alr approach for replacing values below the detection limit

J. Palarea-Albaladejo[1], J. A. Martín-Fernández[2] , and J. Gómez-García[3]

[1] jpalarea@pdi.ucam.edu, Departamento de Informática de Sistemas,
Universidad Católica San Antonio de Murcia

[2] josepantoni.martin@udg.es, Departament d'Informàtica i Matemàtica Aplicada,
Universitat de Girona

[3]jgomezg@um.es, Departamento de Métodos Cuantitativos para la Economía,
Universidad de Murcia

Abstract

All of the imputation techniques usually applied for replacing values below the detection limit in compositional data sets have adverse effects on the variability. In this work we propose a modification of the EM algorithm that is applied using the additive log-ratio transformation. This new strategy is applied to a compositional data set and the results are compared with the usual imputation techniques.

Key words: compositional data, EM algorithm, log-ratio, rounded zeros.

1 Introduction

In compositional data analysis multivariate statistical methods based on log-ratio methodology are applied. Then one realises that zeros are a problem because ratios and logarithms can't be made. Presence of rounded zeros is very common when a compositional data set is analyzed. These rounded zeros are produced by the values below to the detection limit of the process of measure. In compositional data analysis other different kind of zeros is considered. A zero is called essential zero or structural zero if this null value means that the part is completely absent. Essential zeros must be treated in a different way (Aitchison, 1986; Martín-Fernández, Barceló-Vidal y Pawlowsky-Glahn, 2003) than the rounded zeros. For these ones, imputation techniques are commonly used because, in essence, a rounded zero can be considered as a missing value. It is very important to emphasize that the decision to apply specific techniques for missing values to deal with the rounded zeros is independent to the statistical methodology (Euclidean, log-ratio,...) that one has selected. In other words, even when one scientist selects the Euclidean option he has the "rounded zeros problem" because he should to deal with the missing values.

All the papers and books related to imputation techniques recommend that one should be careful to use a replacement strategy because the general structure of the data could be seriously distorted. In particular, the covariance structure and the metric properties of the data set should be preserved in order to avoid that further analysis on sub-populations be misleading. Note that last sentence express the clue of the replacement techniques for rounded zeros in compositional data. The specific nature of compositional data forces to decide in advance which kind of covariance structure and metric properties one wants to preserve. According to the existing possibilities, one has to decide between the preservation either the classical –Euclidean– covariance and metric or the covariance and the metrics induced by the log-ratio methodology. As is well known, compositional data are formed by continuous variables, which scale of measurement is ratio scale and their main operations are perturbation and subcomposition. Consequently, at least for compositional data, the replacement strategies must be coherent with all these basic aspects.

From a non-parametric point of view, Aitchison (1986, p. 269) proposes an additive replacement method. Zhou (1997) and Tauber (1999), from descriptive point of view, illustrate that this additive replacement could produce spurious groups in cluster analysis when the imputed values tend to zero. Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2000) and Fry, Fry and McLaren (2000) show that the additive replacement doesn't preserves the ratio of non-zero values and, consequently distort the covariance structure of the data set. Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2003) analyse in depth the additive replacement and propose a multiplicative replacement which has better behaviour in relation to the compositional nature of the data.

As is stated in Little y Rubin (2002) and Schafer (1997), the EM algorithm (Dempster, Laird and Rubin, 1977) and the multiple imputation method (Rubin, 1987) are the most reliable methods applied in a parametric context of missing data. Both techniques rely on fully parametric models for multivariate data, usually the normal distribution, and contain in its formulation the variance-covariance matrix. If one directly applies to compositional data a serious distortion of the structure of data is produced. It is easy to show how this strategy can impute (Martín, Palarea and Gómez, 2003), negative values or observations which sum vector is greater than one (or 100%). Following Aitchison (1986), if one applies the additive log-ratio transformation (*alr*) a normal multivariate model could be considered because the data are in the real space. This strategy is applied in Buccianti and Rosso (1999), where from empirical point of view the authors describe the performance of the EM algorithm and, its extension, the Sandford's method (Sandford, Pierson and Crovelli, 1993). In Martín, Palarea and Gómez (2003) a first application of multiple imputation via Markov Chain Monte Carlo (MCMC) to rounded zeros is made, and its behaviour is described. All of these strategies have adverse effects in relation to the covariance structure of compositional data set.

In the following section we present a modification of the EM algorithm in order to be applied in the rounded zeros problem. The performance of this new method takes into account that the rounded zeros must be replaced by small values below the detection limit. Next, we apply all the methods to a compositional data set and compare their behaviour. Finally, main conclusions and future lines of research are presented.

2 A *modified* EM algorithm for dealing with rounded compositional zeros

The *common* EM (Expectation-Maximization) algorithm for missing data problems is a broadly applicable iterative algorithm for computing maximum-likelihood estimates for parametric models in situations where portions of a data matrix Y are missing. The data matrix Y contains a random sample of size n on (Y_1,...,Y_p). The observed part of Y is denoted by Y_{obs}, and the missing part by Y_{miss}, so that Y = (Y_{obs}, Y_{miss}). As is well known, the E step finds the conditional expectation of the missing data, or functions of missing data, given the observed data and current estimated parameters, and then substitutes these expectations for the missing data. The M step performs maximum-likelihood estimation of θ just if there were no missing data in the matrix Y.

