Statistical Modelling 2015; **15**(**2**): 134–158

Bayesian-multiplicative treatment of count zeros in compositional data sets

Josep-Antoni Martín-Fernández,[1] Karel Hron,[2,3] Matthias Templ,[4,5,3] Peter Filzmoser[4,3] and Javier Palarea-Albaladejo[6]

[1] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

[2] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Czech Republic

[3] Department of Geoinformatics, Faculty of Science, Palacký University, Czech Republic

[4] Department of Statistics and Probability Theory, Vienna University of Technology, Austria

[5] Department of Methodology, Statistics Austria, Austria

[6] Biomathematics & Statistics Scotland, UK

Abstract: Compositional count data are discrete vectors representing the numbers of outcomes falling into any of several mutually exclusive categories. Compositional techniques based on the log-ratio methodology are appropriate in those cases where the total sum of the vector elements is not of interest. Such compositional count data sets can contain zero values which are often the result of insufficiently large samples. That is, they refer to unobserved positive values that may have been observed with a larger number of trials or with a different sampling design. Because the log-ratio transformations require data with positive values, any statistical analysis of count compositions must be preceded by a proper replacement of the zeros. A Bayesian-multiplicative treatment has been proposed for addressing this *count zero problem* in several case studies. This treatment involves the Dirichlet prior distribution as the conjugate distribution of the multinomial distribution and a multiplicative modification of the non-zero values. Different parameterizations of the prior distribution provide different zero replacement results, whose coherence with the vector space structure of the simplex is stated. Their performance is evaluated from both the theoretical and the computational point of view.

Key words: Dirichlet distribution; discrete composition; log-ratio transformations; posterior estimate; zero replacement

1 Introduction

In a broad sense, compositional count data are discrete vectors representing the numbers of outcomes falling into any of several mutually exclusive categories. When the ratios between their components are of interest, rather than the total sum

Address for correspondence: Josep-Antoni Martín-Fernández, Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17071 Girona, Spain.
E-mail: josepantoni.martin@udg.edu

10.1177/1471082X14535524

of the vectors, a *compositional* approach is appropriate. An example arises when the population structure of municipalities (small children, young people, middle generation and elderly inhabitants) is investigated. Although there might be quite different sizes of municipalities, they are not important when the relative amounts of the age groups are of interest. The analysis of compositional data (CoDa) deals with vectors of positive values quantitatively describing contributions of D parts to some whole (Aitchison, 1986; Egozcue, 2009). These vectors are known as *compositions* and their variables, or columns of the data matrix, are known as *parts*. According to Pearson (1897), there is a general agreement that the application of standard data analysis techniques (like correlation analysis) to CoDa may yield misleading results. These difficulties of standard techniques, that rely on the assumption of the Euclidean geometry in real space (Eaton, 1983), are due to the special properties of the sample space of representations of compositions, the simplex $\mathcal{S}^D$, defined by

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D) : x_j > 0; \sum_j x_j = \kappa\},$$

where κ is the total sum of the vector, an irrelevant value in CoDa. Compositions are frequently collected in many applied fields (such as geochemistry, nutritional or behavioural sciences) and are represented as vectors with a constant sum constraint, such as proportions ($\kappa = 1$), percentages ($\kappa = 100$) or ppm ($\kappa = 10^6$). Note that the simplex $\mathcal{S}^D$ is a $(D - 1)$-dimensional subset of R^D because $\sum_j x_j = \kappa$, i.e., $x_D = \kappa - (x_1 + \ldots + x_{D-1})$. In addition, the specific properties of CoDa induce their own geometry, the Aitchison geometry (e.g., Egozcue and Pawlowsky-Glahn, 2006), with the structure of a $(D - 1)$-dimensional Euclidean vector space. Aitchison (1986) proposed a powerful set of log-ratio (logarithm of a ratio) transformations to move CoDa from the simplex to real space via constructing new coordinates, on which the usual statistical methods can then be applied. These coordinates are for example obtained by the *centred log-ratio* (clr) *transformation* (clr):

$$\mathrm{clr}(\mathbf{x}) = \mathbf{y} = (y_1, \ldots, y_D) = \left(\ln\frac{x_1}{g(\mathbf{x})}, \ldots, \ln\frac{x_D}{g(\mathbf{x})}\right) \in \mathsf{R}^D,$$

where $g(\mathbf{x}) = (\prod_{j=1}^{D} x_j)^{1/D}$ is the geometric mean of the parts of the vector $\mathbf{x}$. This transformation projects the simplex $\mathcal{S}^D$ to $\{\mathbf{y} \in R^D : \sum_j y_j = 0\}$ in real space R^D, i.e., the hyperplane of constant sum equal zero. The clr transformation can be used to define the Aitchison distance on the simplex between two compositions $\mathbf{x}$ and $\mathbf{x}^*$ as $d_a(\mathbf{x}, \mathbf{x}^*) = d_e(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{x}^*))$, where d_e stands for the Euclidean distance; the Aitchison distance is popularly used, e.g., in cluster analysis for CoDa (Palarea-Albaladejo *et al.*, 2012). Orthonormal log-ratio coordinates are provided by *isometric log-ratio* (ilr) *transformations* $\mathrm{ilr}(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1}) \in R^{D-1}$ (Egozcue et al., 2003), e.g., the transformation defined in Hron *et al.* (2010) where

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \; i = 1, \ldots, D-1. \tag{1.1}$$

Since the 1990s, numerous publications have extended this methodology in both theoretical and practical aspects. Most of these new ideas and strategies to CoDa have been presented over the last decade at the four CoDaWork meetings (e.g., Egozcue *et al.*, 2011) held so far, and collected in special monographs (e.g., Pawlowsky-Glahn and Buccianti, 2011).

The log-ratio methodology must obviously be preceded by a proper treatment of zero values. In a different context, Richardson (1997, p. 628) stated: 'The history of the zero recognition problem is somewhat confused by the fact that many people do not recognize it as a problem at all'. Fortunately, as comprehensively described in Martín-Fernández *et al.* (2011), it is well known that there are different *zero problems* in CoDa: rounded, essential and count zeros. At the present moment, there is not a general methodology for dealing with essential zeros, also known as absolute or structural zeros. In any case, it is clear that strategies for replacing the essential zero by a small value are not appropriate (Martín-Fernández *et al.*, 2011). A rounded zero is a zero entry in the data matrix that is not a true zero, i.e., it has sense to replace it by an appropriate small value. A rounded zero can represent an observed proportion below a particular maximum possible rounding-off error ($\epsilon = 10^{-d}$) or can represent a very small proportion that cannot be recorded due to the low concentration of a substance or element, below a detection limit ($DL = \epsilon$). The *multiplicative replacement* for dealing with a rounded zero (Martín-Fernández *et al.*, 2003) simply consists in replacing it by a small proportion and, then, modifying the non-zero values in the vector of proportions in a multiplicative way. When the number of these zero values is not large (less than 10% of the values in the data matrix), a replacement which uses an imputation value equal to 65% of the threshold value ϵ can be used (Martín-Fernández *et al.*, 2003). In other cases, the more sophisticated model-based replacements of rounded zeros are recommended (Martín-Fernández *et al.*, 2012).

