Mathematical Geology, Vol. 35, No. 3, April 2003 (© 2003)

Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation[1]

J. A. Martín-Fernández,[2] C. Barceló-Vidal,[2] and V. Pawlowsky-Glahn[2]

The statistical analysis of compositional data based on logratios of parts is not suitable when zeros are present in a data set. Nevertheless, if there is interest in using this modeling approach, several strategies have been published in the specialized literature which can be used. In particular, substitution or imputation strategies are available for rounded zeros. In this paper, existing nonparametric imputation methods—both for the additive and the multiplicative approach—are revised and essential properties of the last method are given. For missing values a generalization of the multiplicative approach is proposed.

KEY WORDS: Aitchison distance, detection limit, logratio transformation, simplex, stress, threshold.

INTRODUCTION

To understand the "zero problem" related to compositional data, we must understand the nature of this type of data. As stated in Aitchison (1986), the sample space of compositional data is the simplex $\mathcal{S}^D$ defined as

$$\mathcal{S}^D = \{[x_1, x_2, \ldots, x_D] : x_j > 0; j = 1, 2, \ldots, D; x_1 + x_2 + \cdots + x_D = c\}, \tag{1}$$

where c can be 100, 1, 10^6, or any other constant depending on the units of measurement. Although the value of c is irrelevant from a mathematical point of view, and was therefore set equal to 1 in the original definition, we keep it to avoid confusion in daily practice.

The definition of $\mathcal{S}^D$ in (1) reflects two characteristics of compositional observations. One, which receives general agreement, is that they are proportions of

[1]Received 1 May 2002; accepted 19 February 2003.

[2]Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi P4, E-17071, Girona, Spain, e-mail: josepantoni.martin@udg.es.

some whole and therefore positive and of constant sum. The second, frequently discussed, is associated to the metric, which is considered to be appropriate for this type of data. We agree with Aitchison (1986) in that compositional data reflect only relative magnitude, and thus interest lies in relative—and not absolute—changes. The usual Euclidean metric in real space measures absolute changes, whereas relative changes can be measured using some logarithmic scale. Thus, a proper study of relative variation in a data set can be based on logratios, and dealing with logratios excludes dealing with zeros, hence the condition of observations are strictly positive. Nevertheless, it is clear that zero observations might be present in real data sets, either because the corresponding part is completely absent, or because it is below detection limit. Therefore, a strategy—compatible with our perception of compositional data—is needed on how to deal with zeros in a given data set.

To develop this strategy we first introduce basic concepts related to the vector space structure of the simplex, proceed afterwards with the classification of zeros into essential zeros and rounded zeros, and then we review existing replacement methods for rounded zeros, to analyze their properties and to compare their behavior. To situate the reader, we start with a summary of the most usual nonparametric approaches for missing values with noncompositional data. With the main features of these approaches in mind, we revise first, from a theoretical point of view, the additive replacement method suggested by Aitchison (1986), whose drawbacks have been described from an empirical point of view by Tauber (1999) in a hierarchical cluster analysis context. Next, we present the multiplicative replacement method suggested in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000) and we analyze its properties. In particular, it is shown that the simple replacement method, which is frequently used in experimental sciences as an heuristic strategy, is just an equivalent form of multiplicative replacement. To illustrate the methods, we present three examples where we compare the behavior of the approach proposed in Aitchison (1986) with the alternative approach presented in this paper. Finally, we propose a generalization of the multiplicative replacement of zeros to the substitution of missing values in compositional data sets.

LINEAR VECTOR SPACE STRUCTURE OF THE SIMPLEX

As stated in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000), it is easy to see that the *perturbation operation*, $\mathbf{p} \oplus \mathbf{x} = \mathcal{C}[p_1 x_1, p_2 x_2, \ldots, p_D x_D]$, defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation*, $\alpha \otimes \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]$, defined on $\mathbb{R} \times \mathcal{S}^D$, induce a vector space structure in the simplex. In both operations we consider the *closure operator* "$\mathcal{C}$" defined by $\mathcal{C}(\mathbf{w}) = c[w_1/\sum w_j, w_2/\sum w_j, \ldots, w_D/\sum w_j]$, where $\mathbf{w} \in \mathbb{R}_+^D$. Once we have a vector space structure, we need a compatible distance, as in most statistical techniques a distance between observations is explicitly or implicitly assumed. A distance, compatible with the

vector space structure of the simplex, is the Aitchison distance

$$d_a(\mathbf{x}, \mathbf{x}^*) = d_e(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{x}^*)), \tag{2}$$

where $\mathbf{x}, \mathbf{x}^*$ are compositions in $\mathcal{S}^D$, d_e represents the Euclidean distance in $\mathbb{R}^D$, and the centered logratio transformation is defined by $\mathrm{clr}(\mathbf{x}) = [\ln(x_1/g(\mathbf{x})), \ldots, \ln(x_D/g(\mathbf{x}))]$, with $g(\mathbf{x}) = (\prod_{j=1}^{D} x_j)^{1/D}$. The simplex, with the perturbation operation, the power transformation and the Aitchison distance, is then a linear vector space. See Aitchison (2002), Billheimer, Guttorp, and Fagan (2001) and Pawlowsky-Glahn and Egozcue (2001, 2002) for further details.

The properties of the Aitchison distance (2) have been extensively discussed in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998a) and in Aitchison and others (2000). Nevertheless, this distance has its "Achilles heel." The presence of zero values in a data set prevents its application, as it makes it impossible to consider all the ratios $(x_j/x_k, 1 \leq j < k \leq D)$. Hence we have a powerful mathematical structure on which to build on, and one weakness associated to the presence of zeros in a data set, thus justifying the need for replacement strategies.

Another important concept in our approach is the concept of subcomposition. If our interest is focused on some of the parts of the composition we must select these parts and form their *subcomposition*. The subcomposition $\mathbf{x}_S \in \mathcal{S}^S$ of a composition $\mathbf{x} \in \mathcal{S}^D$ is defined as $\mathbf{x}_S = \mathcal{C}(\mathbf{S}\mathbf{x})$, where $\mathbf{S}$ is a $(S \times D)$ selecting matrix with all elements zero except one in each row and at the most one in each column (Aitchison, 1986). Note that a subcomposition preserves the ratios between the selected parts.

Before proceeding, note that the definition of a proper sample space for compositional data can be based on the concept of classes of equivalence. In fact, a compositional observation is nothing else but a ray from the origin into the positive orthant of D-dimensional real space $\mathbb{R}_+^D$ (Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn, 2001). Any point $\mathbf{w} \in \mathbb{R}_+^D$ on one of those rays can be projected on an arbitrary surface, as long as the projection is one to one. A possible surface is given by the hyperplane containing $\mathcal{S}^D$, thus reducing $\mathcal{S}^D$ to a representation of those classes. Another possible representation would be given by an hyperbolic surface; this representation, although unusual, is important for interpretation of oncoming results (Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn, 2001). Projection on $\mathcal{S}^D$ is performed applying the *closure operator* "$\mathcal{C}$" to $\mathbf{w} \in \mathbb{R}_+^D$. In general $\mathbf{w} \neq \mathcal{C}(\mathbf{w})$, and this property will hold for any other representation as well. Nevertheless, despite the fact that the representation might change, the fact that the important issue is quantification of relative differences will not.

ESSENTIAL ZEROS AND ROUNDED ZEROS

In compositional data analysis we distinguish two kinds of zeros: *essential zeros*—or absolute absence of the part in the observation—and *rounded zeros*—or

presence of a component, but below detection limit. Its different nature implies that the treatment of the two kinds of zeros should be different. Essential zeros are present in many applications. For example, in an analysis of an household budget, as described in Aitchison (1986), we can find some family with zero value on the commodity group "tobacco and alcohol." For many problems which require a statistical approach, like factor analysis or hierarchical cluster analysis, it seems reasonable to interpret the presence of an essential zero in a part either as an indication that the composition belongs to a different group or population, or as an indication that the component has no significance for the purpose of the study. Thus, in the above example relative to the household budget, we would either divide the sample in households that spend or spend not some part of their budget in tobacco or alcohol, or we would amalgamate the proportion spend in this commodity group to another commodity group. If the approach assumed as suitable is to divide the sample, then a statistical analysis of any kind would be applied to each subsample separately. Otherwise, after amalgamation of different parts, the analysis would be performed on the full data set.

