MARKOV CHAIN MONTE CARLO METHOD APPLIED TO ROUNDING ZEROS OF COMPOSITIONAL DATA: FIRST APPROACH

J. A. Martín-Fernández[1], J. Palarea-Albaladejo[2], J. Gómez-García[3]
[1]Universitat de Girona, Departament d'Informàtica i Matemàtica Aplicada, 17071 Girona, Spain, josepantoni.martin@udg.es
[2]Universidad Católica San Antonio, Departamento de Informática de Sistemas, 30107 Murcia, Spain, jpalarea@pdi.ucam.edu
[3]Universidad de Murcia, Departamento de Métodos Cuantitativos, 30100 Murcia, Spain, jgomezg@um.es

1 Introduction

As stated in Aitchison (1986), a proper study of relative variation in a compositional data set should be based on logratios, and dealing with logratios excludes dealing with zeros. Nevertheless, it is clear that zero observations might be present in real data sets, either because the corresponding part is completely absent –*essential zeros*– or because it is below detection limit –*rounded zeros*. Because the second kind of zeros is usually understood as "a trace too small to measure", it seems reasonable to replace them by a suitable small value, and this has been the traditional approach. As stated, *e.g*. by Tauber (1999) and by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000), the principal problem in compositional data analysis is related to rounded zeros. One should be careful to use a replacement strategy that does not seriously distort the general structure of the data. In particular, the covariance structure of the involved parts –and thus the metric properties– should be preserved, as otherwise further analysis on sub-populations could be misleading. Following this point of view, a non-parametric imputation method is introduced in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000). This method is analyzed in depth by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003) where it is shown that the theoretical drawbacks of the additive zero replacement method proposed in Aitchison (1986) can be overcome using a new multiplicative approach on the non-zero parts of a composition. The new approach has reasonable properties from a compositional point of view. In particular, it is "natural" in the sense that it recovers the "true" composition if replacement values are identical to the missing values, and it is coherent with the basic operations on the simplex. This coherence implies that the covariance structure of subcompositions with no zeros is preserved. As a generalization of the multiplicative replacement, in the same paper a substitution method for missing values on compositional data sets is introduced.

Among parametric techniques to treat the inference problem in presence of missing data (Allison, 2001; Little and Rubin, 2002; Schafer, 1997) we find several methods, but the EM algorithm (Dempster, Laird and Rubin, 1977), and its extensions, and the multiple imputation method (Rubin, 1987) represent the most based approaches and the actual state of the art. They all provide a set of flexible and reliable tools for inference in large classes of missing-data problems, and rely on fully parametric models for multivariate data, usually the normal distribution. Using the additive logratio transformation, Buccianti and Rosso (1999) describes from an empirical point of view the performances of the EM algorithm and of the Sandford's method. This method was proposed in Sandford, Pierson and Crovelli (1993) as an extension of the EM algorithm to 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

$$\mu_{\text{missing data}} = \frac{n\mu_{\text{whole data set}} - m\mu_{\text{observed data}}}{n-m}, \tag{1}$$

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.

Gómez and Palarea (2003) review the most important features of the multiple imputation techniques and analyze the role of Markov Chain Monte Carlo (MCMC) simulation algorithms within this methodology applied to missing data in a real space. In the next section we describe briefly the multiple imputation method via MCMC. Next we present a numerical example in order to illustrate the performance of this method and of all above methods, to process rounded zeros of a compositional data set. The aim of this work is to do a first approach to the "zeros problem" using parametric tools.

2 Multiple imputation via Markov Chain Monte Carlo simulation

The goal of the MCMC method (Hastings, 1970) is to generate values of a random vector X with probability distribution $\pi(x)$, often multidimensional. With them we can calculate Monte Carlo approximations of any quantity expressed by an integral, generally with no analytical solution. In order to do this, an ergodic Markov chain with $\pi(x)$ is generated as stationary distribution. Then, under mild conditions and after a large number of transitions of the chain, we obtain approximate samples from $\pi(x)$. In the missing data context, we use MCMC algorithms to generate values of the missing part of the problem. Namely, we use the *Gibbs sampler* in a particular form known as *data augmentation* algorithm (Tanner and Wong, 1987). Several recent references can be consulted for a detailed description of these algorithms, the theoretical background and practical aspects: Geyer (1992); Gilks, Richardson and Spiegelhalter (1996); Robert and Casella (1999), Tierney (1994) and others.

