2O_3, Fe_2O_3 or H_2O as the divisor (Table 2). It must be checked whether the results produced by the different parametric methods are independent of the selected divisor. As it is well known (Aitchison 1986, p. 142), the divisor is not important when we apply the common EM algorithm, since this algorithm is invariant under the group of permutations of the parts of a composition. Table 4 shows descriptive measures of the data set resulting from this algorithm when we use the part Fe_2O_3 as divisor for the alr transformation (2).

Although the relative structure of the parts containing non-zero values are preserved (bold values in Tables 1 and 4), it can be seen that the distortion of the data structure of $\mathbf{X}$ is larger than the distortion produced by multiplicative replacement (see MSD and STRESS in Tables 3 and 4). This conclusion is confirmed when we compare the descriptive measures (Table 4) with the true values (Table 1). Note that the values of $\text{var}(\ln(x_j/x_k))$ (upper triangle in variation matrix) underestimate the true values. From Fig. 3 we can interpret that the common EM algorithm has mainly replaced the zero values by values that are larger (10^{-2} on the Y axis) than the true values because many of the subtractions in the parts SiO_2 and Res_6 are negative. Thus, in this case, the common EM algorithm replaces most of the fictitious zeros by values above the detection limit.

When the Sandford method or the MI method is applied to the alr-transformed data set, we observe that the resulting descriptive measures (MSD, STRESS, ...) are different depending on the part selected as divisor in the alr transformation (2). For the Sandford method, Martín-Fernández et al. (2003b) state that the differences are caused by the different values of the observed mean. For the MI method, further studies are necessary before it can be established whether this dependence is due to the selected divisor or it is only an effect of the simulation process. As for the common EM algorithm, we observe that this strategy does not take into account that the rounded zeros should be replaced by values below the detection limit. Both methods provide a similar outcome to the common EM algorithm. Thus, we do not report here their results and refer the reader to Martín-Fernández et al. (2003b) or Palarea-Albaladejo et al. (2005) for more details.

Modified EM Algorithm

The modified EM algorithm is applied to replace zeros in the data set $\mathbf{X}^*$. Table 5 shows the resulting descriptive measures. Since this algorithm is invariant under the group of permutations of the parts of a composition, we only report here the results when the part Fe_2O_3 is the divisor in the alr transformation (2).

Table 5 Descriptive measures of the completed data set resulting from the modified EM algorithm

$g(\mathbf{X})$: (0.5646, 0.0241, 0.2422, 0.0282, 0.1242, 0.0168); totvar($\mathbf{X}$): 0.9738

Variation array

j \ k	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.9167	**0.1288**	0.1784	**0.0885**	0.5817
SiO_2	3.1531	0	0.9329	0.9922	0.8733	0.9223
Fe_2O_3	**0.8464**	−2.3067	0	0.1908	**0.1519**	0.5912
TiO_2	2.9979	−0.1552	2.1515	0	0.2207	0.6383
H_2O	**1.5140**	−1.6391	**0.6676**	−1.4839	0	0.5240
Res_6	3.5169	0.3638	2.6705	0.5190	2.0029	0