The treatment of rounded and essential zeros is beyond the scope of our approach. Following Walley (1996), we assume that the compositional count data set contains zero values resulting from insufficiently large samples. That is, they refer to unobserved positive values that may have been observed with a larger number of trials. For example, after 10 trials in an experiment, the number of outcomes falling into three categories has been $(A_1, B_1, C_1) = (0, 4, 6)$. From this result, an estimate of the probabilities is the vector of proportions $\mathbf{x}_1 = (0, 0.4, 0.6)$. The value $A_1 = 0$ indicates a small probability for the first category. When the experiment is identically repeated the second time, doing 36 trials, the count vector obtained was $(A_2, B_2, C_2) = (1, 14, 21)$. Now, the estimate is $\mathbf{x}_2 = (0.028, 0.389, 0.583)$, corroborating a small *positive* value for the first category. Note that in our particular example the ratio between the second and third categories takes the same value: $4/6 = 14/21 \approx 0.667$ (or $0.4/0.6 = 0.389/0.583$), i.e., both samples give the same relative information in these categories. However, log-ratio techniques only can be applied to the second sample because it has no zero values. Similarly, it might happen that in small municipalities, some of the age categories yield zero values, what would most probably not be so if the municipalities would be larger.

In our approach, the new proposed treatment of zero values involves *Bayesian* inference on the zero values and a *multiplicative* modification (Martín-Fernández *et al*., 2003) of the non-zero values in the vector of counts. This approach, known in the literature by the term *Bayesian-multiplicative* (BM) treatment (e.g., Martín-Fernández *et al*., 2011), has been applied in several case studies (e.g., Pierotti *et al*., 2009), comprising *count zeros*. Other studies dealt with other different treatments (e.g., Aebischer *et al*., 1993; Elston *et al*., 1996; Friedman and Alm, 2012; Rodrigues and Lima, 2009) or with other types of zeros (e.g., Butler and Glasbey, 2008; Martín-Fernández *et al*., 2012; Stewart and Field, 2010). However, the existing literature reveals a gap in the theoretic and empirical knowledge with respect to the count zero treatment.

In the following section, basic elements of multinomial and Dirichlet models are reviewed and Bayesian estimation is described. Section 3 presents the new BM treatment, which builds on the proper preservation of principles of CoDa. Section 4 consists of practical examples to illustrate the performance of the proposed techniques. Section 5 provides some remarks about a model-based approach, originally developed for imputation of rounded zeros (Martín-Fernández *et al*., 2012). Finally, conclusions are summarized in Section 6. The programming of the data analyses discussed in this work has been conducted using the open-source R statistical environment (R development core team, 2012). Computer routines implementing the methods can be obtained from the R package 'zCompositions' and also from the website http://www.compositionaldata.com.

2 Posterior Bayesian estimates using the uniform prior

Rodrigues and Lima (2009) and Martín-Fernández *et al*. (2011) presented typical examples of compositional count data. Similarly, the results of the election survey for the November 2012 elections of the parliament of Catalonia, an autonomous community of Spain, can be considered as count data. The answers were recorded for 41 regions, a subdivision of the electoral provinces. The categories were: null vote, abstention, different parties and coalitions and none the above. Because the number of seats in the parliament of Catalonia assigned to each party is related to its percentage of valid votes, the focus was on the counts in the subcomposition formed by categories of parties and coalitions.

Commonly, the multinomial and Dirichlet models are used in the analysis of this type of data. Note that the total sum of the vectors does not provide relevant information; instead, the relative information between categories is of primary interest. In the next section, we develop these popular approaches in order to accommodate them as close as possible to principles of CoDa (Egozcue, 2009).

Following Agresti (2003), suppose that the outcomes in each of n identical and independent trials can fall in any of D mutually exclusive categories. Let $\mathbf{c} = (c_1, \ldots, c_D)$ denote the vector of counts, where c_j stands for the number of trials having outcome in category j. For a fixed n (number of trials), this vector belongs to a $(D-1)$-dimensional space because $\sum_j c_j = n$, i.e., $c_D = n - (c_1 + \ldots + c_{D-1})$. The random vector $\mathbf{C}$ has a multinomial distribution with parameters $(n; \pi_1, \ldots, \pi_D)$, whose probability mass function is

$$p(\mathbf{C}=\mathbf{c}) = \left(\frac{n!}{c_1!c_2!\cdots c_D!}\right)\pi_1^{c_1}\pi_2^{c_2}\cdots\pi_D^{c_D},$$

where the maximum likelihood (ML) estimates of $\{\pi_j\}$ are $\hat{\pi}_j = c_j/n$, the sample proportions. Note that the value of the total sum of the vector (n) is relevant in this model. In addition, the negative correlation between two categories

$$corr[C_i, C_j] = -\frac{\pi_i\pi_j}{\sqrt{\pi_i(1-\pi_i)\pi_j(1-\pi_j)}}\,(i \neq j),$$

suggests that only a particular (namely, a negative) association between them is considered.

The Dirichlet distribution is a common model for the random vector $\boldsymbol{\pi}$ of probabilities. In general, a random vector $\mathbf{X}$, where $\sum_j X_j = 1$, follows a $Dir(D; \boldsymbol{\alpha})$, if its density function equals

$$f(\mathbf{x}) = \frac{\prod_{j=1}^{D}\Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{D}\alpha_j\right)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_D^{\alpha_D-1},$$

where Γ is the Gamma function and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$, $\alpha_j > 0$, is known as the *concentration* vector because $E[X_j] = \dfrac{\alpha_j}{\sum_k \alpha_k}$. The correlation

$$corr[X_i, X_j] = -\frac{\alpha_i\alpha_j}{\sqrt{\alpha_i\left(\sum_k \alpha_k - \alpha_i\right)\alpha_j\left(\sum_k \alpha_k - \alpha_j\right)}}\,(i \neq j),$$

also suggests a particular association between the variables. However, Aitchison (1986, p. 60) stated that the most important difficulty is that '*Dir*(D; $\boldsymbol{\alpha}$) has a very strong *implied independence structure*'. In other words, each ratio X_i/X_j is independent of any other ratio X_k/X_m.

The above limitations of multinomial and Dirichlet models suggest that the log-ratio methodology represents a more general approach to the analysis of count data, as long as the relative information rather than the absolute one is of interest. However, the fact that the Dirichlet model is the conjugate distribution of the multinomial model constitutes a useful tool in the preprocessing of the data. In particular, it holds for the *imprecise Dirichlet model* presented in Walley (1996) which assumes a Dirichlet model $Dir(D; s\mathbf{t})$ as a *prior* distribution of the random vector $\boldsymbol{\pi}$. Here

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D) = s \cdot \boldsymbol{t} = s \cdot (t_1, \ldots, t_D)$, where s is called the *strength* and $\boldsymbol{t} = (t_1, \ldots, t_D)$, where $\sum t_k = 1$, refers to the *prior estimates* of $\boldsymbol{\pi}$, because $E[\pi_j] = t_j$. In this framework, using the Bayes formula, the prior density function of $\boldsymbol{\pi}$, $p(\boldsymbol{\pi}) \propto \prod_{j=1}^{D} \pi_j^{s \cdot t_j - 1}$, multiplied by the likelihood function $L(\boldsymbol{\pi} \mid \mathrm{c}) \propto \prod_{j=1}^{D} \boldsymbol{\pi}_j^{c_j}$, is equal to the posterior density function $Dir(D;\, s \cdot \boldsymbol{t} + \mathrm{c})$ of $\boldsymbol{\pi}$: $p(\boldsymbol{\pi} \mid \mathrm{c}) = (L(\boldsymbol{\pi} \mid \mathrm{c}) \cdot p(\boldsymbol{\pi})) \propto \prod_{j=1}^{D} \pi_j^{c_j + s \cdot t_j - 1}$. Consequently, the *posterior Bayesian estimate* of π_j equals to

$$E[\pi_j \mid \mathrm{c}] = \frac{c_j + s \cdot t_j}{n + s}, \tag{2.1}$$

where the strength s regulates the importance of the prior estimate t_j. For the case of $c_j = 0$, one can write the posterior estimate as $t_j \cdot \frac{s}{n+s}$, i.e., the prior estimate t_j multiplied by a reduction factor $\frac{s}{n+s}$. In another context, Palarea-Albaladejo *et al.* (2007) introduced a similar idea to deal with values below a detection limit, where in essence the reduction of the estimate of an expected value is performed.