On the contrary, when we consider that one component has a rounded zero, i.e. when we consider that this value denotes not that the part is completely absent, but rather that no quantifiable proportion could be recorded to the accuracy of the measurement process, then it seems not sensible to divide the sample into subsets according to the presence or absence of zeros. One typical example of rounded zero is the zero in a component of a particular mineral which indicates that no quantifiable proportion of the mineral has been recorded by the measurement process. Because this 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. But this replacement is not without problems, as stated e.g. by Tauber (1999) and by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000) for the additive approach. Thus, 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 subpopulations could be misleading.

REPLACEMENT STRATEGIES FOR NONCOMPOSITIONAL DATA

Parametric and Nonparametric Techniques

Let $\mathbf{Y}$ be a data set with missing values in real space $\mathbb{R}^D$. If the goal is to perform a cluster analysis based on a hierarchical clustering method using the Euclidean distance d_e, first it is necessary to complete the matrix of distances between observations. Several strategies have been suggested in the literature for

that purpose, which can be classified into parametric and nonparametric techniques. Among the first ones we find the EM algorithm and its extensions, the multiple imputation techniques and the Markov Chain Monte Carlo method (Allison, 2001; Little and Rubin, 1987; Shafer, 1997). They all provide a set of flexible and reliable tools for inference in large classes of missing-data problems. All these tools rely on fully parametric models for multivariate data.

The group of nonparametric techniques consists, essentially in a family, of strategies known as "imputation." Imputation is equivalent to forcing the incomplete data set into a rectangular complete-data format by inserting a quantity for each missing value. Then, from the completed data set, the multivariate analysis can be performed. In those cases where the matrix of distances is calculated from the completed data set, such "imputation" procedures are not recommended because, whenever very similar or identical estimates are used on different observations, the similarity between these observations are grossly exaggerated (Krzanowski, 1988). In this case, an alternative strategy to "imputation" procedures is suggested by Krzanowski (1988), which can be synthesized as follows:

(i) omit any component that has a missing value when computing the distance between two observations and work only with those components that have all values present for both the observations concerned;
(ii) if the previous step means working with S components instead of D, inflate the resulting distance by a factor D/S.

Imputation Methods: Properties

As exposed in Little and Rubin (1987), the principal nonparametric imputation methods in survey practice include several strategies: mean imputation, hot deck imputation, cold deck imputation, and composite methods. Following Sandford, Pierson, and Crovelli (1993), when the missing values are actually censored data, that is, when the values for some components are reported as "less than" a given threshold value, a simple imputation can be considered. For a "small" proportion of "less than" values (not more than 10%) a simple-substitution method using a δ value equal to 0.55 of the threshold value is suggested. But what is important for our purposes is that all these imputation methods have the following properties in common:

P1. The canonical projection $\Pi(\mathbf{y})$ on the nonmissing components of observation $\mathbf{y} \in \mathbb{R}^D$ is identical to the same projection $\Pi(\mathbf{z})$ of the replaced observation $\mathbf{z} \in \mathbb{R}^D$. Thus, the covariance structure of the components without missing values is preserved.
P2. Consider two observations $\mathbf{y}, \mathbf{y}^* \in \mathbb{R}^D$ having "common" missing values, and $\mathbf{z}, \mathbf{z}^*$ their replaced observations. It holds that $y_j - y_j^* = z_j - z_j^*$, where y_j,

y_j^*, are nonmissing values, and z_j, z_j^* the corresponding replacements. Furthermore, if the imputation method assigns the same replacement value to every missing component y_j of the two observations, then $d_e(\mathbf{z}, \mathbf{z}^*)$ does not depend on the imputed values and it is identical to the Euclidean distance between the projections $d_e(\Pi(\mathbf{y}), \Pi(\mathbf{y}^*))$.

P3. Consider that $\mathbf{y}$ or $\mathbf{y}^*$ have censored values and these censored values are not in the same component. It holds that $\lim_{\delta \to \pm\infty} d_e(\mathbf{z}, \mathbf{z}^*) = +\infty$, where $\mathbf{z}$, $\mathbf{z}^*$ are their replaced observations and δ is the imputed value. Note that in this property we are considering that the sample space of the components is the real space, and then we can consider that $\delta \to \pm\infty$.

REPLACEMENT STRATEGIES FOR COMPOSITIONAL DATA

General Remarks

Certainly, any replacement strategy—parametric and nonparametric—has its advantages and disadvantages. However, in this paper we do not discuss a best strategy. Our goal is to provide a suitable imputation procedure for compositional data. This goal is motivated by the observation that the main difference between parametric and nonparametric strategies is that the first one decides the imputed value based on parametric models. Nevertheless, both strategies must be based on a suitable imputation procedure. Therefore, as a first step, we focus our interest in nonparametric imputation.

Many authors (Allison, 2001; Little and Rubin, 1987; Shafer, 1997) do not recommend the imputation procedures because they can distort the covariance structure, biasing estimated variances and covariances towards zero. Although this weakness of imputation methods is certainly a critical problem, the strategy suggested by Krzanowski (1988) has an even more important weakness: it is not suitable at all for compositional data. To see that this is so, consider the following example. Take three compositional observations $\mathbf{x} = [0, 0.8, 0.2]$, $\mathbf{x}^* = [0.95, 0.04, 0.01]$, and $\mathbf{x}' = [0.06, 0.76, 0.18]$. The strategy of Krzanowski implies comparing the subcompositions formed by the second and third components: $\mathbf{x_S} = [0.8, 0.2]$, $\mathbf{x_S}^* = [0.8, 0.2]$, and $\mathbf{x_S}' = [0.81, 0.19]$. Assuming that the zero in sample $\mathbf{x}$ is actually a rounded zero we expect $\mathbf{x}$ and $\mathbf{x}'$ to be more similar than $\mathbf{x}$ and $\mathbf{x}^*$. Nevertheless, we obtain that $d_a(\mathbf{x_S}, \mathbf{x_S}^*) = 0$ and $d_a(\mathbf{x_S}, \mathbf{x_S}') = 0.07$ (Fig. 1). Therefore, in the present paper, we focus our attention on imputation strategies as a nonparametric strategy without the correction suggested by Krzanowski (1988).

By analogy, a suitable imputation method for compositional data must have properties as reasonable as the properties of the methods for noncompositional data—see properties P1, P2, and P3 under Imputation Methods: Properties section. For compositional data we consider the sample space to be $\mathcal{S}^D$ and we know

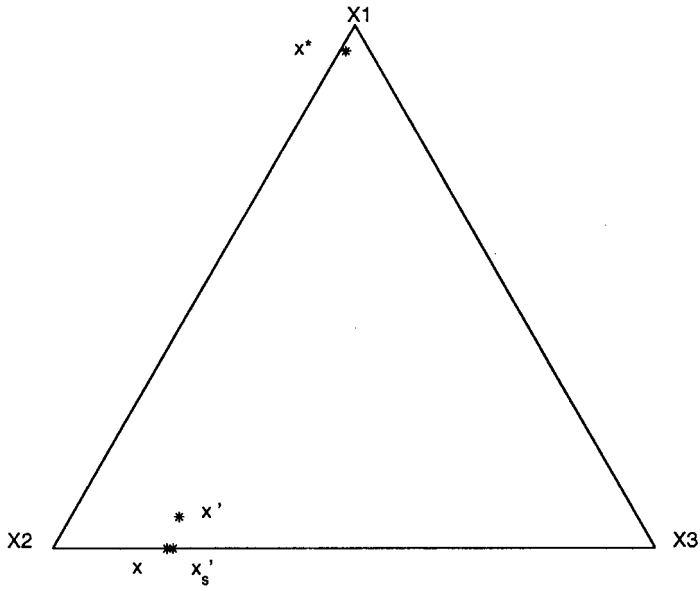

Figure 1. Representation in $\mathcal{S}^3$ of compositions $\mathbf{x} = [0, 0.8, 0.2]$, $\mathbf{x}^* = [0.95, 0.04, 0.01]$, $\mathbf{x}' = [0.06, 0.76, 0.18]$, and their subcompositions $\mathbf{x_S} = [0.8, 0.2]$, $\mathbf{x_S}^* = [0.8, 0.2]$, and $\mathbf{x_S}' = [0.81, 0.19]$. Note that in this representation the three points $\mathbf{x}$, $\mathbf{x_S}$, and $\mathbf{x_S}^*$ overlap.

that, as mentioned above, a measure of difference compatible with the vector space structure of $\mathcal{S}^D$ is the Aitchison distance. Thus, in property P1 the concept of canonical projection on the nonmissing components must be replaced by the concept of subcomposition on the nonmissing parts. Analogously, because perturbation is the internal operation in $\mathcal{S}^D$, the difference invariance in property P2 must be replaced by the perturbation invariance. Moreover, in property P2 it seems logical to expect that a replacement rule of zeros has the property that the Aitchison distance between replaced compositions does not depend on the imputed value when the replaced compositions come from compositions with common zeros. In relation to the suitable sample space (1), in property P3 we must replace, $\delta \to +\infty$ by $\delta \to c^-$ (convergence from the left) and $\delta \to -\infty$ by $\delta \to 0^+$ (convergence from the right), where c is the constant value used for the constraint in (1). Note that any replacement rule of rounded zeros in compositional data is forced to modify the non-zero values because the sum-constrain must be verified.