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 generating the missing data and consider the observed-data likelihood as the relevant likelihood for the vector of unknown parameters θ of the complete-data model. Typically, two ignorability situations are distinguished: MAR (*missing at random*), when the probability that one value is missing depends on the Y_{obs} part of the vector but not on Y_{miss}; 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 some ad hoc missing-data methods can provide proper inferences. The common EM algorithm and the multiple imputation method assume that the missingness mechanism is MAR and its developments are 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 Y_{miss} part of the vector. This case represents the nonignorable situation and, usually, is hardest to deal with analytically and special models and methods are required. For continuous data, one group of nonignorable methods is based on models known as stochastic censoring or selection models (Heckman, 1976; Amemiya, 1984). Other used models are pattern-mixture models and pattern-set mixture models (Little, 1993). Nevertheless, in some empirical NMAR situations the MAR assumption has been found to yield more accurate predictions of the missing values than nonignorable modeling (Rubin, Stern and Vehovar, 1995). Clearly, in a compositional context, the rounded zeros problem is a NMAR case: data have a left censored point in the detection limit. This fact could explain why the *common* EM algorithm and the multiple imputation methods have adverse effects (see subsections 3.2.2 and 3.2.3).

Suppose that Y=(Y_1,...,Y_{D-1}) are the *alr*-transformed variables of the X=(X_1,…,X_D) compositional data set. Assume that Y has a (D-1)-variate normal distribution with mean vector μ and covariance matrix Σ. As is well-known, the complete-data log-likelihood function for μ and Σ based on a sample of size n is

$$\ell(\mu,\Sigma \mid Y) = -n\ln(2\pi) - \frac{1}{2}n\ln\left|\Sigma\right| - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu),$$

Because the normal model belongs to the regular exponential family, the EM algorithm has a particularly simple implementation. In particular, the complete-data log-likelihood is lineal in the following complete-data sufficient statistics needed to estimate the mean and covariance matrix,

$$T_j = \sum_{i=1}^{n} y_{ij},\ \ j = 1,\ldots,D-1 \ \text{ and } \ T_{jh} = \sum_{i=1}^{n} y_{ij}y_{ih},\ \ j,h = 1,\ldots,D-1.$$

Let $\theta^{(t)}$ = ($\mu^{(t)}$, $\Sigma^{(t)}$) denote the estimates of the parameters at the *t*th iteration. In this work, in order to propose a *modified* EM algorithm we consider the matrix C = (c_{ij}) of censoring points for Y, where c_{ij} = ln(δ_j/x_{iD}), i = 1,…,n and j = 1,..,(D-1), with δ_j being the detection limit for X_j. Then, the *modified* E step of the EM algorithm requires the computation of the expectations of T_1 and T_2 conditioning on the observed data, Y_{obs}, and the current parameter estimates, $\theta^{(t)}$, as follows:

$$T_j^{(t)} = E\left(T_j \mid Y_{obs},\theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t)},\ \ j = 1,\ldots,D-1\ ,$$

and

$$T_{jh}^{(t)} = E\left(T_{jh} \mid Y_{obs},\theta^{(t)}\right) = \sum_{i=1}^{n}(y_{ij}^{(t)}y_{ih}^{(t)} + v_{jhi}^{(t)}),\ \ j,h = 1,\ldots,D-1,$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{if } y_{ij} \text{ is observed} \\ E(y_{ij} \mid y_{obs,i}, y_{ij} < c_{ij}, \theta^{(t)}) & \text{if } y_{ij} \text{ is missing} \end{cases},$$

with $y_{obs,i}$ representing the set of variables observed for case i, i = 1,...,n, and

$$v_{jhi}^{(t)} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ E(\sum_{i=1}^{n} y_{ij}y_{ih} \mid y_{obs,i}, y_{ij} < c_{ij}, y_{ih} < c_h, \theta^{(t)}) & \text{if } y_{ij} \text{ and } y_{ih} \text{ is missing} \end{cases}$$

We propose that in this *modified* EM algorithm missing values y_{ij} are thus replaced by the conditional mean of y_{ij} given the set of values $y_{obs,i}$ observed for that case and the censoring point c_{ij}. By analogy with the strategy in Amemiya (1984), when we assume normality this conditional mean can be obtained by

$$E(y_{ij} \mid y_{obs,i}, y_{ij} < c_{ij}, \theta^{(t)}) = y_{obs,i}^{*}\hat{\beta}_j - \hat{\sigma}_j \frac{\phi\left(\dfrac{c_{ij} - y_{obs,i}\hat{\beta}_j}{\hat{\sigma}_j}\right)}{\Phi\left(\dfrac{c_{ij} - y_{obs,i}\hat{\beta}_j}{\hat{\sigma}_j}\right)},$$

where $y_{obs,i}^{*}$ represents the extended vector (1 $y_{obs,i}$), $\hat{\beta}_j$ is the vector of regression coefficients (including the constant term) of Y_j on the observed variables for case i, $\hat{\sigma}_j$ is the estimated standard deviation of Y_j, and Φ and ϕ are the distribution and density function respectively of the standard normal variable.