The most common prior estimate of $\boldsymbol{\pi}$ is $\boldsymbol{t} = (1/D, \ldots, 1/D)$, the *uniform* prior (Bernard, 2005). Table 1 shows the most usual imprecise Dirichlet models when $t_j = 1/D$. For example, the Jeffreys prior was applied in Pierotti *et al.* (2009). According to Walley (1996), 'the value $s = \sqrt{n}$ yields a posterior mean that is equal to the minimax estimator of $\boldsymbol{\pi}$ under quadratic loss'. Note that the Haldane prior is equal to the ML estimates, the sample proportions. In this case, when $c_j = 0$ the posterior estimate is also equal to zero, being useless for dealing with the zero problem in CoDa. The other posterior Bayesian estimates, Perks, Jeffreys, Bayes-Laplace and SQ, respectively, consist of adding $1/D$, $1/2$, 1 and $\sqrt{n}/D$ pseudo-counts to the counts of each category.

Figure 1 illustrates the performance of the Bayesian estimation in relation to the prior selected. In all cases, the prior estimate of $\boldsymbol{\pi}$ is equal to $\boldsymbol{t} = (1/D, \ldots, 1/D)$. The curves show the values of the Bayesian estimate $t_j \cdot \frac{s}{n+s}$ for $c_j = 0$ when the number

Table 1 Most common Dirichlet models and posterior Bayesian estimates when $t_j = 1/D$.

Prior	s	$\alpha_j = s \cdot t_j$	$\frac{c_j + s \cdot t_j}{n+s}$	$c_j = 0 \rightarrow t_j \cdot \frac{s}{n+s}$
Haldane	0	0	c_j/n	0
Perks	1	$\frac{1}{D}$	$\frac{c_j + 1/D}{n+1}$	$\frac{1}{D(n+1)}$
Jeffreys	$\frac{D}{2}$	$\frac{1}{2}$	$\frac{c_j + 1/2}{n + D/2}$	$\frac{1}{2n+D}$
Bayes-Laplace	D	1	$\frac{c_j + 1}{n+D}$	$\frac{1}{n+D}$
Square root (SQ)	$\sqrt{n}$	$\sqrt{n}/D$	$\frac{c_j + \sqrt{n}/D}{n + \sqrt{n}}$	$\frac{1}{D(\sqrt{n}+1)}$

Source: Authors' own.

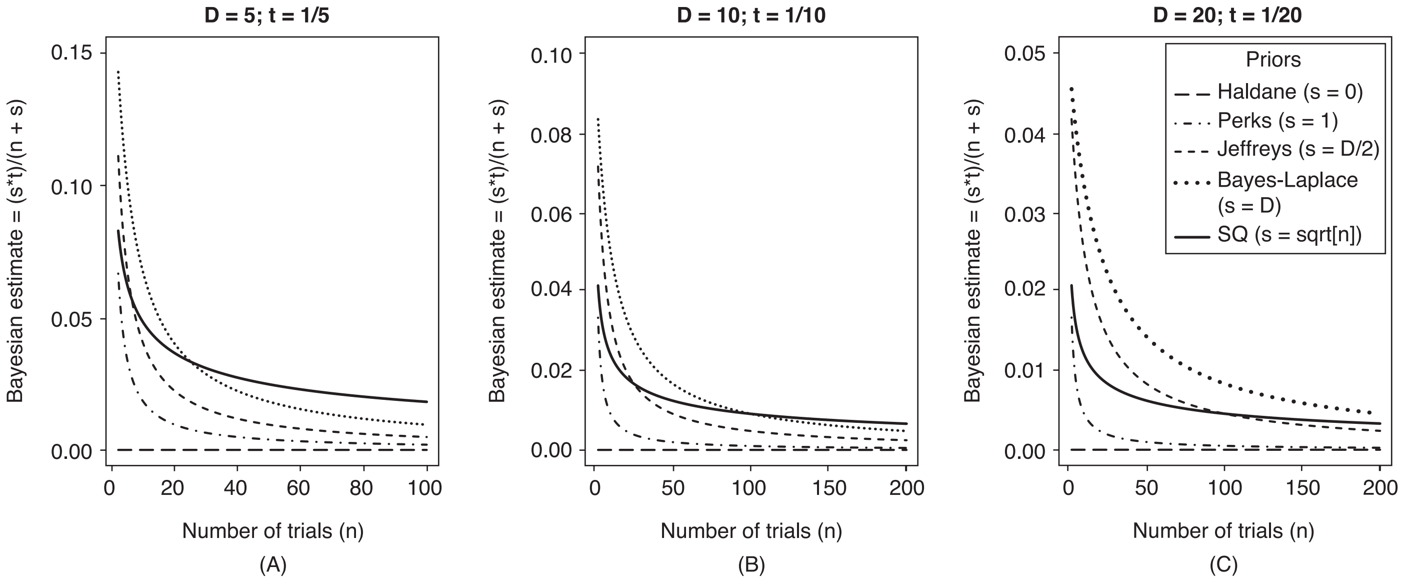

Figure 1 Performance of posterior Bayesian estimation for different priors, when $t_j = 1/D$: (A) for $D = 3$; (B) for $D = 8$; (C) for $D = 20$. Long-dashed line is used for the Haldane prior ($s = 0$), dashed-dotted line for the Perks prior ($s = 1$), dashed line for the Jeffreys prior ($s = D/2$), the Bayes-Laplace prior ($s = D$) is represented by the dotted line and the solid line stands for the SQ prior.

Source: Authors' own.

of trials n varies. The long-dashed horizontal line is the Haldane prior ($s = 0$), the dashed-dotted line is the Perks prior ($s = 1$), the dashed line is the Jeffreys prior ($s = D/2$), Bayes-Laplace ($s = D$) is the dotted line and the solid line is the SQ prior. The estimate provided by the Haldane prior is equal to zero (sample proportion) for any value of n. Regarding the rest of priors, for the three cases considered ($D = 5$, 10 and 20), the Perks prior provides the lowest estimates and the difference is larger for larger values of D. For few numbers of trials n, the largest estimates are provided by the Bayes-Laplace prior. However, when $\sqrt{n} > D$, this role is played by the SQ prior. In the middle we find the estimates from the Jeffreys prior. This one and the Bayes-Laplace prior show a 'parallel' trend in all scenarios. In any case, the larger the strength s, the larger the Bayesian estimates. In other words, for any $D \geq 2$ and n it holds

$$0 < \frac{1}{D(n+1)} \leq \frac{1}{2n+D} \leq \frac{1}{n+D}, \tag{2.2}$$

that states the order between Haldane, Perks, Jeffreys and Bayes-Laplace priors. The posterior estimates provided by the SQ prior are always greater than the Perks estimates. In addition, when $\sqrt{n}$ exceeds the values $D/2$ and D, the SQ estimates exceeds both the Jeffreys and the Bayes-Laplace estimates.

According to Elston *et al.* (1996), a count zero represents censored information where the threshold is equal to $0.5/n$. Note that, when one estimates a zero value in a part, any of the used four priors (Table 1, Haldane excepted) may provide an estimated value greater than the minimum observed value, i.e., above the threshold. For example, when one applies the Bayesian estimation to π_1 from the count vector $(A_1, B_1, C_1) = (0, 4, 6)$, the resulting posterior estimate is respectively equal to 0.03, 0.043, 0.077 and 0.08. In all four cases, the posterior estimate is greater than the ML estimate $\hat{\pi}_1 = 0.028$ obtained from the count vector $(A_2, B_2, C_2) = (1, 14, 21)$. This fact could be a serious difficulty when one replaces zeros by appropriate small values in CoDa.