Recently (Schafer, 1997) the MCMC algorithm has been adapted to resolve an important practical problem of the multiple imputation method. That is, to impute missing values we need to simulate independent realizations of $P[X_{miss} \mid X_{obs}]$, the posterior predictive distribution of the missing part of our sample under some complete-data model and a prior distribution for the vector q of parameters. We can write

$$P[X_{miss} \mid X_{obs}] = \int P[X_{miss} \mid X_{obs}, q]P[q \mid X_{obs}]dq \quad , \tag{2}$$

where X_{miss} and X_{obs} represent, respectively, the missing and observed part of our sample. In practical situations, especially in multidimensional contexts, to generate the imputed values by (2) is a complex question. Then, simulating MCMC samples of (2) using an adapted data augmentation algorithm, the task is acceptably solved for the majority of problems.

In general, a multiple imputation technique can be synthesized as follows:

(i) Firstly, each missing value is replaced by a set of k>1 simulated values. Then we have k "completed" data sets.

(ii) Next, we apply to each data set the statistical technique of our interest (linear regression, discriminant analysis, …) and we obtain k estimations of each quantity of interest.

(iii) Finally, using *simple rules* we combine the k estimations to obtain a unique global estimation.

In the first phase, similarly that it happens in other parametric imputation techniques, we must assume a probability distribution model for our data. As is suggested in Schafer (1997), for a real variables the most usual and robust hypothesis assume normality of our data. In addition, it is crucial to establish the mechanism of missingness. In relation to this mechanism the missing data can be classified as: MAR (missing at random), MCAR (missing completely at random), and NMAR (not missing at random). In the first case, MAR, the probability that an observation is missing may depend on observed part of the data but not on missing part of the data. MCAR is a particular case of MAR because MCAR requires that the missing data values are a simple random sample of all data values. Finally, NMAR consider that the probability that an observation is missing may depend on the unobserved part of the data. That is the mechanism of missingness is nonignorable. As is exposed in Schafer (1997), "models for nonignorable response appropriate for wide classes of data have not been proposed" and "construction and evaluation of general models for nonignorable nonresponse are an important area for future study". In this paper we work assuming that our missing data are MAR. Only under MAR assumption the imputations generated by (2) are proper. Several simulation studies (*e. g.* Graham and Schafer, 1999; Collins, Schafer and Kam, 2001) have showed the robustness of multiple imputation method across deviations from normality or MAR assumption. But it is clear that, as an open question, it appears the convenience to extend the analysis to NMAR and nonnormal situations.

In front of the rest of imputation methods (simple substitution, EM algorithm, …), the multiple imputation incorporates the additional source of uncertainty associate to the presence of missing values as

an observed variability in the k completed sets. Furthermore, is easy to show that in data sets with a moderate amount of missing values, a reduced number of different imputations (k=5) are enough to obtain efficient results. The estimations obtained from the k completed data sets are combined using simple rules proposed by Rubin (1987). These rules can be reported as follows:

(i) Let be Q an unknown quantity of interest that we want to estimate applying some statistical technique (*e.g.* beta coefficient in linear regression model). Let be Q* its estimation and U the associated variance of the estimation. Then, after the simulation phase, we have k estimations $Q^*_1, Q^*_2, \ldots, Q^*_k$ and its variance-covariance matrix $U_1, U_2, \ldots, U_k$.

(ii) To obtain a unique estimation we use the following expressions

$$\bar{Q}_k = \frac{\sum Q^*_i}{k}, \quad T_k = \bar{U}_k + \left(1 + \frac{1}{k}\right) B_k, \quad B_k = \frac{\sum (Q^*_i - \bar{Q}_k)(Q^*_i - \bar{Q}_k)'}{(k-1)}, \tag{3}$$

where T_k is called the variance-covariance matrix of the estimations. Here, the matrix $\bar{U}_k$, average of the variances $U_1, U_2, \ldots, U_k$, is known as the variability within-imputations and B_k is known as the variability between-imputations. By this expressions the uncertainty due to missing data are incorporated to the estimation.