With the above features in mind, we are motivated to define a suitable replacement method for rounded zeros in compositional data. Nevertheless, first we will revise, from a theoretical point of view and for the purpose of comparing properties, existing replacement methods.

Additive Replacement Strategy

In Aitchison (1986) the following replacement strategy for rounded zeros is suggested. Consider that a composition $\mathbf{x} \in \mathcal{S}^D$ contains Z rounded zeros, then $\mathbf{x}$ can be replaced by a new composition $\mathbf{r} \in \mathcal{S}^D$ without zeros according to the following replacement rule:

$$r_j = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & \text{if } x_j = 0\,, \\ x_j - \frac{\delta(Z+1)Z}{D^2}, & \text{if } x_j > 0\,, \end{cases} \tag{3}$$

where δ is a small value, less than a given threshold. Note that the constant sum-constraint (1) of compositional data forces to modify both the zero and the nonzero values. Note also that we can generalize rule (3) using a different threshold δ_j for every part x_j. The problem is that this replacement rule is additive for nonzero values and is thus not coherent with basic operations of the vector space $\mathcal{S}^D$. Replacement rule (3) has the following properties:

i. The part r_j in (3) depends not only on the threshold δ but also on the number of parts D and the number Z of zeros. Two compositions $\mathbf{x}$ and $\mathbf{x}^*$ containing a different amount of zeros but some of them in the same parts, can be replaced by $\mathbf{r}$ and $\mathbf{r}^*$ having different values in these parts.

ii. If $\mathbf{x}$ and $\mathbf{x}^*$ have "common" zero values, i.e. the same amount of zeros and in the same parts, and $\mathbf{r}$, $\mathbf{r}^*$ are their replaced compositions, then it holds that the value of the Aitchison distance (2) is not preserved, $d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*) \neq d_{\mathrm{a}}(\mathbf{x}_s, \mathbf{x}_s^*)$, where $\mathbf{x}_s$ and $\mathbf{x}_s^*$ are subcompositions of $\mathbf{x}$ and $\mathbf{x}^*$ on their nonzero parts. Furthermore, it is easy to show that $d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*)$ depends on δ and that

$$\lim_{\delta \to 0^+} d_{\mathrm{a}}^2(\mathbf{r}, \mathbf{r}^*) = d_{\mathrm{a}}^2(\mathbf{x}_s, \mathbf{x}_s^*) + \frac{Z}{D(D-Z)} \left[\sum_{x_j \neq 0} \ln\left(\frac{x_j}{x_j^*}\right) \right]^2, \tag{4}$$

i.e., the distance only depends on the nonzero parts of the compositions.

iii. If $\mathbf{x}$ or $\mathbf{x}^*$ have some zero values but these are not "common," then

$$\lim_{\delta \to 0^+} d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*) = +\infty, \quad \text{and} \quad \lim_{\delta \to c^-} d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*) = +\infty.$$

iv. If $\mathbf{x}$ has more than one zero value, then

$$\frac{r_j}{r_k} \neq \frac{x_j}{x_k}, \quad \text{for } x_j > 0, x_k > 0.$$

Furthermore, the value of ratios r_j/r_k depends on the thresholds. Therefore, for any data set where some parts do not have zeros, the covariance structure of the subcomposition on these parts is not preserved. Thus, any subcompositional analysis obtained by multivariate methods based on the covariance structure could be seriously distorted.

In a hierarchical cluster analysis spurious classifications may be obtained if very small values of δ are used, as in this case compositions with zero parts tend to group together according to the number and the position of zeros. This is due to the fact that the Aitchison distance between two replaced compositions is extremely sensitive to changes in the threshold δ when using (3), as illustrated empirically in Tauber (1999). This is due to the additive character of the imputation method and not to the logratio approach, as will be shown later. Thus when δ tends to zero the resulting classification tends to the same classification that we would obtain if the zeros were considered as essential zeros (Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 2000).

Simple Replacement Strategy

Actually, the additive replacement rule (3) is not the only method used to replace rounded zeros in the experimental sciences. Many researchers simply replace the rounded zeros in a composition $\mathbf{x}$ by a small quantity to obtain a vector of positive components, $\mathbf{w} \in \mathbb{R}_+^D$. Then, they apply the closure operator, $\mathbf{r} = \mathcal{C}(\mathbf{w})$. This strategy can be expressed by the following replacement rule

$$r_j = \begin{cases} \frac{c}{c+\sum_{k|x_k=0}\hat{\delta}_k}\,\hat{\delta}_j, & \text{if } x_j = 0, \\ \frac{c}{c+\sum_{k|x_k=0}\hat{\delta}_k}\,x_j, & \text{if } x_j > 0, \end{cases} \tag{5}$$

where $\hat{\delta}_j$ is the imputed value on the part x_j and c is the constant of the sum-constraint in (1). Note that with this procedure the resulting imputed value depends not only on the thresholds $\hat{\delta}_j$ but also on the number Z of zeros of $\mathbf{x}$. Thus, when fixing the $\hat{\delta}_j$, two compositions $\mathbf{x}$ and $\mathbf{x}^*$ containing a different amount of zeros but some of them in the same parts, can be replaced by $\mathbf{r}$ and $\mathbf{r}^*$ having different values in these parts, something well known from the additive strategy discussed earlier. Nevertheless, it would be possible to find suitable values for the thresholds such that equal values would result. This possibility, together with the fact that the procedure implies a multiplicative modification of the nonzero values of $\mathbf{x}$, something that can be directly associated to the internal operation of the vector space $\mathcal{S}^D$, i.e. the perturbation operation, suggested an alternative formulation of a multiplicative replacement rule (Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 2000), whose properties are developed in the following section. A similar approach

has been used in Fry, Fry, and McLaren (1996). It is based on changing the additive component in the additive approach to a multiplicative component, but keeping the replacement value for zero components in (3). They justified this alternative by the fact that the additive replacement rule does not preserve ratios, whereas the multiplicative does, although without further discussion of properties.

Multiplicative Replacement Strategy

Let $\mathbf{x} \in \mathcal{S}^D$ and assume it has Z zeros. We propose to replace $\mathbf{x}$ with a composition $\mathbf{r} \in \mathcal{S}^D$ without zeros using the expression

$$r_j = \begin{cases} \delta_j, & \text{if } x_j = 0, \\ (1 - \frac{\sum_{k|x_k=0}\delta_k}{c})x_j, & \text{if } x_j > 0, \end{cases} \tag{6}$$

where δ_j is the imputed value on the part x_j, and c is the constant of the sum-constraint (1). The modification of nonzero values in (6) is multiplicative, as suggested by the simple replacement rule (5). Furthermore, it can be derived from (5) taking $\hat{\delta}_j = \delta_j\ c/(c - \sum_{k|x_k=0}\delta_k)$, and vice-versa, given that if the imputation value $\delta_j = \hat{\delta}_j\ c/(c + \sum_{k|x_k=0}\hat{\delta}_k)$ is used in (6) then the simple replacement rule (5) is obtained. But the expression in (6) enhances the properties of the multiplicative approach in counterposition to the properties of the additive approach, thus making the alternative much more intuitive. It has the following reasonable properties not satisfied by the additive replacement (3):

P1. It is "natural" in the sense that, if the imputed values δ_j in a composition $\mathbf{x}$ are equal to the "true" censored values, then $\mathbf{r}$ recovers the "true" composition. Moreover, the imputed value δ_j does not depend on the amount of parts D neither on the number Z of zeros. This property is not explicitly satisfied by the simple replacement strategy (5), although it could be forced by taking $\hat{\delta}_j = \delta_j\ c/(c - \sum_{k|x_k=0}\delta_k)$ for given δ_j "true" values.