In addition, the Perks, Jeffreys, Bayes-Laplace and SQ priors (Table 1) have another difficulty because they violate the *representation invariance principle* (RIP): the posterior probability should not depend on D, the number of mutually exclusive categories (Walley, 1996). This principle is important in CoDa, because one of the first steps when preprocessing the data set is the amalgamation of some categories which would affect the posterior Bayes estimates. Figure 1 shows how changes of D affect the value of the posterior estimates. In particular, the posterior estimates provided by the Perks prior $\frac{1/D}{n+1}$ are more sensitive when one amalgamates categories, because the estimates increase as D decreases. For example, for fixed $n = 50$, when $D = 20$ decreases to 8 and 3, the posterior estimates increase from 9.8×10^{-4} to, respectively, 2.5×10^{-3} and 6.5×10^{-3}. The difficulty posed by the fact that the imputed value depends on D, the number of parts in the compositions, is not a new issue in the context of the zeros problem for CoDa. Martín-Fernández *et al.* (2003) dealt with a similar situation for the replacement of values below the detection limit, as the existing replacement in the literature at the moment (Aitchison, 1986) replaced zeros by a value depending on D.

3 Bayesian-multiplicative method for CoDa

Compositional techniques based on the log-ratio methodology (e.g., Pawlowsky-Glahn and Buccianti, 2011) are appropriate in those cases where the total sum of the vector is not of interest. Data coming from multinomial trials are called *compositional count data*. As mentioned above, typical examples are the population structure in municipalities and results of an election survey. Because the log-ratio transformations require data with positive values, any statistical analysis of count compositions must be preceded by a proper replacement of the zeros. The above difficulties of the standard Bayesian methods, and the special nature of compositional count data, motivate the introduction of their particular accommodation for CoDa.

Let $\mathbf{c}_i = (c_{i1}, \ldots, c_{iD})$ be a compositional vector of counts with some zero values and $n_i = \sum_j c_{ij}$. The composition $\mathbf{x}_i = \mathbf{c}_i/n_i$ is replaced by the vector $r_i = (r_{i1}, \ldots, r_{iD})$ using the BM replacement

$$r_{ij} = \begin{cases} t_{ij} \cdot \dfrac{s_i}{n_i + s_i}, & \text{if } x_{ij} = 0, \\ x_{ij} \cdot \left(1 - \sum_{k|x_{ik}=0} t_{ik} \cdot \dfrac{s_i}{n_i + s_i}\right), & \text{if } x_{ij} > 0, \end{cases} \tag{3.1}$$

where t_{ij} is related to the prior and s_i corresponds to the strength of this prior. Both parameters could be different along the samples and the parts because the prior information available could vary depending on the trials or the categories. Using BM replacement (3.1), a zero value is replaced by its posterior Bayesian estimate $E[\pi_j|\mathbf{c}]$. To replace a zero value by an expected value is not a novel procedure. In a parametric approach to the rounded zeros problem, other replacement methods (e.g., Palarea-Albaladejo and Martín-Fernández, 2008; Martín-Fernández *et al.*, 2012; Palarea-Albaladejo *et al.*, 2013) replace the zero values by expected values, assuming a normal distribution on the simplex (Mateu-Figueras and Pawlowsky-Glahn, 2008) and applying classical or robust methods, like in Hron *et al.* (2010), to deal with missing values. In addition, according to Martín-Fernández *et al.* (2003), the non-zero parts in BM (3.1) are modified in a *multiplicative* way. This modification preserves the original ratios between parts, as well as the total sum representation of the vector:

$$\frac{r_{ij}}{r_{ik}} = \frac{x_{ij}}{x_{ik}}; \qquad \sum_{j=1}^{D} r_{ij} = 1.$$

A notable effect of this modification is a minor distortion of the association between the parts (Martín-Fernández *et al.*, 2003).

Let us consider the same example from above where, after $n = 10$ trials in an experiment, the number of outcomes falling into three categories was $(A_1, B_1, C_1) = (0, 4, 6)$. The BM procedure (3.1) with the Jeffreys prior ($t_{ij} = \frac{1}{3}$ and $s_i = \frac{3}{2}$) is applied to replace the zero in the category A. First, the vector of counts is normalized to unit constant sum constraint to get the vector of proportions $\mathbf{x}_1 = (0, 0.4, 0.6)$. Then the zero value in the composition is replaced by the Bayesian estimate. Next, the non-zero values are modified accordingly to produce the vector $\mathbf{r} = (0.0435, 0.3826, 0.5739)$, where the constant sum constraint 1 is automatically preserved. Finally, an *artificial* vector of counts could be produced if the replaced composition $\mathbf{r}$ is multiplied by the total $n = 10$. Note that the ratio B/C between the non-zero values is preserved, i.e., the relative information between these two parts is the same as before the replacement. On the other hand, if an analyst decides to apply the typical posterior Bayesian estimation (2.1) to all the values, zeros and non-zero values, this information changes. For example, using replacement (2.1) in the sample $(A_1, B_1, C_1) = (0, 4, 6)$, the compositional vector $\mathbf{x}_1 = (0, 0.4, 0.6)$ should be replaced by the vector $\mathbf{r} = (0.0435, 0.3913, 0.5652)$, where the total sum is equal to 1. In this case, the relative information between the categories B and C is distorted because the ratio B/C increases to 0.6923. In addition, the multiplicative modification in BM replacement (3.1) suggests a more natural behaviour. For example, imagine that the analyst imputes 0.028 in the zero value, i.e., takes $t_{ij} = 1/3$ and $s_i = 10/11$ in replacement (3.1). In this case, the resulting compositional vector is (0.028, 0.389, 0.583), which is equal to the composition of the second trial (A_2, B_2, C_2). This consistent behaviour of the BM replacement (3.1) suggests that the treatment of zeros will cause a minor distortion in the data set.

The formula of BM replacement (3.1) offers the possibility of using valuable information from the past. Although Bernard (2005) stated that the most common prior is $\boldsymbol{t} = (1/D, \ldots, 1/D)$, an analyst could have past specific information about π_j and use it in t_j. In this way, a strategy could be to take the prior equal to the sample estimation of the expected value of the part. Indeed, given a data set $\mathbf{C}$, let $\mathbf{X}$ denote the corresponding data set formed by the vectors of proportions $\mathbf{x}_i = \mathbf{c}_i/n_i$, $i = 1 \ldots N$. When one wants to replace the count zeros in a vector $\mathbf{x}_i$, the parameters considered are

$$\alpha_{ij} = \sum_{\substack{k=1 \\ k \neq i}}^{N} c_{kj} \quad \text{and} \quad t_{ij} = \frac{\alpha_{ij}}{\sum_{k=1}^{D} \alpha_{ik}} = \hat{m}_{ij}. \tag{3.2}$$

Hereafter, we denote by $\boldsymbol{t}_i = \hat{\mathbf{m}}_i = (\hat{m}_{i1}, \ldots, \hat{m}_{iD})$ this prior. Note that $\hat{\mathbf{m}}_i$ is the ML estimate when one assumes a multinomial model (Agresti, 2003) for the analysis of the data set, and considers that the sample $\mathbf{c}_i$ is not yet observed. In other words, this prior is an estimate obtained with a *leave-one-out* scheme, i.e., when one deals with the vector $\mathbf{x}_i$ and assumes that the other samples are the prior information. Table 2 shows the resulting posterior Bayesian estimates when one modifies the values Perks,

Table 2 Dirichlet models and modified posterior Bayesian estimates when $t_j = \hat{m}_j$.