MSD: 0.0339; STRESS: 0.0222

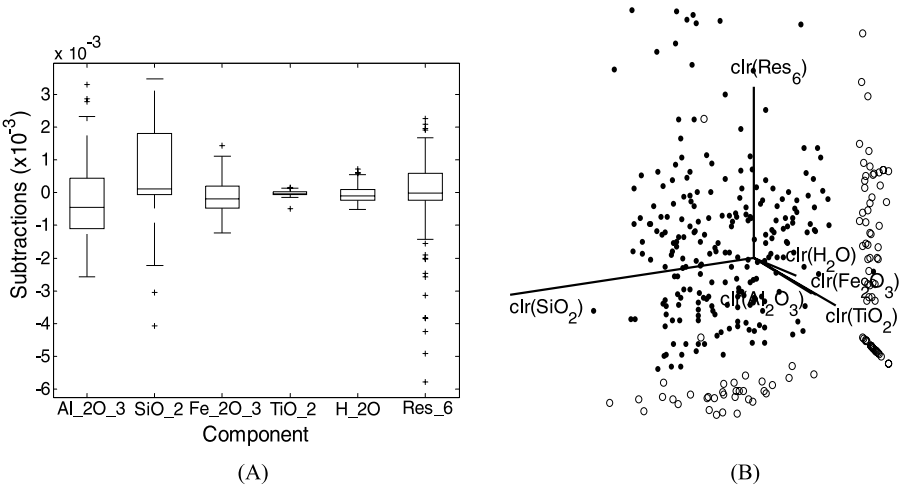


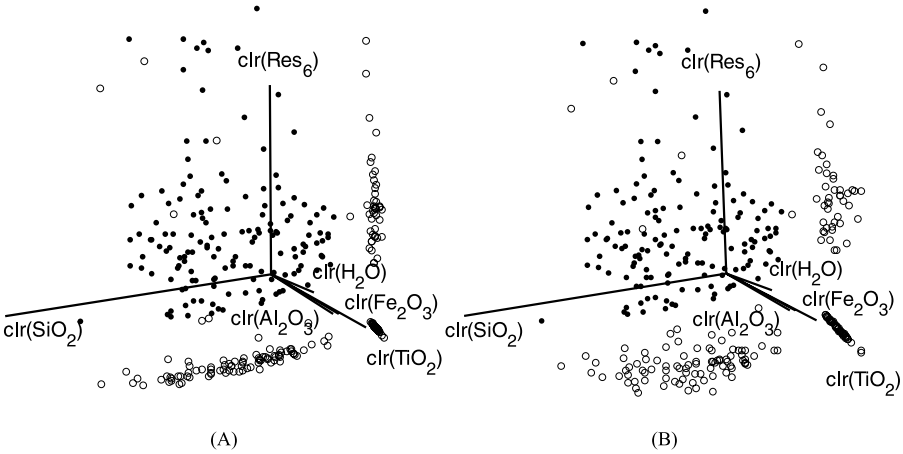
Fig. 4 Plots from the modified EM algorithm: (**A**) boxplot of the subtractions between observations of the data set **X** and observations of the completed data set; (**B**) compositional biplot of completed data set. *Empty circles* correspond to completed samples

Observe that the relative structure of the parts containing non-zero values is preserved (bold values in Tables 1 and 5) and log-ratio variances are somewhat overestimated (Table 1). The MSD and STRESS measures take values close to zero significantly improving the results obtained with the other parametric methods. In addition, we observe in Fig. 4(A) that modified EM takes into account (10^{-3} on the Y axis) that rounded zeros must be replaced by small values. Some of these values are greater than the true value and others are lower, but all of them are below the detection limit. The modified EM algorithm provides very similar results to multiplicative replacement (3). Only the total variance totvar($\mathbf{X}$) = 0.9718 (Table 1) seems to be better estimated by the modified EM (totvar($\mathbf{X}$) = 0.9738, Table 5) than the multiplicative replacement (totvar($\mathbf{X}$) = 0.9568, Table 3). In the case of multiplicative replacement, we have observed (Fig. 2(B)) that completed observations are too similar to each other. From Fig. 4(B), we can interpret that the modified EM algorithm shows a better behaviour because it seems that the completed observations (empty circles) have decreased their similarity. Broadly speaking, the modified EM algorithm does not seem to significantly improve the results obtained with multiplicative replacement. This fact suggests that non-parametric replacement (3) should be applied when the number of zeros is not large. Even so, the modified EM has the advantage over multiplicative replacement that it avoids the need to make a posterior sensitivity analysis in relation to the value δ_j used in (3).