P2. It is coherent with the basic operations in the simplex: if a selecting matrix $\mathbf{S}$ of nonzero parts of composition $\mathbf{x}$ is considered, and $\mathbf{x}_S = \mathcal{C}(\mathbf{Sx})$ is the subcomposition obtained, denoting by $\mathbf{r}_S = \mathcal{C}(\mathbf{Sr})$ the subcomposition derived from the replacement vector, the following properties hold:
   (a) *subcomposition invariance*—$\mathbf{x}_S = \mathbf{r}_S$.
   (b) *perturbation invariance*—for all $\mathbf{p} \in \mathcal{S}^D$, $(\mathbf{p} \oplus \mathbf{r})_S = (\mathbf{p} \oplus \mathbf{x})_S$;
   (c) *power transformation invariance*—for all $\alpha \in \mathbb{R}$, $(\alpha \otimes \mathbf{r})_S = (\alpha \otimes \mathbf{x})_S$;
   These properties imply that this replacement strategy is coherent with the vector space structure defined on $\mathcal{S}^D$. Above relations between the basic operations and replacement rule (6) are illustrated in Figure 2. Consider the compositions $\mathbf{x} = (0, 1/3, 2/3)$ and $\mathbf{x}^* = (0, 0.64, 0.36)$, and their replaced compositions $\mathbf{r} = (0.05, 0.32, 0.63)$ and $\mathbf{r}^* = (0.05, 0.6, 0.35)$ using (6) with $\delta = 0.05$.

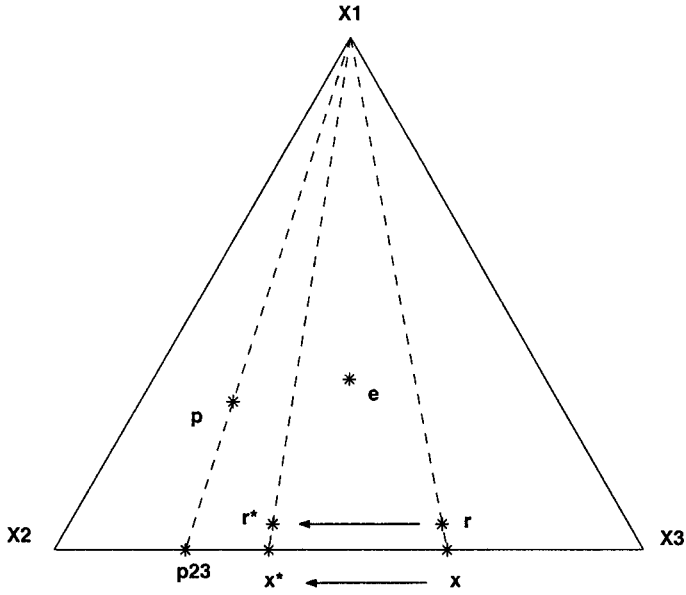

Figure 2. Relations between multiplicative replacement (6) and basic operations of vector space $\mathcal{S}^3$.

Compositions $\mathbf{p}23 = \mathbf{x}^* \oplus \mathbf{x}^{-1}$ and $\mathbf{p} = \mathbf{r}^* \oplus \mathbf{r}^{-1}$ are the perturbations that transform $\mathbf{x}$ to $\mathbf{x}^*$ and $\mathbf{r}$ to $\mathbf{r}^*$, respectively. Certainly, when we calculate the composition $\mathbf{p}23$ we are working only with the nonzero parts of $\mathbf{x}$ and $\mathbf{x}^*$. The dashed lines represent the projections formed by parts X_2 and X_3. We can see that compositions $\mathbf{x}$ and $\mathbf{x}^*$, and their replaced compositions $\mathbf{r}$ and $\mathbf{r}^*$ have the same subcomposition.

P3. Ratios are preserved: $r_j/r_k = x_j/x_k$ for all nonzero values x_j, x_k. We want to emphasize that this preservation of the ratios implies that the covariance structure of components without zeros is preserved. Therefore, any subcompositional analysis based on the nonzero values of the initial data set or based on the replaced data set gives the same results.

P4. When $\mathbf{x}$ and $\mathbf{x}^*$ have "common" zero values, and the replaced compositions $\mathbf{r}$ and $\mathbf{r}^*$ are obtained using identical imputation values $\delta_j = \delta_j^*$, then

(a) $r_j/r_j^* = x_j/x_j^*$ for all nonzero values x_j, x_j^*, and the Aitchison distance $d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*)$ does not depend on the imputed values.

(b) Despite $d_{\mathrm{a}}(\mathbf{r}, \mathbf{r}^*)$ is not equal to $d_{\mathrm{a}}(\mathbf{x}_S, \mathbf{x}_S^*)$, the following equality holds:

$$d_{\mathrm{a}}^2(\mathbf{r}, \mathbf{r}^*) = d_{\mathrm{a}}^2(\mathbf{x}_S, \mathbf{x}_S^*) + \frac{Z}{D(D-Z)} \left[\sum_{x_j \neq 0} \ln \left(\frac{x_j}{x_j^*} \right) \right]^2 ,$$

where Z is the number of common zeros in $\mathbf{x}$ and $\mathbf{x}^*$. It is important to note that above expression is the same as (4) when $\lim_{\delta\to 0^+}$ for the additive replacement.

P5. If $\mathbf{x}$ or $\mathbf{x}^*$ have some rounded zeros but these zeros are not "common," it holds that $\lim_{\delta\to 0^+} d_a(\mathbf{r}, \mathbf{r}^*) = +\infty$, and $\lim_{\delta\to c^-} d_a(\mathbf{r}, \mathbf{r}^*) = +\infty$. Note that, as we have explained above, analogous results are obtained in real space. Thus, this property is not a characteristic weakness of the Aitchison distance. It becomes evident if we represent compositional data on an hyperbolic surface instead of on $\mathcal{S}^D$.

Above properties show that the multiplicative replacement (6) is more suitable than the additive replacement (3) to replace rounded zeros in compositional data. After replacing zeros we can perform any multivariate analysis and, logically, the results of this analysis should be subjected to some form of sensitivity analysis. In hierarchical cluster analysis context, Tauber (1999) and Zhou (1997) have analyzed, from a descriptive point of view, the behavior of the additive replacement (3). The most important conclusion of their work is that, if the value of δ tends to zero, then "spurious" clusters appear in the classification. This conclusion was deduced from the observation that every "spurious" cluster is formed by the data with a "common" zero, i.e. the data with the same amount of zeros and in the same parts. The authors attribute this fact to the logratio transformation, but we can conclude that "spurious" clusters depend on the imputation procedure. Note that this phenomenon happens also when we consider data in real space with the Euclidean distance—see property P3 in page 9. In Aitchison (1986), for a sensitivity analysis the range $\frac{\delta_r}{5} \leq \delta \leq 2\delta_r$, where δ_r is the maximum rounding-off error, is suggested as reasonable. Sandford, Pierson, Crovelli (1993) consider as a suitable imputed value 0.55 of the threshold. Thus, this suggested range seems to be appropriate. In the following case study we use the range given by Aitchison (1986) to compare the behavior of both replacements (3) and (6).

CASE STUDIES

The "Halimba Bauxite Deposit" Data

The first data set, provided by G. Bardossy from the Hungarian Academy of Sciences, and previously used in Mateu-Figueras, Barceló-Vidal, and Pawlowsky-Glahn (1998), 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 $\mathbf{X}$. The sixth part Res_6 consists in a residual part of the composition, i.e., it is equal to $(100 - (Al_2O_3 + \cdots + H_2O))\%$. Some univariate descriptive statistics of the six parts are given in Table 1. Note that the smallest values appear in components SiO_2, TiO_2, and Res_6. Following

Table 1. Univariate Descriptive Statistics for Halimba Data Set

Attribute	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	H_2O	Res_6
Minimum	0.4680	0.0020	0.1400	0.0090	0.1060	0.0020
First quartile (Q_1)	0.5380	0.0125	0.2235	0.0260	0.1175	0.0110
Median (Q_2)	0.5610	0.0280	0.2400	0.0290	0.1215	0.0160
Third quartile (Q_3)	0.5760	0.0490	0.2540	0.0310	0.1255	0.0230
Maximum	0.6220	0.1450	0.3210	0.0390	0.1590	0.0950

Aitchison and Greenacre (2002), the data set can be represented in a biplot diagram (Fig. 3). In this biplot, where the proportion of total variability retained is equal to 96.71%, we can verify the larger variability in the second and sixth components, i.e. $\ln(SiO_2/g)$ and $\ln(Res_6/g)$, where g is the geometric mean of the sample. Circled observations represent compositions considered as outliers of an additive logistic normal distribution (Aitchison, 1986) because their atypicality index is greater than 0.999.