Extensive studies incorporating new and more complex rules and analysing the asymptotic behaviour of the estimators can be found in Meng and Rubin (1992), Rubin (1987, 1996) or Robins and Wang (2000).

3 Case study

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 $(100-(Al_2O_3+\ldots+H_2O))\%$. 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,…,6; k=j,…,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=1,2,…,6; k=1,…,j). 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: $\hat{\xi} = (0.5644,\ 0.0246,\ 0.2421,\ 0.0282,\ 0.1242,\ 0.0166)$ and $\text{totvar}(X) = 0.9718$.

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.

Amount of obs.	Pattern of missing values						Amount of observed obs. if...(a)
	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6	
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, the zeros have to be replaced. Our aim is to compare the performance of the different imputation techniques: non-parametric multiplicative replacement (proposed in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 2003), EM algorithm, Sandford's method, and MCMC multiple imputation (presented in Gómez and Palarea, 2003).

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.

As it is well known (Aitchison, 1986; Buccianti and Rosso, 1999; Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 2003) and it is confirmed here, when not suitable imputation technique is applied to compositional data the general structure of the data can be seriously distorted. Broadly speaking, it's likely that the imputed values will be negatives and that the sum constraint will be not preserved. That's happens, *e.g.* when one apply EM algorithm, Sandford's method or MCMC multiple imputation to compositional data without previous logratio transformation. In this case, Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2003) show that the best results are obtained when the rounded zeros in X* are replaced by the "small" value δ=0.0065 using the multiplicative replacement. 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.

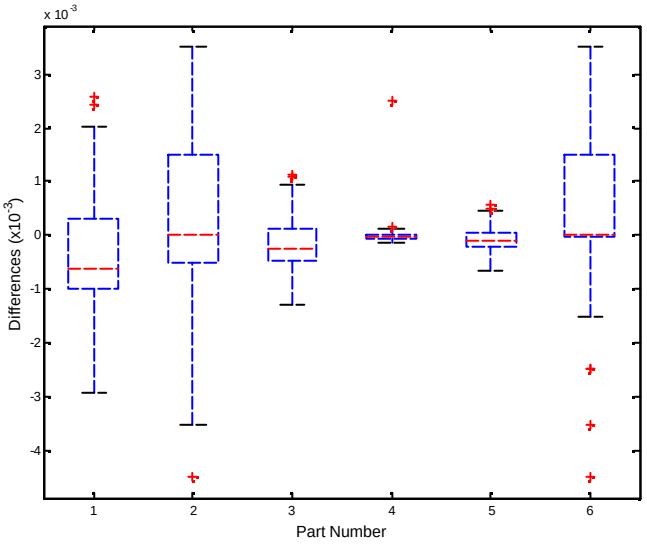

Figure 1. Boxplot of the differences between observations of the data set X and observations of the data set resulting from multiplicative replacement with δ=0.0065. (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

Following Aitchison (1986) if we transform the data set X* using the additive logratio transformation (alr) then we can consider parametric models for our transformed data since the transformed data belongs to a real space. Thus, we transform the data set X* and we assume the normal distribution for the alr-transformed data, as a previous phase to applying a parametric technique of imputation. Here, the alr-transformed values of zero are considered as missing values. 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. These results have been obtained using the procedure integrated into the SPSS® package (release 11.5 for Windows).

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						

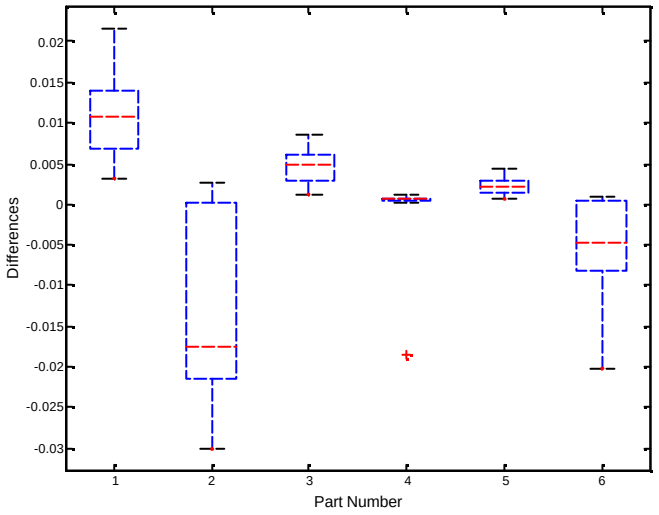

Figure 2. Boxplot of the differences between observations of the data set X and observations of the data set resulting from EM algorithm. (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

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, which 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.