In addition, when the number of zeros is large, the modified EM could well behave better than multiplicative replacement. To confirm this, we increase the fictitious detection limit for the data set Halimba from 1% to 1.5%. Now 11.80% of the values in the data matrix $\mathbf{X}^*$ are rounded zeros, mainly concentrated in Res_6. Table 6 shows that the modified EM algorithm provides better results than multiplicative replacement (with $\delta_j = 0.00975$, 65% of the detection limit) in the descriptive measures

Table 6 Descriptive measures of completed data set when the fictitious detection limit is equal to 1.5%

	Multiplicative replacement	Modified EM algorithm
$g(\mathbf{X})$	(0.5633, 0.0264, 0.2416, 0.0280, 0.1239, 0.0167)	(0.5642, 0.0253, 0.2420, 0.0283, 0.1241, 0.0161)
totvar($\mathbf{X}$)	0.7714	0.8707
MSD	0.0903	0.0779
STRESS	0.0563	0.0444



(A) (B)

Fig. 5 Compositional biplots of completed data set when the fictitious detection limit is equal to 1.5%. *Empty circles* correspond to completed samples: (**A**) from multiplicative replacement; (**B**) from the modified EM algorithm

totvar ($\mathbf{X}$), MSD and STRESS. We compare the completed observations from multiplicative replacement (empty circles in Fig. 5(A)) with the completed observations from modified EM (empty circles in Fig. 5(B)) and conclude that the variability of these observations is increased. This suggests that for compositional data sets with a large number of rounded zeros it might be preferable to apply the modified EM algorithm rather than the multiplicative replacement. A complete study of the behaviour of this algorithm in relation to the number of rounded zeros in the data set is given in Palarea-Albaladejo and Martín-Fernández (2007).

Modified EM in Two Extreme Cases

Now we consider two particular real compositional data sets. One of them has a small number of parts and zeros, and the other a large number of parts and zeros. The first comes from Aitchison (1986, p. 399) and consists of 30 samples of foraminiferal compositions at 30 different depths. The compositions only contain four parts and the total number of zeros is 5. The detection limit for this data set is 0.01. The second data set corresponds to granulometric data from the Darss Sill area in the Baltic Sea. For a better description of the data, see Martín-Fernández et al. (2003a). This set

Table 7 Descriptive measures of the completed Foraminifera and Darss Sill data sets. 'MR' corresponds to multiplicative replacement; and 'mEM' to the modified EM algorithm

		Foraminiferal compositions	Darss Sill
$g(\mathbf{X})$	MR	(0.6830, 0.2465, 0.0400, 0.0305)	(0.0017, 0.0033, 0.0079, 0.0338, 0.3418, 0.5540, 0.0484, 0.0090)
	mEM	(0.6847, 0.2471, 0.0379, 0.0303)	$(7 \times 10^{-10}, 0.0003, 0.0038, 0.0330, 0.3448, 0.5603, 0.0490, 0.0089)$
totvar $(\mathbf{X})$	MR	2.4499	16.9151
	mEM	2.6756	64.4153
MSD		0.0260	221.0945
STRESS		0.0072	0.3146

contains 1281 samples with 8 parts. In 1163 of them there is at least one zero and the total number of zeros in the data matrix is 2852. These zeros are mainly concentrated in the first, second, and third components, respectively, 1156, 925, and 599 zeros, while the sixth component has no null values. The detection limit considered here for all the components is 0.001.

Following the same scheme as in previous sections, the modified EM algorithm is executed and convergence is achieved in 8 and 593 iterations, respectively, for each data set. Table 7 summarizes the main descriptive measures obtained. In this table 'MR' denotes results obtained by multiplicative replacement; and 'mEM' denotes results from the modified EM algorithm. In the Halimba case study the measures MSD and STRESS have been used to compare the true data set $\mathbf{X}$ and the completed data set (Tables 3–6). In contrast, we now use (Table7) these measures to compare the two completed data sets obtained from 'MR' and 'mEM'. Note that all the descriptive measures for the Foraminiferal data (Table7) suggest that both methods—MR and mEM—produce very similar completed data sets.