As suggested in Aitchison (1986), 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 $\text{var}[\ln(X_j/X_k)]$ $(j = 1, 2, \ldots, 5; k =$

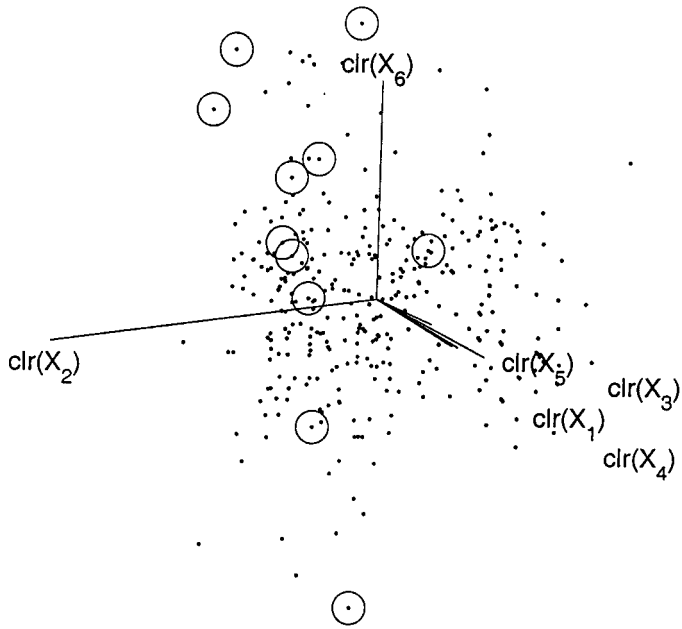

Figure 3. Biplot of the Halimba data set (see text for more details).

Table 2. Variation Array of Halimba Data Set (Estimates): Upper Triangle var[$\ln(X_j/X_k)$]; Lower Triangle E[$\ln(X_k/X_j)$] (See Text for More Details)

	k					
j	1(Al_2O_3)	2(SiO_2)	3(Fe_2O_3)	4(TiO_2)	5(H_2O)	6(Res_6)
1(Al_2O_3)	—	0.8946	0.1288	0.1793	0.0885	0.6105
2(SiO_2)	3.1314	—	0.9095	0.9703	0.8515	0.9321
3(Fe_2O_3)	0.8464	−2.2850	—	0.1915	0.1519	0.6194
4(TiO_2)	2.9981	−0.1333	2.1516	—	0.2214	0.6603
5(H_2O)	1.5140	−1.6174	0.6676	−1.4841	—	0.5566
6(Res_6)	3.5284	0.3970	2.6819	0.5303	2.0144	—

$j+1,\ldots,6$) as an upper triangular array and we use the lower triangle to display in position (j,k) an estimate of the logratio expectation E[$\ln(X_k/X_j)$] ($k=1,2,\ldots,5$; $j=k+1,\ldots,6$). The variation array of the Halimba data set $\mathbf{X}$ is given in Table 2. Observe that the sign of the logratio means corroborate that the components 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. Finally, as introduced in Aitchison (1997), we can compute the compositional geometric mean $\hat{\xi}$ and the total variability, totvar($\mathbf{X}$), of the data set $\mathbf{X}$ defined as

$$\hat{\xi}=\mathcal{C}[g_1,g_2,\ldots,g_D];\quad \text{totvar}(\mathbf{X})=\frac{1}{n}\sum_{i=1}^{n}d_a^2(\mathbf{x}_i,\hat{\xi}), \tag{7}$$

where

$$g_j=\left(\prod_{i=1}^{n}x_{ij}\right)^{1/n}$$

symbolizes the geometric mean of part X_j in data set $\mathbf{X}$. The relation of these measures of central tendency and dispersion with the Aitchison distance d_a have been extensively analyzed in Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998b). In the case of the Halimba data set $\mathbf{X}$ we obtain $\hat{\xi}=[0.5644, 0.0246, 0.2421, 0.0282, 0.1242, 0.0166]$ and totvar($\mathbf{X}$) $=0.9718$.

To illustrate our approach, every observed value of $\mathbf{X}$ smaller than 0.01 is transformed to a zero value. We call $\mathbf{X}^*$ the compositional data set resulting from this procedure. As a consequence, 105 compositions of $\mathbf{X}^*$ have at least one zero. Moreover, out of the 332×6 values in the data matrix, 128 are zero. Note that this amount of zero values is less than 10% of the total amount (332×6). Therefore, it seems reasonable to consider a simple-substitution (Sandford, Pierson, and Crovelli, 1993). These zeros are mainly concentrated in the components

SiO_2 and Res_6. Only one zero appears in the fourth component TiO_2. As can be deduced from Table 1, the components Al_2O_3, Fe_2O_3, and H_2O will have no zeros in $\mathbf{X}^*$.

Sensitivity Analysis

We assume the zeros of $\mathbf{X}^*$ to be nonessential 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 additive replacement approach proposed by Aitchison (3) and the multiplicative approach (6) proposed in this paper. Combining 10 different values $\delta_k = 0.001 * k$, $(k = 1, 2, \ldots, 10)$, with both replacement rules, 20 data sets without zeros are obtained: $\mathbf{R}_{a,k}$, $(k = 1, 2, \ldots, 10)$, using the additive method (3) and threshold δ_k; and $\mathbf{R}_{m,k}$, $(k = 1, 2, \ldots, 10)$, using the multiplicative method (6). Our aim is to compare the sensitivity of both replacement rules in relation to the value δ_k.

Because we know in this case the original data set $\mathbf{X}$, the sensitivity with respect to observed values of $\mathbf{X}$ can be analyzed. We perform this analysis using two different measures. We calculate the Aitchsion distance $d_a(\mathbf{x}_i, \mathbf{r}_i)$ $(i = 1, 2, \ldots, 332)$ between the original composition $\mathbf{x}_i \in \mathbf{X}$ and the replaced composition $\mathbf{r}_i$ obtained from $\mathbf{x}_i^* \in \mathbf{X}^*$. Then, as a first measure of distortion, we consider the mean of these distances squared:

$$\text{msd} = \frac{\sum_{i=1}^{332} d_a^2(\mathbf{x}_i, \mathbf{r}_i)}{332}. \tag{8}$$

By analogy to usual least squares methods, it seems reasonable to assume that the smaller msd for a given replacement value, the better the strategy applied. Note that for compositions without zeros it holds $d_a^2(\mathbf{x}_i, \mathbf{r}_i) = 0$. Thus, msd (8) is a measure of distortion related to compositions with zero values. Actually, we could modify the denominator in (8) by subtracting the amount of compositions without zeros. Nevertheless, it seems to be more sensible that a measure of distortion takes into account the total amount of compositions.

As a second measure of distortion we consider the stress (*st*andardized *re*sidual *s*um of squares) defined by

$$\text{stress} = \frac{\sum_{i<j} (d_a(\mathbf{x}_i, \mathbf{x}_j) - d_a(\mathbf{r}_i, \mathbf{r}_j))^2}{\sum_{i<j} d_a^2(\mathbf{x}_i, \mathbf{x}_j)}. \tag{9}$$

This measure, used for analogous purposes in the same manner in an Martín-Fernández, Olea-Meneses, and Pawlowsky-Glahn (2001), is one of the basic elements of multidimensional scaling theory (Cox and Cox, 1994). Note that if

$\mathbf{x}_i,\ \mathbf{x}_j \in \mathbf{X}$ are two compositions without zeros, it holds that $d_a(\mathbf{x}_i, \mathbf{x}_j) - d_a(\mathbf{r}_i, \mathbf{r}_j) = 0$. Therefore, these distances do not affect the stress. In a different manner than in msd (8), in stress (9) 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.