Prior	Perks	Jeffreys	Bayes-Laplace	SQ	GBM
s	1	$D/2$	D	$\sqrt{n}$	$1/g_i$
$t_j \cdot \frac{s}{n+s}$	$\hat{m}_j \cdot \frac{1}{n+1}$	$\hat{m}_j \cdot \frac{D}{2n+D}$	$\hat{m}_j \cdot \frac{D}{n+D}$	$\hat{m}_j \cdot \frac{1}{\sqrt{n}+1}$	$\hat{m}_j \cdot \frac{1}{g_i \cdot n+1}$

Source: Authors' own.

Jeffreys, Bayes-Laplace and SQ—from Table 1 using the prior $\hat{\mathbf{m}}_i$. Note that Equation (2.2) also holds with priors from Table 2 because they are related to priors from Table 1 by the factor $\hat{m}_j \cdot D$. Because usually $\hat{m}_j < 1/D$, each posterior estimate from Table 2 is smaller than the corresponding estimate in Table 1, and for $\pi_j \approx 1/D$ both treatments will provide similar results. In addition, modified Perks posterior estimates result again in the lowest values, extremely close to zero, suggesting potential outliers when the composition is represented in terms of log-ratio coordinates. Because the prior $\hat{\mathbf{m}}_i$ is very informative, one should consider larger values of the strength parameter s. Note that when $s \to \infty$, the posterior probabilities tend to t_j. From this point of view, when $\sqrt{n} > D$, the SQ prior is the most recommendable among the priors from Table 1. Otherwise, the Bayes-Laplace prior is the best. For this prior, the strength parameter is $s = D = 1/(1/D)$, i.e., the inverse of the average of a uniform prior $\mathbf{t}_i = (1/D, \ldots, 1/D)$. According to Aitchison (1986), the average of a compositional vector is equal to its geometric mean. Consequently, the strength parameter associated with the prior $\hat{\mathbf{m}}_i$ is $s = 1/g_i$, where g_i stands for the geometric mean of $\hat{\mathbf{m}}_i$. The right-most column of Table 2 shows the posterior probabilities provided by this prior, hereafter denoted as Geometric BM (GBM) prior.

Posterior estimates provided by the GBM prior depend on D because the geometric mean involves all parts of a composition. In other words, the GBM prior violates the RIP and is not invariant under amalgamation of categories. GBM posterior estimates have an interesting interpretation in terms of future trials. Consider that after an average of $g_i \cdot n$ trials, no outcome falls in category j, whose prior probability estimate is $\hat{m}_j$. Let V be a random variable taking values $0/(g_i \cdot n + 1)$ and $1/(g_i \cdot n + 1)$ with respective probabilities $1 - \hat{m}_j$ and $\hat{m}_j$. In this case, the expectation of V is equal to $E[V] = \hat{m}_j \cdot \frac{1}{g_i \cdot n + 1}$. Using this approach, the BM replacement (3.1) is now expressed as

$$r_{ij} = \begin{cases} \dfrac{\hat{m}_{ij}}{g_i \cdot n_i + 1}, & \text{if } x_{ij} = 0, \\ x_{ij} \cdot \Bigl(1 - \dfrac{\sum\limits_{k \mid x_{ik}=0} \hat{m}_{ik}}{g_i \cdot n_i + 1}\Bigr), & \text{if } x_{ij} > 0. \end{cases}$$

This expression is very similar to that of the multiplicative replacement for rounded zeros (Martín-Fernández *et al.*, 2003) and the log-normal replacement for values below a detection limit (Palarea-Albaladejo *et al.*, 2013). In all of them, a non-observed proportion is replaced by its estimate, and the observed proportions are modified to accommodate the imputed value preserving the ratios. Consider once again the simple example of a data set **C** only formed by the count vectors $(A_1, B_1, C_1) = (0, 4, 6)$ and $(A_2, B_2, C_2) = (1, 14, 21)$. We apply the GBM replacement to obtain a posterior estimate of the zero value in the first vector. Using the prior estimate $\hat{m}_1 = 0.0278$ from Equation (3.2), the resulting composition is $\mathbf{r} = (0.0098, 0.3961, 0.5941)$. Obviously, in this case the ratio B/C is also preserved and, in contrast to the four used priors in Table 1, the posterior estimate 0.0098 is lower than the observed value 0.028.

The GBM replacement can impute some values greater than the minimum observed proportion in a part. Let l_j be the lowest non-zero observed value in the part X_j. The imputed value $\frac{\hat{m}_{ij}}{n_i + 1}$ may be greater than l_j when $\frac{\hat{m}_{ij}}{l_j} > (n_i + 1)$. For example, if $l_j = 1/100$, $\hat{m}_{ij} = 1/10$ and $g_i = 0.2$, for $n_i < 45$ the GBM replacement will impute a value greater than the minimum. Note that this fact may occur when the number of trials n_i in an experiment is small; however, this problem is less likely as n_i becomes larger. In practice, for these samples, the zeros will be replaced by $0.65 \cdot l_j$ to force an appropriate imputed value below the minimum (Martín-Fernández *et al.*, 2003). To summarize, it seems that the SQ and GBM replacements result in the best behaviour among the Bayesian replacement techniques, which represent the most popular way to deal with count zeros.

These important properties are not satisfied by other strategies applied in the existing literature (e.g., Aebischer *et al.*, 1993; Elston *et al.*, 1996; Friedman and Alm, 2012; Rodrigues and Lima, 2009), where the count zero in the vector of proportions is replaced by a small value, but the constant sum or the ratios are not preserved. Elston *et al.* (1996) stated the simple property that a count zero in the vector of proportions represents censored information and set a threshold equal to $\epsilon = 0.5/n$. Consequently, following Martín-Fernández *et al.* (2003), we can adapt the rounded zero multiplicative replacement for the case of count zeros, hereafter *CZM* replacement. That is, we replace each zero in the proportion vector by the small value $0.65 \cdot 0.5/n$ and, then, modifying the non-zero values in a multiplicative way. If we apply the procedure CZM to the count vector $(A_1, B_1, C_1) = (0, 4, 6)$, the resulting replaced vector of proportions is $\mathbf{r} = (0.0325, 0.3870, 0.5805)$, which preserves the ratio B/C and the constant sum. However, the imputed value 0.0325 is greater than the ML estimate $\hat{\pi}_1 = 0.028$. Although the CZM fulfills the RIP because the imputed value $0.65 \cdot 0.5/n$ does not depend on D, in some cases the modified value turns out to be negative. For example, a simulation of 1000 samples with $D = 100$, $n = 20$ and $\pi_i = 0.01$, $i = 1, \dots D$, shows 18 201 negative values in the imputed data matrix. In addition, this treatment adds spurious correlation between rare parts resulting from adding a fixed value, shared by the parts with count zeros. Note that this distortion of the covariance structure also appears in any BM replacement from Table 1, where the parameter s_{ij} and the prior probabilities t_{ij} are constant for all the parts. This inconvenience is also shared by the multiplicative replacement of rounded zeros, recommended only when the number of zeros in the data matrix is not large (Martín-Fernández *et al.*, 2003).

The above preliminary theoretical conclusions need to be illustrated through numerical examples using both simulated and real data sets. This is the goal of the next section.