The *modified* M step fully coincides with the *common* M step. Here, the new estimates $\theta^{(t+1)}$ of the parameters are computed from the estimated complete-data sufficient statistics:

$$\mu_j^{(t+1)} = \frac{1}{n} T_j^{(t)},\ j = 1,\dots,k \quad \text{and} \quad \sigma_{jh}^{(t+1)} = \frac{1}{n}\left(T_{jh}^{(t)} - \frac{1}{n} T_j^{(t)} T_h^{(t)} \right),\ j,h = 1,\dots D-1.$$

The E and M steps are iteratively repeated until convergence. In particular, we have implemented an usual criterion: the algorithm stops when $\max\{|\mu^{(t+1)} - \mu^{(t)}|, |\Sigma^{(t+1)} - \Sigma^{(t)}|\}$ is lower than the tolerance level ε . In particular, in the case study of subsection 3.2.4 we take $\varepsilon = 0.0001$. After the algorithm stops, we return to the simplex by the inverse alr-transformation obtaining a compositional data set without zeros.

3 Empirical application

3.1 The Halimba data set

The data set, provided by G. Bardossy from the Hungarian Academy of Sciences, and previously used in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003), corresponds to the subcomposition [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). Let us call this data set X. The sixth part Res_6 consists in a residual part of the composition, i.e., it is equal to (100-(Al_2O_3+…+ H_2O))%. A brief descriptive analysis of the data set give us that the smallest values appear in components SiO_2, TiO_2 and Res_6, and that the larger variability appears in the second and sixth components, i.e. SiO_2 and Res_6. As is well known, the compositional variation array provides a useful descriptive summary of the pattern of variability of compositions. In this array we set out the logratio variance var[ln(X_k/X_j)]; (j=1,2,…,5; k=j+1,…,6) as an upper triangular array and we use the lower triangle to display in position (j,k) an estimate of the logratio expectation E[ln(X_k/X_j)]; (j=2,…,6; k=1,…,j-1). The variation array of the Halimba data set X is given in Table 1. Observe that the sign of the logratio means corroborate that the parts SiO_2, TiO_2 and Res_6 take smallest values. The larger values of logratio variance appear when SiO_2 or Res_6 are involved. In this table we have reported in black the values corresponding to the parts without values smaller than 0.01. Finally, we can compute the compositional geometric mean and the total variability of the data set X, respectively:

$$\hat{\xi} = (0.5644,\ 0.0246,\ 0.2421,\ 0.0282,\ 0.1242,\ 0.0166) \text{ and totvar(X)} = 0.9718\ .$$

Table 1. Variation array of Halimba data set: Uppertriangle var[ln(X_k/X_j)]; lower triangle E[ln(X_k/X_j)] (see text for more details).

	k					
j	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

Following Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003), every observed value of X smaller than 0.01 is transformed to a zero value. We call X* the compositional data set resulting from this procedure. As a consequence, out of the 332x6 values in the data matrix X*, 128 are zero, distributed in 105 compositions or rows. Note that this amount of zero values is reduced (less than 10%). Therefore, it seems reasonable to expect that imputation techniques give us suitable results. These zeros are mainly concentrated in the parts SiO_2 and Res_6. Only one zero appears in the fourth part TiO_2. As can be deduced from Table 2, the parts Al_2O_3, Fe_2O_3 and H_2O have no zeros in X*.

Table 2. Pattern of existing missing values in data set X*. Letter "M" symbolizes that the variable contains missing value.

	Pattern of missing values						
Amount of obs.	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6	Amount of observed obs. if...(a)
227							227
1				M			228
34						M	261
23		M				M	331
47		M					274

(a) Amount of observed obs. without missing values if the variables with missing values in this pattern are not considered.

In our study we assume the zeros of X* to be non-essential zeros, *i.e.* rounded zeros. Before applying any multivariate method to the data set X*, the zeros have to be replaced.

3.2 Zero replacement strategies

Our aim is to compare the performance of the different imputation techniques: non-parametric multiplicative replacement, EM algorithm, Sandford's method, MCMC multiple imputation (first approach in Martín, Palarea y Gómez, 2003), and the new EM-type algorithm proposed in this paper.