Figure 4 shows the behavior of the two measures, msd and stress, when either the additive replacement (3) or the multiplicative replacement (6) are applied. Note that the multiplicative replacement (dashed line '$- - -$') shows a better behavior in both measures of distortion. We want to emphasize that the best results are obtained for δ close to 0.0065. This fact is coherent with the procedure followed to introduce artificially zero values in $\mathbf{X}$ because the compositional geometric mean of the compositions belonging to $\mathbf{X}$ with values below 0.01, is the composition (0.5793, 0.0058, 0.2572, 0.0322, 0.1181, 0.0075). Observe that the imputed value 0.0065 is close to the second and sixth part of this geometric mean. This result confirms that the multiplicative replacement (6) is a "natural" substitution (property P1).

The measures msd (8) and stress (9) are only useful when the original data set is known. Thus, we analyze also the sensitivity of the measure of central tendency $\hat{\xi}$ and the measure of dispersion totvar($\mathbf{X}$)—see (7). Figure 5(A) shows the variation of $d_a^2(\hat{\xi}, \mathbf{e})$ when either the additive replacement (3) or the multiplicative replacement (6) are applied, where $\mathbf{e} = \mathcal{C}[1, 1, \ldots, 1] \in \mathcal{S}^6$ is the center of $\mathcal{S}^D$. Note that we are analyzing the norm of the compositional geometric mean. Figure 5(B) shows the variation of totvar($\mathbf{R}_{a,k}$) and totvar($\mathbf{R}_{m,k}$). For comparison purposes we have drawn (dotted line '$\cdots$') the value of these measures for the original data set $\mathbf{X}$. We observe that the multiplicative replacement (6) has a better behavior than the additive replacement (3). Note that the best results are also obtained for δ close to 0.0065. Because our aim is to investigate the reason for this different behavior of both measures with respect to the additive replacement (3) and the multiplicative replacement (6), we analyzed the variation array of the data sets $\mathbf{R}_{a,k}$ and $\mathbf{R}_{m,k}$, $(k = 1, 2, \ldots, 10)$. Table 3 shows only the variation array of data sets $\mathbf{R}_{a,1}$ and $\mathbf{R}_{a,10}$. These data sets are obtained when we apply the additive replacement (3) with, respectively, $\delta = 0.001$ and $\delta = 0.01$. For practical limitations, the variation arrays of the other data sets $\mathbf{R}_{a,k}; (1 < k < 10)$ are omitted. Table 4 shows an analogous information when we apply a multiplicative replacement (6). In comparing the variation array of the data set $\mathbf{X}$ (see Table 2) with the variation array resulting of both replacements, we observe that, in most of the cases, when the multiplicative replacement (6) is applied, the range of sensitivity of log-means and log-variance show better results. It is worthwhile to remark that in Table 4, when the parts involved in the log-means and the log-variance are parts without zeros, i.e. $X_1,\ X_3,\ X_5$, we obtain—as expected—exactly the same results (see bold values in Table 4) as in Table 2, given that the subcompositional covariance structure is preserved.

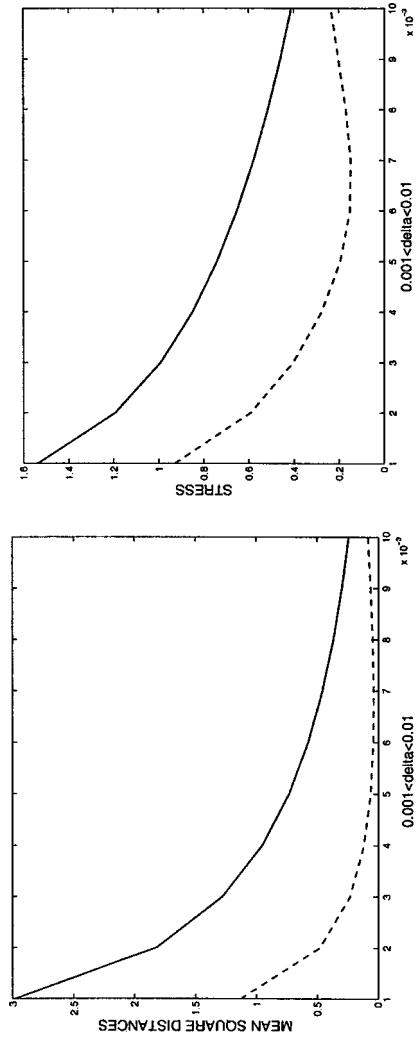

Figure 4. Sensitivity of measures of distortion with the Halimba data set. Dashed line (— —) corresponds to multiplicative replacement; continuous line corresponds to additive replacement: (A) Mean square distances (msd); (B) Stress.

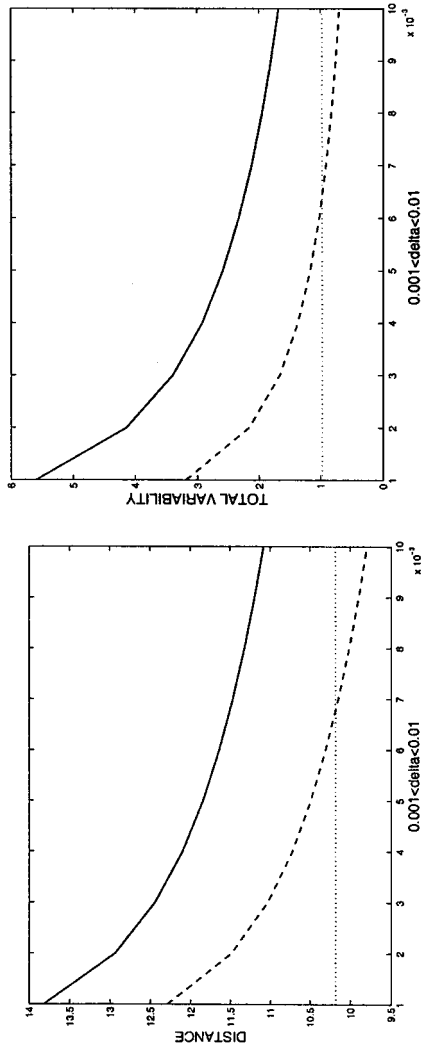

Figure 5. Sensitivity of measures of central tendency and dispersion of Halimba data set (dotted line '. . .'). Dashed line (– – –) corresponds to multiplicative replacement; continuous line to additive replacement; dotted line (· · ·) corresponds to the value of these measures for the original data set Halimba: (A) Distance (square) between compositional geometric mean and $\mathbf{e} = \mathcal{C}[1, 1, \ldots, 1] \in \mathcal{S}^6$, the center of the simplex; (B) Total variability.

Table 3. Variation Array of Data Sets Resulting From Additive Replacement With $\delta = 0.001$ and $\delta = 0.01$ ($\mathbf{R}_{a,1}$–$\mathbf{R}_{a,10}$) (Estimates): Upper Triangle var[ln(X_j/X_k)]; Lower Triangle E[ln(X_k/X_j)] (See Text for More Details)

	k					
j	1(Al_2O_3)	2(SiO_2)	3(Fe_2O_3)	4(TiO_2)	5(H_2O)	6(Res_6)
1(Al_2O_3)	—	2.0387	0.1288	0.3048	0.0886	1.6540
		1.1581	0.1288	0.2098	0.0897	0.8494
2(SiO_2)	3.7858	—	2.0536	2.1201	2.0006	2.3042
	3.3017		1.1729	1.2271	1.1146	1.2427
3(Fe_2O_3)	0.8465	−2.9394	—	0.3101	0.1519	1.6545
	0.8470	−2.4547		0.2198	0.1525	0.8530
4(TiO_2)	3.0093	−0.7765	2.1628	—	0.3320	1.7082
	3.0095	−0.2922	2.1625		0.2461	0.8934
5(H_2O)	1.5142	−2.2717	0.6677	−1.4952	—	1.6195
	1.5157	−1.7860	0.6687	−1.4938		0.8006
6(Res_6)	4.0718	0.2859	3.2253	1.0625	2.5576	—
	3.6807	0.3790	2.8337	0.6712	2.1650	

Sensitivity of Outliers

Another important question to analyze when we replace zeros of a data set is the sensitivity of outliers. For our purposes, a composition $\mathbf{x}_i \in \mathbf{X}$ is considered as