4 Practical examples

4.1 Real data set

We used a data set kindly provided by Scotland's Rural College (SRUC) in Scotland, UK. It consists of scan sample behavioural observations of a group of 29 sows during 10 days from 7:30 a.m. to 3:30 p.m. and recorded every 5 minutes (97 times). In order to illustrate how to deal with the count zero problem, we focused on the location behaviour during the third day. The possible locations were: straw bed (BED), half in the straw bed (HALF BED), dunging passage (PASSAGE), half in the dunging passage (HALF PASS.), feeder (FEEDER) and half in the feeder (HALF FEED.). Thus, we worked with a count data matrix of 29 rows (sows) by six columns (categories, locations), representing the number of times each sow was seen at each location. The total sum for the 29 count vectors is always 97, i.e., it is not relevant. The data can be considered as compositional count data in $\mathcal{S}^6$ focusing the interest on the ratios between the counts from different categories. In addition, we have information about a binary grouping variable (treatment) with levels MIX (social mix) and CON (control), which splits the animals in two groups of 14 and 15 sows, respectively. The idea is studying differences in behaviour between animals that underwent social mix and control animals that remained in their home pen.

No relevant differences in the patterns of zeros were found between groups. Table 3 shows the information regarding the zero patterns in the full data set. A '0' represents a zero value in a particular category (location) within the pattern, whereas '1' means no zero value. The first column, on the left, shows the number of animals that have such as particular pattern of zeros. There are 10 different patterns of the presence of zeros in the rows, where the most frequent (six rows) is the pattern with zero values in BED, PASSAGE and FEEDER. Only four count vectors in the data set have no zeros. The last row, at the bottom, displays the percentage of zero values in each part. Overall, 28.74% of the entries in the data matrix are equal to zero, where the parts PASSAGE and FEEDER have no zeros. The zeros are mainly concentrated in the parts HALF BED, HALF PASS. and HALF FEED. These columns have 48.30%, 65.60%, and 51.70% zero entries, respectively.

The crucial feature for our study is the presence of count zeros, as a zero value in a particular category (animal location) is not an absolute zero. That is, it is possible that if one observes the animal in a different moment, one may see the animal in

Table 3 Count zero patterns in the sows data set. The value '1' represents that the count is not a zero value. The first column shows the number of rows exhibiting each zero pattern. Here 'H.' stands for 'HALF'.

Number of rows	BED	H. BED	PASSAGE	H. PASS.	FEEDER	H. FEED.
1	0	0	1	0	1	1
1	0	0	1	1	1	0
6	1	0	1	0	1	0
3	1	0	1	0	1	1
2	1	0	1	1	1	0
1	1	0	1	1	1	1
4	1	1	1	0	1	0
5	1	1	1	0	1	1
2	1	1	1	1	1	0
4	1	1	1	1	1	1
Zeros (%)	6.90	48.30	0	65.60	0	51.70

Source: Authors' own.

such a particular location. Consequently, a replacement method is appropriate. We selected the SQ and GBM methods because they have the best properties among the BM treatments. Note that $D = 6$ and $n = 97$ for all compositions, i.e., the SQ strength is greater than the Bayes-Laplace strength (Table 1). In addition, we also applied the CZM replacement, as it represents one of the most used treatments among non-parametric methods (Martín-Fernández *et al.*, 2011). The binary grouping variable is actually not relevant to apply the CZM procedure. However, the SQ and GBM replacements were applied separately for each group (MIX, CON), because the Bayesian estimates are calculated from the counts in each group. To compare these methods, we firstly focused on the parts HALF PASS. and HALF FEED. which have the highest number of zeros. Figure 2 shows the profile of the replaced parts HALF PASS. and HALF FEED. after the zero replacements. Figure 2A shows the imputed values according to CZM replacement. It can be seen that there are many similar values, which results in higher distortion of the covariance structure (Martín-Fernández *et al.*, 2003). Note that, in this case, the imputed values in both parts are $0.65 \cdot 0.5/97 = 0.003$. The posterior estimates using the SQ prior (Figure 2B) are more different, but are extremely close to zero. The GBM replacement imputes the largest variety of values which are greater than the SQ estimates.

These differences between the profiles of the parts are summarized in Table 4. For the parts HALF BED, HALF PASS. and HALF FED., this table shows the more affected univariate statistics (minimum, first quartile Q1 and median), resulting from the imputation. Note that the profiles of the parts HALF PASS. and HALF FEED. are the same when the CZM replacement is applied. This is not the case for the part HALF BED, because the percentage of zeros in this part is 48.30%, lower than the median. For the GBM replacement, the percentiles show more differences between the parts. The values imputed by the SQ replacement are extremely small; the median of the parts HALF PASS. and HALF FEED. are lower than the minimum

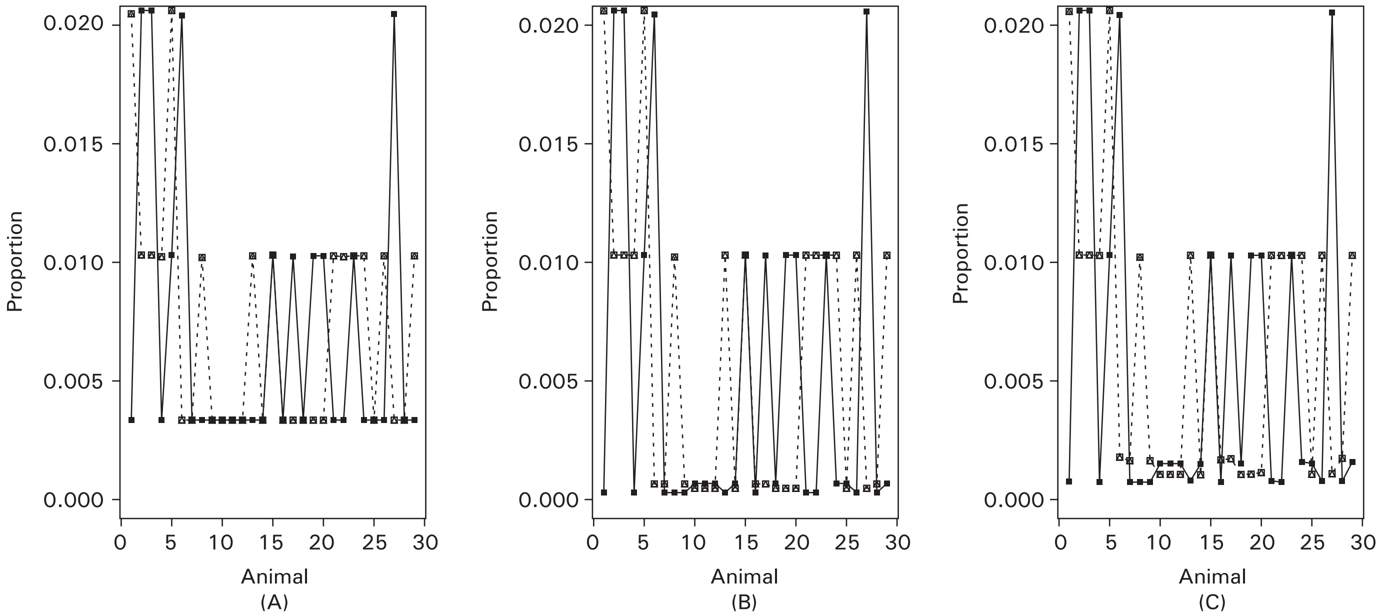

Figure 2 Performance of count zero replacements in the sows data set: (A) CZM; (B) SQ; (C) GBM. Profiles of parts HALF PASS (solid line) and HALF FEED (dashed line).

Source: Authors' own.

Table 4 Some univariate statistics (minimum, first quartile Q1, median—in %) of the parts HALF BED, HALF PASS., and HALF FEED. after replacements CZM, SQ, and GBM. Here 'H.' stands for 'HALF'.