For the Darss Sill data set the conclusion is the opposite. Observe that for those components with many zeros (1st, 2nd and 3rd) the compositional geometrical mean $g(\mathbf{X})$ is very different. The variability of the completed data set is larger when modified EM is applied, and the MSD and STRESS measures suggest that the two completed data sets are very different. This different pattern is shown in Fig. 6, where the 1163 completed samples obtained from multiplicative replacement (3) are plotted with those 1163 completed observations produced from the modified EM algorithm in a compositional biplot. This biplot is of very good quality because it explains 92.3% of the total variability, which mainly appears to be due to the different behaviour of the two methods in relation to the first component of the composition. We observe that the longest ray corresponds to the variable 'clr1' and the completed observations for multiplicative replacement (squares □) take on a similar value in this variable. This behaviour states that the similarity between the completed observations is grossly exaggerated. Note that there are few completed observations where multiplicative replacement and modified EM have the same behaviour. These observations have no zero in the first, second and third component and correspond in the biplot to the points where square □ and plus + overlap.

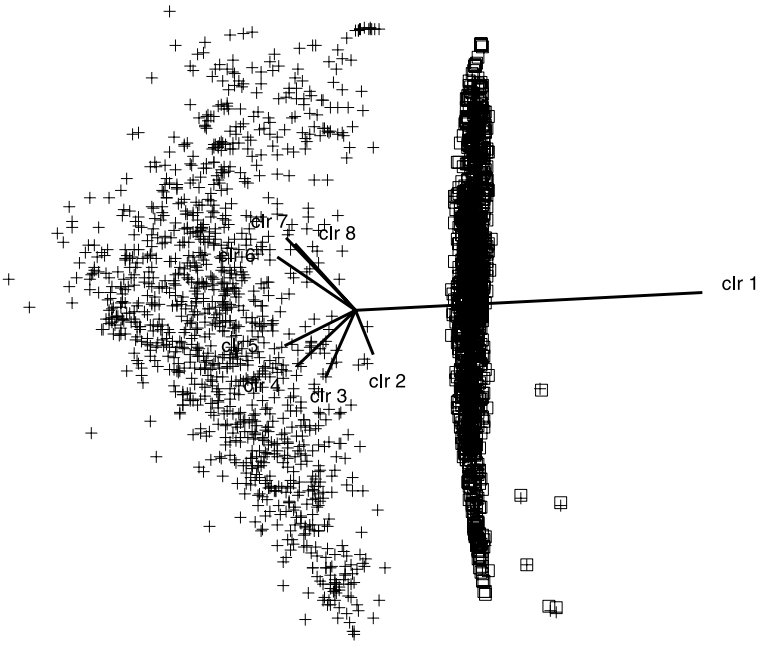


Fig. 6 Compositional biplot of completed samples of Darss Sill data set. □: for multiplicative replacement; +: for the modified EM algorithm

Conclusions

The presence of rounded zeros is a practical problem in compositional data analysis based on log-ratio methodology. A proper replacement of zero values by small non-zero values may overcome this problem. In this work, we have reviewed and compared the main parametric strategies for rounded compositional zeros. All methods analyzed are coherent with the basic operations on the simplex. This coherence implies that the covariance structure of subcompositions without zeros is preserved. Nevertheless, the common EM algorithm, the Sandford method and the MI method in its standard formulation do not take into account that the rounded zeros should be replaced by small values below the detection limit, and this is an important deficiency.

As an alternative, we have introduced a modification of the EM algorithm, which is applied in combination with the additive log-ratio transformation. This method takes into account that the imputed value must be lower than the given detection limit and is independent on the selected divisor in the alr transformation. Additionally, the zero values are imputed conditionally on the information included in the observed data. Reasonable estimations are obtained and minimum distortion is produced. When the number of rounded zeros is small, the modified EM does not significantly improve the performance of the non-parametric multiplicative replacement. However, case studies reveal its potential in the presence of a large amount of null values.

The method we have analyzed in this paper can be mainly extended in two ways: first, by obtaining measures of variability for the estimates, which must incorporate

the additional uncertainty due to the presence of compositional rounded zeros; second, by generalizing the modified EM algorithm to compositional data sets simultaneously containing rounded zeros and missing data.

Acknowledgements This research has been financially supported by the Spanish Ministry of Education and Science through the project MTM2006-03040. The authors would like to thank Dr. G. Bardossy of the Hungarian Geological Society for providing the Halimba data set, and the reviewers for their helpful and valuable comments on the paper.