Because we know in this case the original observations $x_i \in X$, we perform this analysis using descriptive measures (boxplot, compositional geometric mean, total variability, and variation array) of the replaced compositions r_i obtained from $x^*_i \in X^*$. In addition, following Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003), we calculate the Aitchison's distance $d_a(x_i, r_i)$, i=1,2,…,332 between the original composition $x_i \in X$ and the replaced composition r_i. As a first measure of distortion, we consider the mean of these distances squared

$$\mathrm{MSD} = \frac{\sum d_a^2(x_i, r_i)}{332}$$

and, as a second measure of distortion we consider the STRESS (standardized residual sum of squares) defined by

$$\mathrm{STRESS} = \frac{\sum_{i<j} \left(d_a(x_i, x_j) - d_a(r_i, r_j) \right)^2}{\sum_{i<j} d_a^2(x_i, x_j)}$$

Note that in a different manner than in MSD, in STRESS we measure the distortion due to compositions where both have zero values, as well as the distortion due to compositions where only one of them has zero values.

3.2.1 Multiplicative replacement

Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003) show that the best results with multiplicative replacement are obtained when the rounded zeros in X* are replaced by the "small" value δ=0.0065 (65% of the detection limit). Table 3 shows descriptive measures of data set resulting from this replacement.

Table 3. Descriptive measures of data set resulting from multiplicative replacement with δ=0.0065. (see text for more details).

Compositional geometric mean: (0.5645, 0.0246, 0.2421, 0.0281, 0.1242, 0.0164)

Total variability: 0.9602

Variation array						
	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.8829	**0.1288**	0.1864	**0.0885**	0.6166
SiO_2	3.1318	0	0.8984	0.9598	0.8397	0.9153
Fe_2O_3	**0.8464**	-2.2853	0	0.1979	**0.1519**	0.6246
TiO_2	2.9990	-0.1327	2.1526	0	0.2273	0.6727
H_2O	**1.5140**	-1.6178	**0.6676**	-1.4850	0	0.5612
Res_6	3.5411	0.4093	2.6947	0.5421	2.0271	0

MSD: 0.0328

STRESS: 0.0210

In Table 3 we can observe that the values of MSD and STRESS are reasonably close to zero. Thus we can conclude that the distortion of the data structure of X has not been large. The same conclusion is obtained when we compare the true values of the compositional geometric mean, total variability and the elements of the variation array with the values in Table 3. Note that the relative structure of the parts containing no zero values is preserved (black values in Table 3).

To major description of the distortion we can analyze the percentiles of the differences between the true values of the data in X and the values of the data resulting from the multiplicative replacement. Figure 1 shows these percentiles in the boxplot diagrams of the differences for each part. Remember that the zeros are concentrated in the parts SiO_2, TiO_2, and Res_6. In this figure we can observe that the distortion is not large and is symmetric, *i.e.* the true values have been replaced by larger or smaller values in approximately the same proportion.

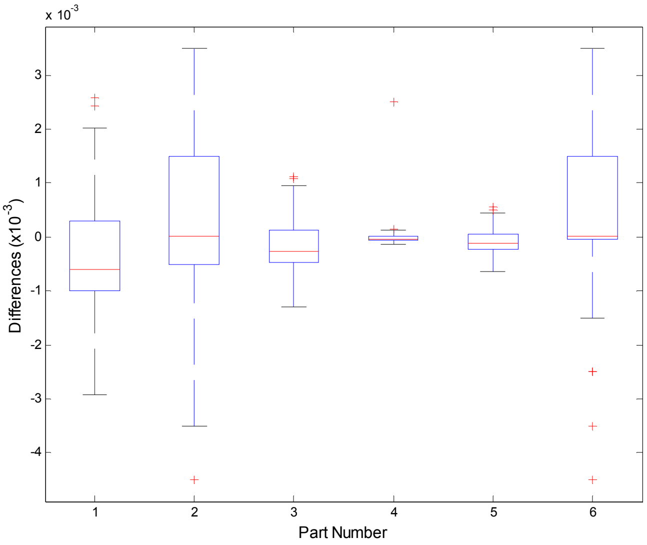

Figure 1. Boxplot of the differences between observations of the data set X and observations of the data set resulting from multiplicative replacement with δ=0.0065. (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

3.2.2 The *common* EM algorithm and the Sandford's method

The *common* EM algorithm and, its extension, the Sandford's method are applied to the *alr*-transformated data set. Then we return to the simplex. This strategy, used for the same proposes in Buccianti and Rosso (1999), needs one part free of zero values in data set X* in order to use it as divisor of the *alr*-

transformation. Remember that the parts Al_2O_3, Fe_2O_3, and H_2O_6 have not zero values. Therefore, as we can choose one of them as a divisor then we must analyze if the results are independent in relation to the selected divisor. As it is well known (Aitchison, 1986; Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 1999; Buccianti and Rosso, 1999) the choice of the part as the divisor is not important when we apply EM algorithm since this algorithm is invariant under the group of permutations of the parts of the compositions. Table 4 shows descriptive measures of data set resulting from this algorithm when we use the part Fe_2O_3 as divisor for the alr-transformation. Figure 2 shows the pattern of the differences between observations of the data set X and observations of the data set resulting from EM algorithm.

Table 4. Descriptive measures of data set resulting from EM algorithm (see text for more details).