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. Certainly, as it is showed in Table 5, when we analyze the members of the expression (1) we confirm that the mean $\mu_{\text{whole data set}}$ is invariant under the group of permutations of the parts since this mean is produced by the EM algorithm. Thus, we conclude that the differences must to be caused by the different values of the mean $\mu_{\text{observed data}}$ (see Table 5).

Table 5. Amount of observations without zero and with missing values depending on the selected divisor (**Div.**). Compositional geometric mean of the observed data and of the whole data set.

	Al_2O_3 (**Div.**)	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Amount observ. (m)	--	262	332	331	332	275
Amount miss. (n-m)	--	70	0	1	0	57
Comp. geom. $mean_{obs.data}$	0.55655	0.03483	0.23872	0.02786	0.12246	0.01958
Comp. geom. $mean_{whole data}$	0.55805	0.03296	0.23937	0.02793	0.12279	0.01890
	Al_2O_3	SiO_2	Fe_2O_3 (**Div.**)	TiO_2	H_2O	Res_6
Amount observ. (m)	332	262	--	331	332	275
Amount miss. (n-m)	0	70	--	1	0	57
Comp. geom. $mean_{obs.data}$	0.55643	0.03507	0.23868	0.02785	0.12243	0.01955
Comp. geom. $mean_{whole data}$	0.55805	0.03296	0.23937	0.02793	0.12279	0.01890
	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O (**Div.**)	Res_6
Amount observ. (m)	332	262	332	331	--	275
Amount miss. (n-m)	0	70	0	1	--	57
Comp. geom. $mean_{obs.data}$	0.55695	0.03431	0.23890	0.02788	0.12255	0.01941
Comp. geom. $mean_{whole data}$	0.55805	0.03296	0.23937	0.02793	0.12279	0.01890

A simple overview of the results in Table 5 show us that the $\mu_{\text{observed data}}$ of variable $\ln(SiO_2/Fe_2O_3)$, *i.e.* $\hat{E}[\ln(SiO_2/Fe_2O_3)]$ for the observed data (m=262; Table 5), takes the value −1,91783437. Nevertheless, if we consider the estimation $\hat{E}[\ln(SiO_2/Al_2O_3)]$ (m=262; Table 5) and the estimation $\hat{E}[\ln(Fe_2O_3/Al_2O_3)]$ (m=332; Table 5) and we calculate the subtraction $\hat{E}[\ln(SiO_2/Al_2O_3)]-\hat{E}[\ln(Fe_2O_3/Al_2O_3)]$ we obtain the value −1,92469852. Analogously, we calculate $\hat{E}[\ln(SiO_2/H_2O)]-\hat{E}[\ln(Fe_2O_3/H_2O)]$ and obtain the value −1,94072832. The cause of this divergence lies in the different amount of observations that we use to calculate the estimation of the mean depending on the selected divisor (see Table 5). Thus, in order to eliminate the "divisor effect" we would have to estimate $\mu_{\text{observed data}}$ using the subset of observations without zero values in all parts. Obviously, this procedure will give us lower efficiency in the estimation.

Table 6 shows descriptive measures of data set resulting from Sandford's method for each selected divisor in the alr-transformation. Note that the divergences between the three cases are not so large and are mainly concentrated on the logratio variances (uppertriangle of the variation array).

Table 6. Descriptive measures of data set resulting from Sandford's method.

Part as divisor in alr-transformation: 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						