Table 4. Variation Array of Data Sets Resulting From Multiplicative Replacement With $\delta = 0.001$ and $\delta = 0.01$ ($\mathbf{R}_{m,1}$–$\mathbf{R}_{m,10}$) (Estimates): Upper Triangle var[ln(X_j/X_k)]; Lower Triangle E[ln(X_k/X_j)] (See Text for More Details)

	k					
j	1(Al_2O_3)	2(SiO_2)	3(Fe_2O_3)	4(TiO_2)	5(H_2O)	6(Res_6)
1(Al_2O_3)	—	1.5626	**0.1288**	0.2492	**0.0885**	1.2205
		0.7500	**0.1288**	0.1773	**0.0885**	0.5167
2(SiO_2)	3.5280	—	1.5779	1.6414	1.5230	1.7068
	3.0400		0.7651	0.8264	0.7065	0.7757
3(Fe_2O_3)	**0.8464**	−2.6815	—	0.2566	**0.1519**	1.2228
	0.8464	−2.1935		0.1898	**0.1519**	0.5271
4(TiO_2)	3.0047	−0.5233	2.1583	—	0.2816	1.2732
	2.9977	−0.0422	2.1513		0.2198	0.5741
5(H_2O)	**1.5140**	−2.0140	**0.6676**	−1.4907	—	1.1812
	1.5140	−1.5259	**0.6676**	−1.4837		0.4565
6(Res_6)	3.8638	0.3358	3.0174	0.8591	2.3498	—
	3.4663	0.4263	2.6199	0.4686	1.9523	

Table 5. Sensitivity of Atypicality Index (%) of Outliers in Halimba Data Set $\mathbf{X}$ for $\delta = 0.001$, 0.0025, 0.005, 0.0075, 0.01. "T" Symbolizes That the Atypicality Index is Not Greater Than 0.999 (Additive Replacement/Multiplicative Replacement)

		δ				
Obs.	Aty. ind.	0.001	0.0025	0.005	0.0075	0.01
55	99.91	T/T	T/T	T/T	T/T	T/T
223	99.92	T/T	T/T	T/T	T/T	T/99.95
42	99.94	T/T	T/T	T/99.93	T/99.94	T/99.96
62	99.95	T/T	T/99.90	T/99.94	99.90/99.96	99.92/99.97
10	100	T/T	T/99.96	99.91/99.9	99.95/99.99	99.96/100
50	100	99.99/99.99	99.99/100	100/100	100/100	100/100
12	100	T/T	T/99.92	T/99.99	T/100	99.94/100
9	100	100/100	100/100	100/100	100/100	100/100
15	100	T/T	T/99.97	T/100	99.94/100	99.98/100
14	100	T/99.96	99.91/100	99.99/100	100/100	100/100
13	100	100/100	100/100	100/100	100/100	100/100

an outlier if its atypicality index is greater than 0.999 (99.9%) (Aitchison, 1986). Table 5 shows the atypicality index (in percent) of outliers of $\mathbf{X}$ and their variation when we consider the data set $\mathbf{R}_{a,k}$ or $\mathbf{R}_{m,k}$. As suggested by (Aitchison, 1986), for this analysis we have considered only the values $\delta = 0.001$, 0.0025, 0.005, 0.0075, 0.01. Compositions of Table 5 are identified by their row number in data matrix $\mathbf{X}$. The letter "T" symbolizes that the atypicality index of a composition is not greater than 0.999. Note that no big differences are detected between the behavior of both replacement rules with respect to outliers.

Two More Data Sets

Because our aim is to convince the reader of the usefulness of above results, we have studied two more data sets. The second data set has been cited in Aitchison (1986) and was first used for comparison purposes by Bacon-Shone (1992). It consists of 30 samples of foraminiferal composition at 30 different depths. The composition contains four parts, and three compositions have a zero in part three, while two different compositions have a zero in part four. Note that this data set is a set with few compositions and a small number of zeros. The suggested range for sensitivity analysis (Aitchison, 1986) is $0.001 \leq \delta \leq 0.01$. We have analyzed the behavior of the two replacement rules and we have obtained similar results to those obtained with the Halimba bauxite deposit data. As a summary, Figure 6 shows the sensitivity of measures of central tendency and total variability (7) in relation to the value δ used in the replacement rule. Observe that we obtain the same pattern as in the first case (Fig. 5).

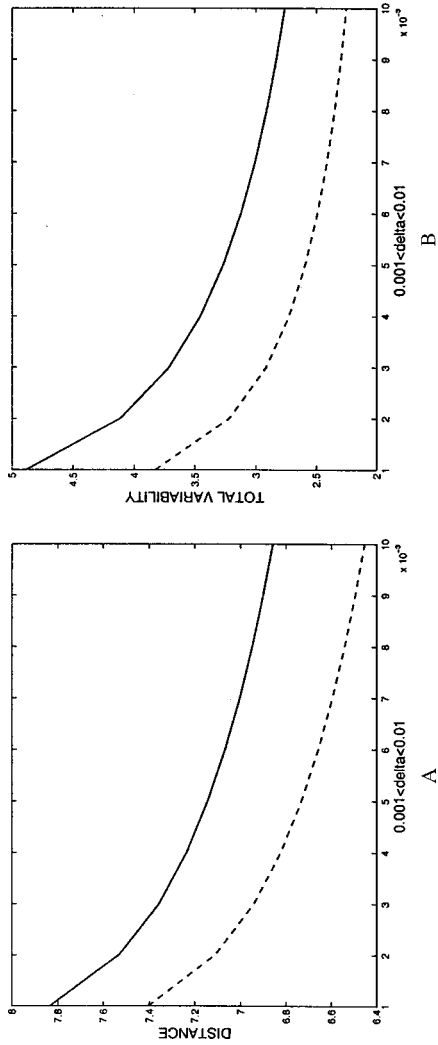

Figure 6. Sensitivity of measures of central tendency and dispersion of Foramin data set. Dashed line (– – –) corresponds to multiplicative replacement; continuous line to additive replacement: (A) Distance (square) between compositional geometric mean and $\mathbf{e} = \mathcal{C}[1, 1, \ldots, 1] \in \mathcal{S}^4$, the center of the simplex; (B) Total variability.

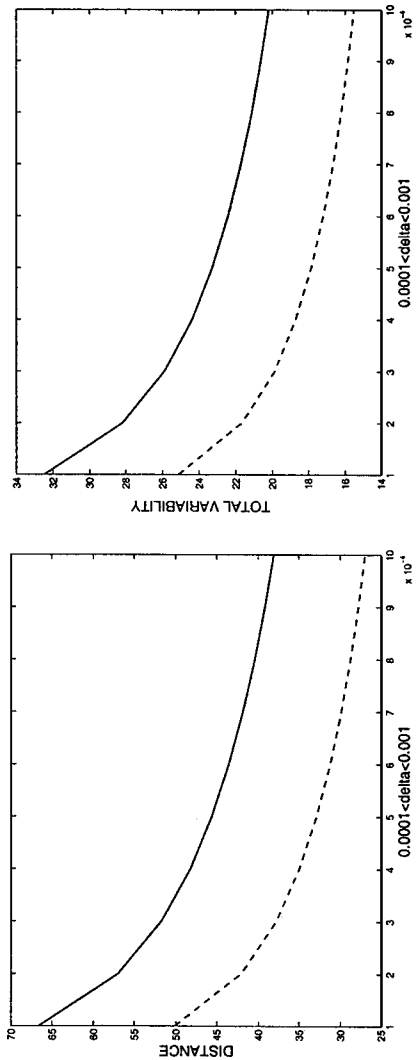

Figure 7. Sensitivity of measures of central tendency and dispersion of the Darss Sill data set. Dashed line (– – –) corresponds to multiplicative replacement; continuous line to additive replacement: (A) Distance (square) between compositional geometric mean and $\mathbf{e} = C[1, 1, \ldots, 1] \in \mathcal{S}^8$, the center of the simplex; (B) Total variability.

The third example is a data set with many compositions and many zero values. It corresponds to granulometric data from the Darss Sill area in the Baltic Sea. This set has been analyzed by many authors for different purposes (Bohling and others, 1996; Davis and others, 1995; Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn, 1997; Tauber, 1999; Zhou, 1997) It has a total number of 1281 grain size samples. Out of 1281 samples, 1163 have at least one zero. Moreover, out of the 1281×8 values in the data matrix, 2852 are zero. Following Aitchison (1986), in this case an appropriate range for sensitivity analysis is $0.0001 \leq \delta \leq 0.001$. Figure 7 shows the sensitivity of measures of central tendency and total variability (7) in the case of the Darss Sill data set. Note that for both measures the same pattern of variation is obtained. We want to remark that these repeated pattern suggests a strong relation between both replacement rules. This relation can be explained noting that in the additive replacement (3) the δ value is subjected to a multiplicative modification by a factor $(Z + 1)(D - Z)/D^2$. Therefore, when we analyze the sensitivity in an appropriate range, we can conclude that the multiplicative replacement (6) has a better behavior than the additive replacement (3).