Compositional geometric mean: (0.5581, 0.0330, 0.2394, 0.0279, 0.1228, 0.0189)						
Total variability: 0.4477						
			Variation array			
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5544	**0.1288**	0.1677	**0.0885**	0.4584
SiO_2	2.8293	0	0.5673	0.6235	0.5178	0.624
Fe_2O_3	**0.8464**	-1.9828	0	0.1819	**0.1519**	0.4736
TiO_2	2.9946	0.1654	2.1482	0	0.2119	0.5159
H_2O	**1.5140**	-1.3152	**0.6676**	-1.4806	0	0.3954
Res_6	3.3851	0.5558	2.5386	0.3904	1.8711	0
MSD: 0.4795						
STRESS: 0.2354						

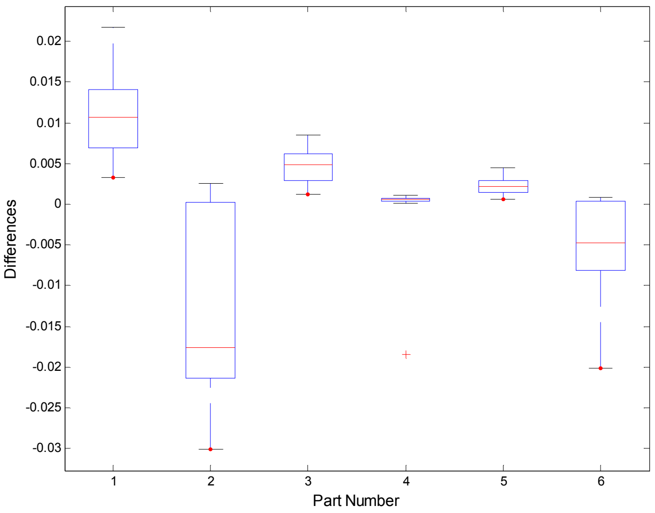

Figure 2. Boxplot of the differences between observations of the data set X and observations of the data set resulting from EM algorithm. (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

Despite the relative structure of the parts containing no zero values are preserved (black values in Table 4), we can observe in Table 4 that the values of MSD and STRESS are larger than the values in Table 3. Thus we can conclude that the distortion of the data structure of X has been larger than the distortion by the multiplicative replacement. This conclusion is confirmed when we compare the true values of the compositional geometric mean, total variability and the elements of the variation array (see Table 1) with the values in Table 4. Note that the values of var[ln(X_j/X_k)] (uppertriangle in variation matrix) give us underestimations of the true values. As can be deduced from Figure 2, the EM algorithm has mainly replaced the zero values by values that are larger than the true values. Thus, we can confirm that the EM algorithm not takes into account that the zero values should be replaced by "small" values.

The Sandford's method substitute censored data in a variable –data under the detection limit– by the mean value $\mu_{\text{missing data}}$ in this variable. This mean value is deduced from

$$\mu_{\text{mis sin g data}} = \frac{n\mu_{\text{whole data set}} - m\mu_{\text{observed data}}}{n - m},$$

where n is the size of the sample with m recorded values. The mean value $\mu_{\text{whole data set}}$ is the estimation of the mean value of the whole distribution produced by the EM algorithm.

When the Sandford's 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. As is showed in Martín, Palarea y Gómez (2003), the differences are caused by the different values of the mean $\mu_{\text{observed data}}$.

Table 5 shows descriptive measures of data set resulting from Sandford's method using Al_2O_3 as divisor in the alr-transformation. Note that the divergences between each selected divisor are not so large and are mainly concentrated on the logratio variances (uppertriangle of the variation array).

Table 5. Descriptive measures of data set resulting from Sandford's method (alr divisor: Al_2O_3).

Compositional geometric mean: (0.5580, 0.0330, 0.2394, 0.0279, 0.1228, 0.01899)						
Total variability: 0.4418						
Variation array						
k						
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5510	**0.1288**	0.1677	**0.0885**	0.4527
SiO_2	2.8292	0	0.5631	0.6177	0.5154	0.6290
Fe_2O_3	**0.8464**	-1.9827	0	0.1819	**0.1519**	0.4672
TiO_2	2.9946	0.1654	2.1482	0	0.2119	0.5063
H_2O	**1.5140**	-1.3152	**0.6676**	-1.4806	0	0.3926
Res_6	3.3851	0.5559	2.5386	0.3904	1.8711	0
MSD: 0.4676						
STRESS: 0.2403						