Part as divisor in alr-transformation: Fe_2O_3						
Compositional geometric mean: (0.5580, 0.0330, 0.2394, 0.0279, 0.1228, 0.01899)						
Total variability: 0.4432						
Variation array						
	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5534	**0.1288**	0.1677	**0.0885**	0.4555
SiO_2	2.8292	0	0.5608	0.6188	0.5182	0.6302
Fe_2O_3	**0.8464**	-1.9828	0	0.1819	**0.1519**	0.4645
TiO_2	2.9946	0.1654	2.1482	0	0.2119	0.5062
H_2O	**1.5140**	-1.3152	**0.6676**	-1.4806	0	0.396
Res_6	3.3851	0.5558	2.5386	0.3904	1.8711	0
MSD: 0.4685						
STRESS: 0.2393						
Part as divisor in alr-transformation: H_2O						
Compositional geometric mean: (0.5580, 0.0330, 0.2394, 0.0279, 0.1228, 0.01899)						
Total variability: 0.4412						
Variation array						
	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5512	**0.1288**	0.1677	**0.0885**	0.4532
SiO_2	2.8292	0	0.5635	0.6182	0.5152	0.6230
Fe_2O_3	**0.8464**	-1.9828	0	0.1819	**0.1519**	0.4678
TiO_2	2.9946	0.1654	2.1482	0	0.2119	0.5083
H_2O	**1.5140**	-1.3152	**0.6676**	-1.4806	0	0.3921
Res_6	3.3851	0.5559	2.5386	0.3904	1.8711	0
MSD: 0.4695						
STRESS: 0.2403						

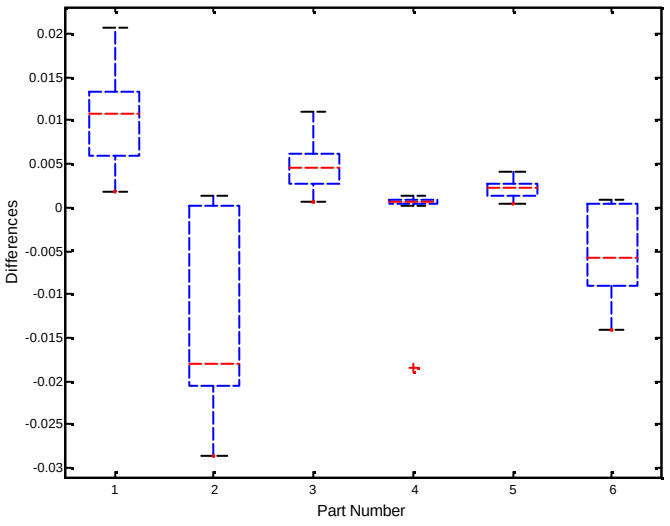

Figure 3. Boxp lot of the differences between observations of the data set X and observations of the data set resulting from Sandford's method (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

As can be deduced from comparison of values in Table 4 and Table 6, 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.

Finally, we apply the MCMC algorithm to perform multiple imputation of zeros values of data set X*. Obviously, as previous phase, we alr-transform our data and we consider the alr-transformed values of zeros as a missing values. Then, we must to decide which part is selected as divisor for the transformation and for each different divisor we will have to analyze the obtained results. In addition, following Gómez and Palarea (2003), because the low percentage of missing values in data set X* and the quick improve on efficiency with a few simulations, we decide to make 4 different imputations for each missing. In order to make comparisons, we consider two possibilities: first, apply a simple rule exposed above in expression (3) to the descriptive measures; second, make the compositional geometric mean of the different imputation and calculate its descriptive measures. We know that this second possibility is not so efficient than the first one. Nevertheless, we present here its results to better illustrate the analysis. Moreover, in order to simplify the analysis, in the first strategy we consider the simple rule consisting on calculate the mean and the standard deviation of the set of 4 descriptive measures of different imputations.

Table 7 shows the results produced following the first strategy, *i.e.* shows the mean and the standard deviation of the descriptive measures of data set resulting from multiple imputation with MCMC algorithm where the selected divisor in the alr-transformation is Fe_2O_3. 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 part Fe_2O_3 is the divisor. Table 8 shows the results produced following the second strategy. A simple comparison of both tables with the true descriptive measures (Table 1) shows us that the first strategy is more efficient than the second since the first strategy give us better estimations and incorporates the variability of these estimations. Figure 4 shows the pattern of the differences between observations of the data set X and observations of the data set resulting from the second strategy of the 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 the each multiple imputation because the pattern of all of them is very similar to the Figure 4.

Table 7. Mean and standard deviation of descriptive measures of data set resulting from MCMC method following the first strategy when the selected divisor is Fe_2O_3 (see text for more details).