	CZM			SQ			GBM		
Part	Min	Q1	Median	Min	Q1	Median	Min	Q1	Median
H. BED	0.335	0.335	1.024	0.044	0.044	1.030	0.109	0.115	1.028
H. PASS.	0.335	0.335	0.335	0.022	0.029	0.068	0.073	0.077	0.152
H. FEED.	0.335	0.335	0.335	0.047	0.047	0.066	0.104	0.109	0.179

Source: Authors' own.

value imputed by the GBM method. In summary, the CZM replacement method gives a more concentrated distribution of points, because it replaces the zero value using the same imputed values in all parts. This particularity is corroborated by the total log-ratio variability (e.g., Pawlowsky-Glahn and Egozcue, 2002) of the three-part subcomposition for only those 25 sows with at least one zero in these parts. The total variability, i.e., the trace of the covariance matrix of the ilr-transformed data, is equal to 1.09 for the CZM replacement, 5.80 for the SQ procedure and 3.18 for the GBM treatment. The largest log-ratio variability of the SQ estimates is due to the extremely small imputed values. This fact is reproduced when one calculates the total variability in each group: 0.95, 5.67 and 2.95, respectively, for 'MIX', and 0.91, 3.06 and 1.74, respectively, for 'CON'.

After the zero values were replaced in the data set, any statistical analysis could be applied to the log-ratio coordinates (Egozcue *et al.*, 2003). For example, as an exploratory step for the completed data matrix using the various techniques, the biplots of the clr-transformed data (clr-biplot, see Figure 3) were constructed. Figures 3A, B and C show biplots for the data set after CZM, SQ and GBM replacements, respectively. From these clr-biplots, which explain 84%, 74% and 76% of the total variance, respectively, we concluded that sows from the group 'CON' were more linked to location BED, and the sows from the group 'MIX' were mostly associated with location FEEDER. The effect of the CZM imputation is also evident, because the parts with a large number of zeros have a smaller log-ratio variability (Figure 3A). Although the explained variance is lower, Figures 3B and C show that the covariance structures are similar, but the HALF BED, HALF PASS. and HALF FEED. categories have more log-ratio variability for the SQ estimates. To illustrate the difference between the groups, we show barplots (Figure 3D) of the profiles of clr-transformed geometric means for both groups, 'MIX' and 'CON'. We can see that the animals in the group 'MIX' were found more times in the FEEDER and HALF FEEDER, and less times in the other locations than the rest of sows. On the contrary, sows in the 'CON' group were found more times in BED and less in FEEDER. This observed separation between the groups could be further analyzed with some inferential statistical tool such as discriminant analysis (e.g., Filzmoser *et al.*, 2012).

4.2 Artificial data set

To illustrate the behaviour of the BM replacements we designed a Monte Carlo study to simulate a collection of multinomial data sets. Table 5 shows a selection of nine probability vectors $\boldsymbol{\pi}_i$, $i = 1, \ldots, 9$, extracted from Davis (1993), where the authors proposed 28 vectors to cover a wide range of practical situations. For our purpose, we selected those vectors including proportions close to zero. Note that the probability vectors $\boldsymbol{\pi}_i$ range in dimension from $D = 9$ to $D = 50$. Following other comparative studies of zero replacement methods (Palarea-Albaladejo *et al.*, 2013, and references therein), we simulated data sets without zeros first and considered them our reference. Next, defining different scenarios, we generated data sets with zeros. Finally, we applied the methods to replace zero values and evaluated their performance comparing the original reference statistics—centre and variability—with their estimates from the imputed data sets.

The right column in Table 5 shows the initial numbers of trials n_0 that we used to obtain count data sets containing $N = 1000$ multinomial vectors without zeros. For each π_i, we simulated 10 000 data sets and calculated the compositional geometric mean (centre) and the log-ratio covariance matrix. The averages of these 10 000 statistics were considered as the reference values for our comparative study.

We compared a total of 10 replacement methods: the CZM method, four common posterior estimates (Table 1), the corresponding four modified methods (Table 2) and the GBM method. The performance of the zero replacement strategies was compared using different scenarios defined by the combinations of different numbers of trials ($n = 50, 100, 200, 500$) and nine probability vectors $\boldsymbol{\pi}_i$ (Table 5). That is, we employed 36 multinomial (n, $\boldsymbol{\pi}_i$) distributions. To reduce the effect of randomness, for each scenario (n, $\boldsymbol{\pi}_i$) we simulated 20 data sets. Table 6 shows the average number of parts with at least one count zero in each scenario (n, $\boldsymbol{\pi}_i$), $i = 1, \ldots, 9$. Note that the six scenarios without zeros (bottom left corner) were eliminated from the analysis. These scenarios correspond to the greater values of n and lower dimensions of $\boldsymbol{\pi}_i$, where it is less likely to have a count zero. On the other hand, for $\boldsymbol{\pi}_9$ all 50 parts have zeros except when $n_4 = 500$. For this number of trials, we only found a relevant number of parts containing at least one zero (40%) for π_6. The percentage of entries that are zeros in the different scenarios varies from 0, e.g. (n_3, $\boldsymbol{\pi}_1$), to 39.96% in scenario (n_1, $\boldsymbol{\pi}_6$), thus, really covering a wide range of practical situations (Davis, 1993).

For each data set, we imputed the zeros using the 10 different replacement methods and, after imputation, the compositional centre and the log-ratio variability were estimated and the error with respect to the corresponding reference value was evaluated. According to previous similar studies (e.g., Martín-Fernández *et al.*, 2012), for the comparison of centres, we calculated their Aitchison distance, averaged by the number of parts D to avoid the effect of dimensionality. To evaluate the difference between log-ratio covariance matrices we averaged the squared error