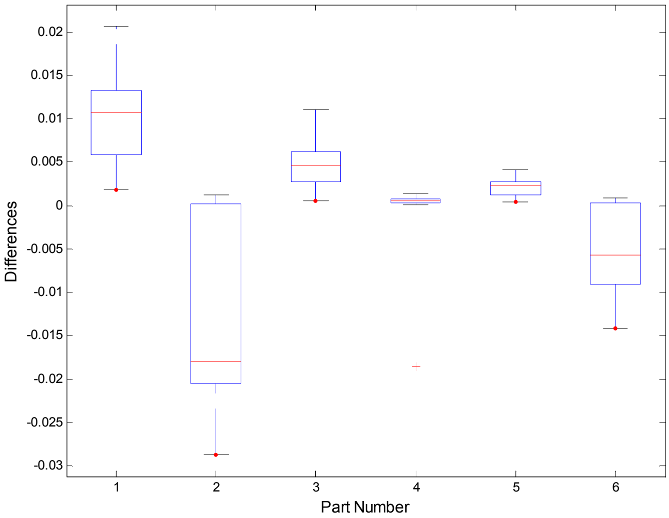

Figure 3. Boxplot of the differences between observations of the data set X and observations of the data set resulting from Sandford's method (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

As can be deduced from comparison of values in Table 4 and Table 5, and from comparison of Figure 2 and Figure 3, the EM algorithm and the Sandford's method give us very similar results. Note that *e.g.* the compositional geometric mean of the respectively resulting data sets is coincident. Thus, Sandford's method has the same behaviour than the EM algorithm: it not takes into account that the zero values should to be replaced by “small” values.

3.2.3 MCMC multiple imputation

The multiple imputation method generates, for each missing value, k simulated values obtained from the posterior predictive distribution of the missing part of the data set given the observed part. To simulate the k values in multivariate spaces, MCMC algorithms are needed (for more details see *e.g.* Schafer, 1997; Martín, Palarea y Gómez, 2003). Next, from each of the k “completed” data sets, we obtain k estimations of each quantity of interest. Finally, we combine the k estimations to obtain a unique global estimation and its combined variance using simple rules (Rubin, 1987) that can be reported as follows:

(i) Let be Q an unknown quantity of interest that we want to estimate. Let be $\hat{Q}$ its point estimator and U the associated variance of $\hat{Q}$. Then, after the simulation and estimation phase, we have k estimations $\{\hat{Q}_1,\ldots,\hat{Q}_k\}$ and its estimated variances $\{U_1,\ldots,U_k\}$.

(ii) The unique estimation ($\overline{Q}_k$) and its variance (T_k) are obtained by:

$$\overline{Q}_k = \frac{1}{k}\sum_{i=1}^{k}\hat{Q}_i \qquad \text{and} \qquad T_k = \overline{U}_k + \left(1+\frac{1}{k}\right)B_k\,,$$

where $\overline{U}_k = \sum_{i=1}^{k} U_i/k$, average of the variances $\{U_1,\ldots,U_k\}$, is known as the within-imputations variability and $B_k = \sum_{i=1}^{k}(\hat{Q}_i - \overline{Q})(\hat{Q}_i - \overline{Q})'/(k-1)$ is known as the between-imputations variability. Without missing data, B = 0 and $T = \overline{U}$.

In the same way as with the EM algorithm, we *alr*-transform our data set X*, and we apply MCMC for imputing the missing values in the real space, and we return to the simplex by the inverse transformation. We must assume a probability distribution model for our transformed data. As is suggested in Schafer (1997), for a real variables the most usual and robust hypothesis assume normality of our data. Additionally, we assume that our missing data are MAR. Only under MAR assumption the imputations generated by the posterior predictive distribution are proper. Several simulation studies (Collins, Schafer and Kam, 2001; Schafer and Graham, 2002; Palarea, Gómez and Martín, 2004) have showed the robustness of multiple imputation method across deviations from normality or from MAR assumption. Nevertheless, it is clear that the rounded compositional zeros are a NMAR situation.

Because of the low percentage of missing values in data set X*, four or five different imputations for each missing are sufficient to achieve an optimal degree of efficiency for the estimation (Rubin, 1987). In this work we decide to take k=5. Concerning the variances of the estimation, for the compositional geometric mean we calculate the estimated variances U_i from the clr-variance, that is, an usual estimated variance of the mean of the *clr*-transformed data. In order to simplify the analysis and illustrate the performance of multiple imputation, we measure the variability of the total variability and the variation array global estimations by its variances between the k imputed data sets, that is, considering only $(1+1/k)B_k$ the second factor of T_k. More research is needed in order to establish a method for calculating the estimated variances U_i for these statistics.

We have observed that the resulting descriptive measures are different depending on the selected part as divisor. Further studies are necessary in order to establish if this dependence is due to the selected divisor or only it is an effect of the simulation process. Nevertheless, since the divergences between the three cases are not so large we have decided only report here the case when the selected divisor in the *alr*-transformation is Fe_2O_3.

Table 6. Descriptive measures of data set resulting from MCMC multiple imputation when the selected divisor in the alr-transformation is Fe_2O_3 (see text for more details).