Compositional geometric mean: (0.5580, 0.0330, 0.2394, 0.0279, 0.1228, 0.0189)

Total variability: mean: 0.5235; st.dev. 0.0144

Variation array (mean)

	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.6111	**0.1288**	0.1679	**0.0885**	0.4826
SiO_2	2.8272	0	0.6240	0.6750	0.5783	0.6965
Fe_2O_3	**0.8464**	-1.9807	0	0.1821	**0.1519**	0.4989
TiO_2	2.9945	0.1673	2.1481	0	0.2120	0.5368
H_2O	**1.5140**	-1.3131	**0.6676**	-1.4805	0	0.4227
Res_6	3.3877	0.5605	2.5412	0.3932	1.8737	0

Variation array (st.dev.)

	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.0118	**0**	0.0002	**0**	0.0053
SiO_2	0.0136	0	0.0079	0.0108	0.0126	0.0173
Fe_2O_3	**0**	0.0137	0	0.0002	**0**	0.0059
TiO_2	0.0004	0.0133	0.0004	0	0.0001	0.0072
H_2O	**0**	0.0137	**0**	0.0004	0	0.0052
Res_6	0.0084	0.0127	0.0084	0.0081	0.0084	0

MSD: mean 0.5450; st.dev. 0.0428

STRESS: mean 0.2109; st.dev. 0.0070

Table 8. Descriptive measures of data set resulting from MCMC method following the second strategy when the selected divisor is Fe_2O_3 (see text for more details).

Compositional geometric mean: (0.5580, 0.0330, 0.2394, 0.0279, 0.1228, 0.0189)						
Total variability: 0.4655						
Variation array						
	k					
j	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Al_2O_3	0	0.5664	**0.1288**	0.1677	**0.0885**	0.4655
SiO_2	2.8272	0	0.5803	0.6348	0.5310	0.6427
Fe_2O_3	**0.8464**	-1.9807	0	0.1820	**0.1519**	0.4825
TiO_2	2.9945	0.1673	2.1481	0	0.2118	0.5215
H_2O	**1.5140**	-1.3132	**0.6676**	-1.4805	0	0.4031
Res_6	3.3877	0.5605	2.5412	0.3932	1.8737	0
MSD: 0.4870						
STRESS: 0.2243						

A simple comparison of the values in Table 7 and Table 8 with the true values of the descriptive measures of data set X (Table 1), confirm us that we obtain better estimations when the simple rule is applied. Nevertheless, note that this better behaviour is mainly in relation of the variability measures, since the measures corresponding to the compositional geometric mean and corresponding to the estimations $\hat{E}[\ln(X_j/X_k)]$ the results are coincident. The MCMC multiple imputation shows a similar behaviour to the rest of parametric methods of imputation, in the sense that this algorithm not takes into account that the missing values should be replaced by "small" values. This pattern is showed in Figure 4.

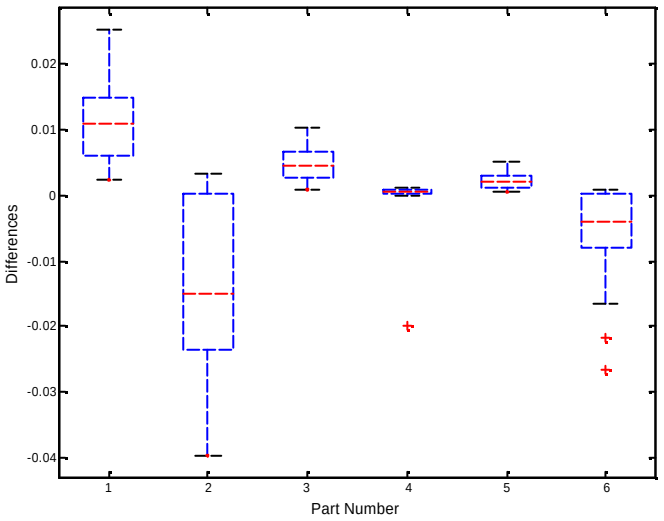

Figure 4. Boxplot of the differences between observations of the data set X and observations of the data set resulting from MCMC method following the second strategy (Part Number: 1.Al_2O_3; 2.SiO_2; 3.Fe_2O_3; 4.TiO_2; 5.H_2O; 6.Res_6).