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London. Reprinted in 2003 by Blackburn Press, 416 p

Aitchison J, Greenacre M (2002) Biplots of compositional data. Appl Stat 51(4):375–392

Aitchison J, Kay JW (2004) Possible solutions of some essential zero problems in compositional data analysis. In: Thió-Henestrosa S, Martín-Fernández JA (eds) Compositional data analysis workshop, Girona, Spain. http://ima.udg.es/Activitats/CoDaWork03/

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32(3):271–275

Amemiya T (1984) Tobit models: a survey. J Econom. 24:3–61

Bacon-Shone J (2003) Modelling structural zeros in compositional data. In: Thió-Henestrosa S, Martín-Fernández JA (eds) Compositional data analysis workshop, Girona, Spain. http://ima.udg.es/Activitats/CoDaWork03/

Buccianti A, Rosso F (1999) A new approach to the statistical analysis of compositional (closed) data with observations below the "detection limit". Geoinformatica 3:17–31

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc Ser B 39:1–38

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal (2003) Isometric logratio transformation for compositional data analysis. Math Geol 35(3):279–300

Fry JM, Fry TRL, McLaren KR (2000) Compositional data analysis and zeros in micro data. Appl Econom 32:953–959

Gómez-García J, Palarea-Albaladejo J, Martín-Fernández JA (2006) Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. Revista Estadística Española 48(162):241–270

Heckman J (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. Ann Econom Soc Meas 5:475–492

Honaker J, Katz JN, King G (2002) A fast, easy, and efficient estimator for multiparty electoral data. Political Anal 10(1):84–100

King G, Honaker J, Joseph A, Scheve K (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am Political Sci Rev 95(1):49–69

Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York, 381 p

Martín-Fernández JA, 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: from theory to practice, vol 264. The Geological Society, London, pp 191–201

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2000) Zero replacement in compositional data sets. In: Kiers H, Rasson J, Groenen P, Shader M (eds) Studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 155–160

Martín-Fernández JA, Olea-Mensese R, Pawlowsky-Glahn V (2001) Criteria to compare estimation methods of regionalized compositions. Math Geol 33(8):889–909

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003a) Dealing with zeros and missing values in compositional data sets. Math Geol 35(3):253–278

Martín-Fernández JA, Palarea-Albaladejo J, Gómez-García J (2003b) Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach. In: Thió-Henestrosa S, Martín-Fernández JA (eds) Compositional data analysis workshop, Girona, Spain. http://ima.udg.es/Activitats/CoDaWork03/

Mateu-Figueras G, Barceló-Vidal C (eds) (2005) Second compositional data analysis workshop—CoDaWork'05, Proceedings, Universitat de Girona, CD-ROM, ISBN: 84-8458-222-1; available at http://ima.udg.es/Activitats/CoDaWork05/

Mateu-Figueras G, Pawlowsky-Glahn V (2007) The skew-normal distribution on SD. Special issue: Skew-elliptical distributions and their application. Commun Stat Theory Methods 36(9):1787–1802

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York, 274 p

Palarea-Albaladejo J, Martín-Fernández JA, Gómez-García J (2005) ALR approach for replacing values below the detection limit. In: Mateu-Figueras G, Barceló-Vidal C (eds) Compositional data analysis workshop, Girona, Spain, 2005. http://ima.udg.es/Activitats/CoDaWork05/

Palarea-Albaladejo J, Martín-Fernández JA (2007) A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Comput Geosci (submitted)

Pawlowsky-Glahn V (guest ed) (2005) Special issue: Advances in compositional data. Math Geol 37(7): 671–850

Rubin DB (1987) Multiple imputation for nonresponse in survey. Wiley, New York, 258 p

Sandford RF, Pierson CT, Crovelli RA (1993) An objective replacement method for censored geochemical data. Math Geol 25(1):59–80

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall, London, 430 p

Thió-Henestrosa S, Martín-Fernández JA (eds) (2003) Compositional data analysis workshop—CoDaWork'03, Proceedings, Universitat de Girona, CD-ROM, ISBN: 84-8458-111-X; available at http://ima.udg.es/Activitats/CoDaWork03/

Wu CFJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11:95–103