Compositional geometric mean [std. error]: (0.5578[0.0085], 0.0333[0.0297], 0.2393[0.0098], 0.0279[0.0130], 0.1227[0.0062], 0.0189[0.0218])						
Total variability: 0.5097 (std. dev.: 0.0077)						
Variation array (std. dev.)						
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.6017 (0.0062)	**0.1288**	0.1681 (0.0008)	**0.0885**	0.4771 (0.0076)
SiO_2	2.8175 (0.0155)	0	0.6112 (0.0036)	0.6638 (0.0059)	0.5702 (0.0058)	0.6903 (0.0151)
Fe_2O_3	**0.8464**	-1.9711 (0.0155)	0	0.1824 (0.0009)	**0.1519**	0.4913 (0.0056)
TiO_2	2.9942 (0.0006)	0.1767 (0.0161)	2.1477 (0.0006)	0	0.2121 (0.0006)	0.5319 (0.0074)
H_2O	**1.5140**	-1.3035 (0.0155)	**0.6676**	-1.4802 (0.0006)	0	0.4168 (0.0076)
Res_6	3.3835 (0.0059)	0.5660 (0.0166)	2.5370 (0.0059)	0.3893 (0.0060)	1.8695 (0.0059)	0
MSD[a]: 0.5319						
STRESS[b]: 0.1981						

(a) As Aitchison's distance $d_a(x_i, r_i)$ between the original composition $x_i \in X$ and the replaced composition r_i we have used the mean distance between each one of the imputed matrices and the original matrix.

(b) As Aitchison's distance $d_a(r_i, r_j)$ between the replaced compositions r_i we have used the mean distance matrix from the five imputed matrices.

Figure 4 shows the pattern of the differences between observations of the data set X and the mean values of the observations across the five imputed data sets resulting from MCMC multiple imputation method when the part selected as divisor is Fe_2O_3. Figures corresponding to the others divisors are not reported here since them are very similar and not informative. Analogously, we have decided not report here the figures corresponding to each of the five imputed data sets because the pattern of all of them is very similar to the Figure 4.

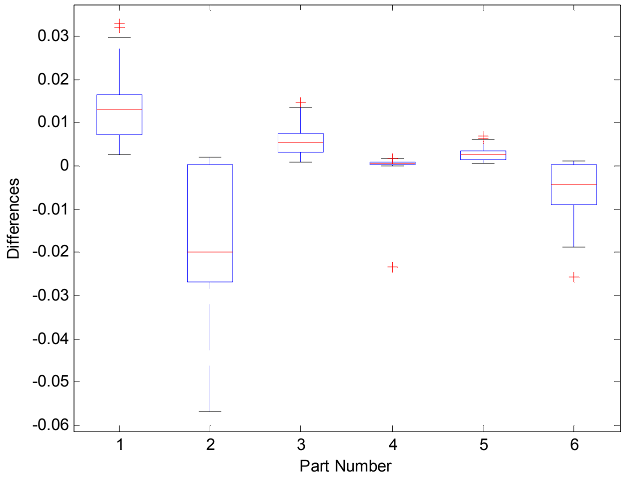

Figure 4. Boxplot of the differences between observations of the data set X and observations of the data set resulting from MCMC method (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

In addition to the problems related to compute the variability associated with the global estimators, the MCMC multiple imputation shows a similar behaviour to the rest of parametric methods of imputation in the sense that this strategy not takes into account that the missing values should be replaced by "small" values. This pattern is showed in Figure 4.

3.2.4 *Modified* EM algorithm for rounded compositional zeros

In this subsection the new algorithm introduced in section 2 is applied. We state that the resulting descriptive measures are slightly different depending on the selected part as divisor in the *alr*-transformation of X*. For comparison purposes, as in previous sections, we decide to use Fe_2O_3 as divisor. Nevertheless, it is important to remark that for the other divisors the produced results are not equal but their differences are of order up to 10^{-3} for the geometric mean, of order up to 10^{-1} for the total variability and the variation array, and of order 10^{-2} for the MSD and STRESS.

Table 7 shows the resulting descriptive measures. The method has a reasonable behaviour: the log-ratio variances are overestimated; the MSD and STRESS measures take values closer to zero than the other parametric methods; and the relative structure of the parts containing no zero values is preserved (black values in Table 7).

Table 7. Descriptive measures of data set resulting from modified EM algorithm when the selected divisor is Fe_2O_3 (see text for more details).

Compositional geometric mean: (0.5650, 0.0236, 0.2424, 0.0282, 0.1243, 0.0164)						
Total variability: 1.0507						
			Variation array			
			k			
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.9504	**0.1288**	0.1779	**0.0885**	0.6115
SiO_2	3.1747	0	0.9668	1.0255	0.9072	0.9581
Fe_2O_3	**0.8464**	-2.3283	0	0.1903	**0.1519**	0.6200
TiO_2	2.9978	-0.1769	2.1514	0	0.2203	0.6657
H_2O	**1.5140**	-1.6607	**0.6676**	-1.4838	0	0.5551
Res_6	3.5373	0.3626	2.6909	0.5395	2.0233	0
MSD: 0.0395						
STRESS: 0.0259						

In Figure 5 one can observe that this method takes into account that the rounded zeros must be replaced by “small” values. Some of these “small” values are greater than the true value and other “small” values are lower. The complete effect is symmetric boxplots around the zero value. The other parametric methods (*common* EM, Sandford’s, and MCMC) have not this suitable behaviour.