4 Conclusions

In this work, it is shown that multiple imputation via MCMC is a suitable tool to replace rounded zeros in compositional data sets. In a different way that EM algorithm or that Sandford's method, the MCMC multiple imputation algorithms incorporates information about the variability of the estimations due to missing part. This information is very useful in order to analyze the efficiency of these estimations. All methods analyzed in this work are coherent with the basic operations on the simplex. This coherence

implies that the covariance structure of subcompositions with no zeros is preserved. Nevertheless, in contrary to the non-parametric multiplicative replacement, all the parametric methods not take into account in its procedure that the rounded zeros should be replaced by "small" values. Future studies will turn our effort to introduce this restriction in MCMC multiple imputation and incorporate the more appropriate, in this context, NMAR assumption.

Acknowledgments

This research has been partially financed by the Direcció General de Enseñanza Superior e Investigación Científica (DGESIC) of the Spanish Ministry for Education and Culture through the project BFM2000-0540; and partially financed by the Direcció General de Recerca de la Generalitat de Catalunya through the project 2001XT00057.

References

Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, London, 416 p.

Allison, P.D., 2001, Missing Data: Thousand Oaks (USA): Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136, 93 p.

Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V., 1999, Comment on 'Singularity and Nonnormality in the Classification of Compositional Data' by G.C. Bohiling et al.: Mathematical Geology, v. 31 no. 5, p. 581–586.

Buccianti, A. and Rosso, F., 1999, A new approach to the statistical analysis of compositional (closed) data with observations below the "detection limit": Geoinformatica, v. 3, p. 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, p. 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): Journal of the Royal Statistical Society, Series B, 39, p. 1-38.

Geyer, C. J., 1992, Practical Markov chain Monte Carlo: Statistical Science, 7, p. 473–511.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., 1996, Markov Chain Monte Carlo in Practice: Chapman &Hall.

Gómez García, J. and Palarea Albaladejo, J., 2003, Algoritmos Monte Carlo basados en cadenas de Markov aplicados a la imputación múltiple de datos faltantes: Actas del 27 Congreso Nacional de Estadística e Investigación Operativa Lleida, 8–11 de abril de 2003, CD-ROM publication, p. 1691-1705.

Graham, J. W. and Schafer, J. L., 1999, On the performance of multiple imputation for multivariate data with small sample size: Statistical Strategies for Small Sample Research, Thousand Oaks: Sage, p. 1-29.

Hastings, W. K., 1970, Monte Carlo sampling methods using Markov chains and their applications: Biometrika, 57, p. 97-109.

Little, R.J.A. and Rubin, D.B., 2002, Statistical Analysis with Missing Data: John Wiley and Sons, second edition, New York, 381 p.

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets: in Kiers, H., Rasson, J., Groenen, P., and Shader, M., eds., Studies in

Classification, Data Analysis, and Knowledge Organization, Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000), University of Namur, Namur: Springer-Verlag, Berlin (D), p. 155–160.

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2003, Dealing with Zeros and Missing Values in Compositional Data Sets: Mathematical Geology, v. 35, no. 3, p. 253–278

Meng, X. L. and Rubin, D. B., 1992, Performing likelihood ratio test with multiply imputed data sets: Biometrika, 79, p. 103-111.

Robert, C. P. and Casella, G., 1999, Monte Carlo Statistical Methods: Springer.

Robins, J. M. and Wang, N., 2000, Inference for imputation estimators: Biometrika, v. 87, p. 113–124.

Rubin, D. B., 1987, Multiple Imputation for Nonresponse in Surveys: Wiley & Sons.

Rubin, D. B., 1996, Multiple imputation after 18+ years: Journal of the American Statistical Association v. 91, p. 473–489.

Sandford, R.F., Pierson, C.T., and Crovelli, R.A., 1993, An objective replacement method for censored geochemical data: Mathematical Geology, v. 25, no. 1, p. 59–80.

Schafer, J. L., 1997, Analysis of Incomplete Multivariate Data: Chapman and Hall, London, 430 p.

Tanner, M. and Wong, W., 1987, The calculation of posterior distributions by data augmentation: Journal of the American Statistical Association, 82, p. 528-550.

Tierney, L., 1994, Markov chains for exploring posterior distributions: The Annals of Statistics, 22 (4), p. 1701-1762.

Tauber, F., 1999, Spurious clusters in granulometric data caused by logratio transformation: Mathematical Geology, v. 31, no. 5, p. 491–504.