MULTIPLICATIVE REPLACEMENT OF MISSING VALUES FOR COMPOSITIONAL DATA

According to Aitchison (1997), different types of irregularities can be considered. To illustrate them, consider Table 6. It contains three compositions suffering from different irregularities. Composition 1 has one zero, which can be assumed to be a rounded zero. In that case we apply a multiplicative replacement (6) for a suitable δ. Composition 2 and composition 3 have missing values, but there is a difference between them: the sum of the nonmissing values. In fact, composition 2 suggests that the preliminary determination fails, for whatever reason, to record the quantity of the missing part. Then, what is reported is the composition of the nonmissing subcomposition. A different situation can be observed for composition 3. Because the sum of the nonmissing values in composition 3 is less than one, we can assume that two values (percentages) are lost. Observe that for composition 2 modification of nonmissing values is mandatory, while for composition 3 it is voluntary.

Table 6. Three 5-Part Compositions Suffering From Irregularities

	X_1	X_2	X_3	X_4	X_5
Obs. 1	0.0000	0.1250	0.1237	0.7253	0.0260
Obs. 2	0.1304	0.3151	missing	0.2002	0.3543
Obs. 3	0.1963	missing	missing	0.0819	0.0114

In general, let $\mathbf{x} \in \mathcal{S}^D$ and assume it has Z missing values. If our interest is to impute a value, we propose to replace $\mathbf{x}$ with a composition $\mathbf{r} \in \mathcal{S}^D$ without missing values using the expression

$$r_j = \begin{cases} m_j, & \text{if } x_j \text{ is missing,} \\ \frac{(c - \sum_{k|x_k \text{ missing}} m_k)}{\sum_{k|x_k \text{ nonmissing}} x_k}\, x_j, & \text{if } x_j \text{ is nonmissing ,} \end{cases} \tag{10}$$

where m_j is the imputed value on the missing part x_j and c is the constant of the sum-constraint (1). Obviously, we can use several methods (Allison, 2001; Little and Rubin, 1987; Shafer, 1997) to choose the value m_j. The easiest option is to choose a constant value. The geometric mean of nonmissing values in part X_j is another simple option to be considered. Of course, any imputation method has its advantages and disadvantages. But here we are not discussing the best value for m_j. What we are advocating here is that the multiplicative strategy is suitable to perform any kind of imputation.

Note that, in replacement (10), the modification of nonzero values is multiplicative. For the case of composition 2 it holds that $\sum_{k|x_k \text{ nonmissing}} x_k = c$. Thus, in this case, replacement (10) is identical to replacement (6). Going on to the case illustrated by composition 3, if the imputated values m_j verify the equality $\sum_{k|x_k \text{ missing}} m_k = c - \sum_{k|x_k \text{ nonmissing}} x_k$, then the nonmissing values of composition 3 are not modified. Otherwise, i.e. if $\sum_{k|x_k \text{ missing}} m_k > c - \sum_{k|x_k \text{ nonmissing}} x_k$, these values change by a reducing factor. In any case, as in the zero replacement strategy (6), the multiplicative character of replacement (10) induces reasonable properties: it is "natural," it is coherent with the basic operations in the simplex, and it preserves the ratios between nonmissing parts. These properties are not developed here further, because they are analogous to the properties P1, P2, and P3 of replacement (6).

CONCLUSIONS

In this paper, it is shown that the theoretical drawbacks of the additive zero replacement method proposed in Aitchison (1986) can be overcome using a multiplicative approach on the nonzero 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, a substitution method for missing values on compositional data sets results which has analogous reasonable properties.

ACKNOWLEDGMENTS

The authors thank the reviewers for useful comments which helped improve the paper. This research has been partially financed by the Dirección General de Enseñanza Superior e Investigación Científica (DGESIC) of the Spanish Ministry for Education and Culture through the project BFM2000-0540.

REFERENCES

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, London, 416 p.

Aitchison, J., 1997, The one-hour course in compositional data analysis or compositional data analysis is simple, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology, Vol. 1: International Center for Numerical Methods in Engineering (CIMNE); Barcelona, Spain, p. 3–35.

Aitchison, J., 2002, Simplicial inference, *in* Viana, M. A. G., and Richards, D. S. P., eds., Contemporary mathematics series, Vol. 287: Algebraic methods in statistics and probability, American Mathematical Society, Providence, RI, p. 1–22.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V., 2000, Logratio analysis and compositional distance: Math. Geol., v. 32, no. 3, p. 271–275.

Aitchison, J., and Greenacre, M., 2002, Biplots of compositional data: Appl. Stat., v. 51, no. 4, p. 375–392.

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

Bacon-Shone, J., 1992, Ranking methods for compositional data: Appl. Stat., v. 41, no. 3, p. 533–537.

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V., 2001, Mathematical foundations of compositional data analysis, *in* Ross, G., ed., Proceedings of IAMG'01, The sixth annual conference of the International Association for Mathematical Geology: Cancun, Mexico, 20 p. (CD, electronic publication).

Billheimer, D., Guttorp, P., and Fagan, W., 2001, Statistical interpretation of species composition: J. Am. Stat. Assoc., v. 96, p. 1205–1214.

Bohling, G. C., Davis, J. C., Olea, R. A., and Harff, J., 1996, Singularity and nonnormality in the classification of compositional data: Math. Geol., v. 30, no. 1, p. 5–20.

Cox, T. F., and Cox, M. A., 1994, Multidimensional Scaling: Monographs on statistics and applied probability: Chapman and Hall, London, 213 p.

Davis, J. C., Harff, J., Olea, R., and Bohling, G. C., 1995, Regionalized classification of the Darss Sill sediments, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG'97, The Third Annual Conference of the International Association for Mathematical Geology, Vol. 1: International Center for Numerical Methods in Engineering (CIMNE), Barcelona, p. 145–150.

Fry, J. M., Fry, T. R. L., and McLaren, K. R., 1996, Compositional data analysis and zeros in micro data: Centre of Policy Studies (COPS), General Paper no. G-120, Monash University, Clayton, Australia.

Krzanowski, W. J., 1988, Principles of multivariate analysis: A user's perspective: Clarendon Press, Oxford, 563 p. (reprinted 1996).

Little, R. J. A., and Rubin, D. B., 1987, Statistical analysis with missing data: Wiley, New York, 278 p.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1997, Different classifications of the Darss Sill data set based on mixture models for compositional data, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG'97, The Third Annual Conference of the International Association

for Mathematical Geology, Vol. 1: International Center for Numerical Methods in Engineering (CIMNE), Barcelona, p. 151–158.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998a, Measures of difference for compositional data and hierarchical clustering methods, *in* Buccianti, A., Nardi, G., and Potenza, R., eds., Proceedings of IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology, Vol. 2: De Frede Editore, Napoli, p. 526–531.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998b, A critical approach to nonparametric classification of compositional data, *in* Rizzi, A., Vichi, M., and Bock, H. H., eds., Advances in data science and classification, Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università La Sapienza, Roma: Springer-Verlag, Berlin, p. 49–56.

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, p. 155–160.

Martín-Fernández, J. A., Olea-Meneses, R., and Pawlowsky-Glahn, V., 2001, Criteria to compare estimation methods of regionalized compositions: Math. Geol., v. 33, no. 8, p. 889–909.

Mateu-Figueras, G., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998, Modeling compositional data with multivariate skew-normal distributions, *in* Buccianti, A., Nardi, G., and Potenza, R., eds., Proceedings of IAMG'98, The Fourth Annual Conference of the International Association for Mathematical Geology, Vol. 1: De Frede Editore, Napoli, p. 532–537.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2001, Geometric approach to statistical analysis on the simplex: SERRA, v. 15, no. 5, p. 384–398.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2002, BLU estimators and compositional data: Math. Geol., v. 34, no. 3, p. 259–274.

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

Shafer, J. L., 1997, Analysis of incomplete multivariate data: Chapman and Hall, London, 430 p.

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

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