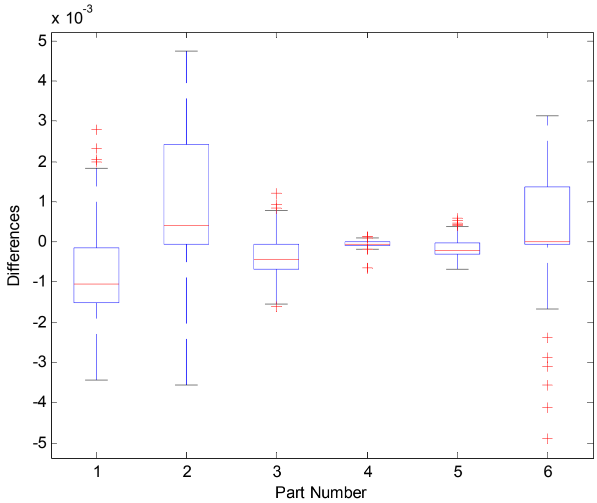

Figure 5. Boxplot of the differences between observations of the data set X and observations of the data set resulting from *modified* EM algorithm (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

Results from modified EM algorithm (Table 7, Fig. 5) are similar to the results produced by the multiplicative method (Table 3; Fig. 1). Note that in this example the amount of rounded zeros is small.

We have stated that if the detection limit is increased from 0.01% to 0.02% the modified EM algorithm produces better results than the multiplicative replacement. This fact suggests that for those compositional data sets with a large amount of rounded zeros it could be preferable to applying the modified EM algorithm than the multiplicative replacement. Nevertheless, further studies are needed in order to analyze this aspect.

4 Conclusions

In this work we have presented and compared the main parametric strategies for the rounded compositional zeros. All methods analyzed are coherent with the basic operations on the simplex. This coherence implies that the covariance structure of subcompositions with no zeros is preserved. Nevertheless, EM algorithm and multiple imputation method, in its standard formulation, not take into account that the rounded zeros should be replaced by “small” values, and this is an important deficiency.

As alternative we have introduced a modification of the EM algorithm applied in combination to the additive log-ratio transformation. This method takes into account that the imputed value must be lower than the detection limit of the part. Moreover, reasonable estimations and minimum distortion is produced. Also, this method improves its behaviour, in relation to the multiplicative replacement, in presence of large amount of null values.

These good features indicate that further research is needed in order to: complete and extend the multiple imputation method; complete and extend the *modified* EM algorithm.

Acknowledgements

This work has received financial support from the Dirección General de Investigación of the Spanish Ministry for Science and Technology through the project BFM2003-05640/MATE.

References

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

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics*, 24, 3–61.

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

Collins, L. M., Schafer, J. L. and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330–351.

Dempster, A. P., Laird N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc., Series B*, 39, 1–38.

Fry, J. M., Fry, T. R. L. and McLaren, K. R. (2000). Compositional data analysis and zeros in micro data. *Appl. Economics*, 32, 953 – 959.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475–492.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Assoc*., 88, 125–134.

Little, R. J. A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. Second edition. New York: Wiley & Sons.

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2000). Zero replacement in compositional data sets. In H. Kiers, J. Rasson, P. Groenen and M. Shader (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization,* Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), pp. 155–160. Berlin: Springer-Verlag.

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

Martín-Fernández, J. A., Palarea-Albaladejo, J. and Gómez-García, J. (2003). Markov chain Monte Carlo method applied to rounding zeros of compositional data: first approach. In S. Thió-Henestrosa and J.A. Martín-Fernández (Eds.), Proceedings of CODAWORK'03 - Compositional Data Analysis Workshop. ISBN 84-8458-111-X. Girona.

Palarea-Albaladejo, J. (2003). Algoritmos Monte Carlo basados en cadenas de Markov. Aplicación a la inferencia mediante imputación múltiple. Master thesis, Universidad de Murcia.

Palarea-Albaladejo, J., Gómez-García, J. and Martín-Fernández, J. A. (2004). Inferencia estadística con datos faltantes: análisis comparado de métodos según patrones y mecanismos de no respuesta. Proceedings of XXVIII National Congress of Statistics and Operations Research, Cádiz. ISBN 84-609-0438-5.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley & Sons.

Rubin, D. B., Stern, H. and Vehovar, V. (1995). Handling "don´t know" survey responses: the case of the Slovenian plebiscite. *J. Am. Statist. Assoc.*, 90, 822–828.

Sandford, R.F., Pierson, C.T., and Crovelli, R.A. (1993). An objective replacement method for censored geochemical data. *Mathematical Geology*, 25(1), 59–80.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147–177.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology*, 31(5), 491–504.

Zhou, D. (1997). Logratio statistical classification and estimation of hydrodynamic parameters from Darss Sill grain-size data. In V. Pawlowsky-Glahn (Eds.), Procedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology, 1, International Center for Numerical Methods in Engineering, pp. 139–144. Barcelona.