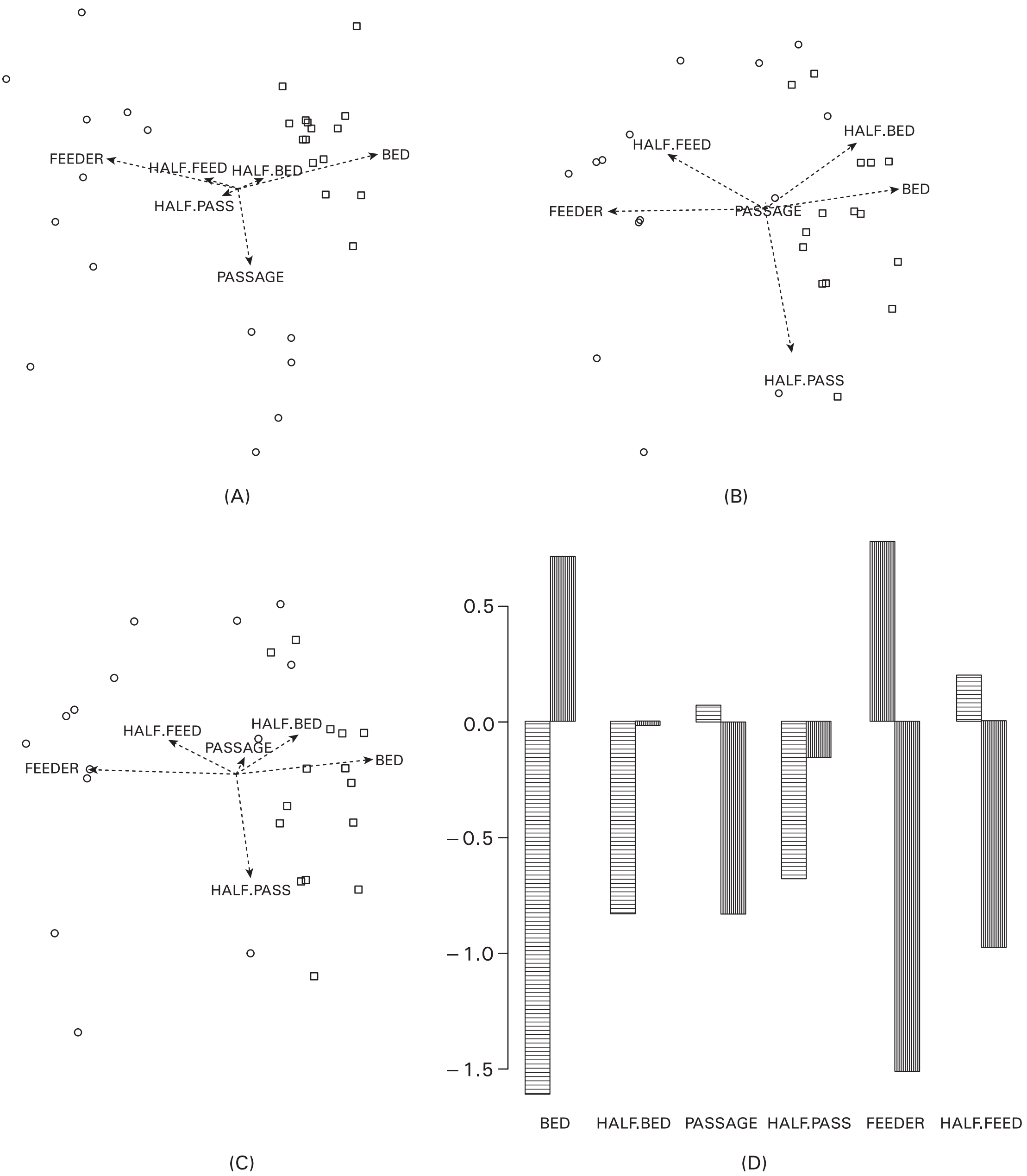

Figure 3 Clr-biplot and barplot of sows data set after the zero replacement: (A) CZM; (B) SQ; (C) GBM. Empty circles (○) and squares (□) correspond to 'MIX' and 'CON' group, respectively; (D) Barplot of the centred geometric means of 'MIX' (horizontal shaded) and 'CON' (vertical shaded) group after GBM replacement.

Source: Authors' own.

Table 5 Probability vectors $\boldsymbol{\pi}_i$ used in the comparative study of Bayesian zero replacement strategies. Column '*i*' shows the number of the scenario, column '*D*' the number of parts. right column shows the initial number of trials.

i	*D*	$\boldsymbol{\pi}_i$	n_0
1	9	0.057 0.077 0.078 0.105 0.105 0.105 0.141 0.141 0.191	200
2	12	0.066 0.071 0.072 0.076 0.078 0.078 0.084 0.086 0.087 0.096 0.097 0.109	200
3	15	0.024 0.033 0.033 0.044 0.044 0.044 0.059 0.059 0.059 0.080 0.080 0.080 0.108 0.108 0.145	400
4	16	0.024 0.031 0.031 0.041 0.041 0.041 0.056 0.056 0.056 0.056 0.075 0.075 0.075 0.102 0.102 0.13	400
5	20	0.016 0.020 0.020 0.028 0.028 0.028 0.037 0.037 0.037 0.037 0.050 0.050 0.050 0.050 0.068 0.068 0.068 0.092 0.092 0.124	600
6	25	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.05 0.05 0.05 0.05 0.50	1100
7	30	0.021 0.023 0.023 0.025 0.025 0.026 0.028 0.028 0.028 0.028 0.030 0.030 0.031 0.031 0.031 0.033 0.034 0.034 0.034 0.034 0.037 0.038 0.038 0.038 0.042 0.042 0.042 0.047 0.047 0.052	2200
8	36	0.019 0.019 0.020 0.020 0.021 0.021 0.021 0.021 0.022 0.022 0.023 0.023 0.024 0.024 0.024 0.024 0.024 0.026 0.026 0.027 0.027 0.028 0.028 0.029 0.029 0.030 0.032 0.032 0.033 0.033 0.037 0.037 0.038 0.043 0.043 0.050	700
9	50	0.02 (50 times)	600

Source: Authors' own.

between the corresponding matrix entries. Finally, we averaged these measures over the 20 data sets and calculated their standard deviations.

In those scenarios, where the number of count zeros was low, all the methods provided similar reasonable results. For example, in scenario (n_2, $\boldsymbol{\pi}_1$), where on average only 0.04% of the 9000 entries in the data matrix were zeros, the few imputed values caused a negligible distortion whatever method was applied. Indeed, across the 10 replacement methods, the mean and standard deviation of the difference between the geometric centres were 2.9×10^{-4} and 2.2×10^{-5}, respectively. Similarly, the resulting values for the difference between the log-ratio covariance matrices were 3.5×10^{-5} and 1.8×10^{-4}. On the other hand, when the number of zeros was higher, the distortion increased. For example, for the most extreme scenario (n_1, $\boldsymbol{\pi}_6$), the mean and standard deviation for the centres were 0.086 and 0.119, and the corresponding values for the covariance structure were 0.01 and 0.012. The large variability among the zero replacements suggests different performance among the methods when the number of zeros becomes large. From the comparison (see Table 7 for details), we concluded that the Perks and modified Perks methods provided the

Table 6 Average number of parts with zeros in each scenario (n, $\boldsymbol{\pi}$).

	$\boldsymbol{\pi}_1$	$\boldsymbol{\pi}_2$	$\boldsymbol{\pi}_3$	$\boldsymbol{\pi}_4$	$\boldsymbol{\pi}_5$	$\boldsymbol{\pi}_6$	$\boldsymbol{\pi}_7$	$\boldsymbol{\pi}_8$	$\boldsymbol{\pi}_9$
n_1	7.15	11.95	14.40	15.55	19.85	24.00	30.00	36.00	50.00
n_2	1.65	2.85	9.10	10.95	15.60	24.00	30.00	36.00	50.00
n_3	0.00	0.00	2.45	3.35	7.60	20.15	19.55	29.85	50.00
n_4	0.00	0.00	0.00	0.00	0.25	10.45	0.10	0.45	1.45

Source: Authors' own.

Table 7 Means and standard deviations ($\cdot 10^{-2}$) of differences between centres and log-ratio covariance matrices for 10 zero replacement methods: CZM, Jeffreys (Jef.), Bayes-Laplace (B-L), Perks (Per.), SQ and GBM. Here 'm.' stands for modified methods (Table 2).

	CZM	Jef.	m. Jef.	B-L	m. B-L	Per.	m. Per.	SQ	m. SQ	GBM
Centre										
Mean	0.78	0.66	0.92	0.80	0.61	1.67	2.36	0.61	1.07	0.53
sd	1.06	0.68	0.93	1.12	0.46	1.84	3.13	0.44	1.19	0.32
Covariance										
Mean	0.20	0.33	0.79	0.18	0.36	6.53	11.10	0.65	1.32	0.26
sd	0.21	0.34	1.24	0.18	0.48	10.70	19.74	0.97	2.24	0.28

Source: Authors' own.

worst estimates for both, the centre and the covariance structure. These methods imputed very small values acting as potential outliers and, consequently, producing large log-ratio differences between centres and covariance matrices. The standard deviations were very large for all the methods, confirming that their performance varies when the number of zeros increases. Concerning the geometric mean (centre), the GBM method showed the best performance, followed by the modified Bayes-Laplace method and the SQ replacement. For the log-ratio covariance matrix, the best method was the Bayes-Laplace one, but CZM, GBM, Jeffreys and modified Bayes-Laplace methods performed nearly as good as the former.

Because the Perks methods resulted in the largest values of the statistics in Table 7, these methods were eliminated from Figure 4. The scenarios were ordered according

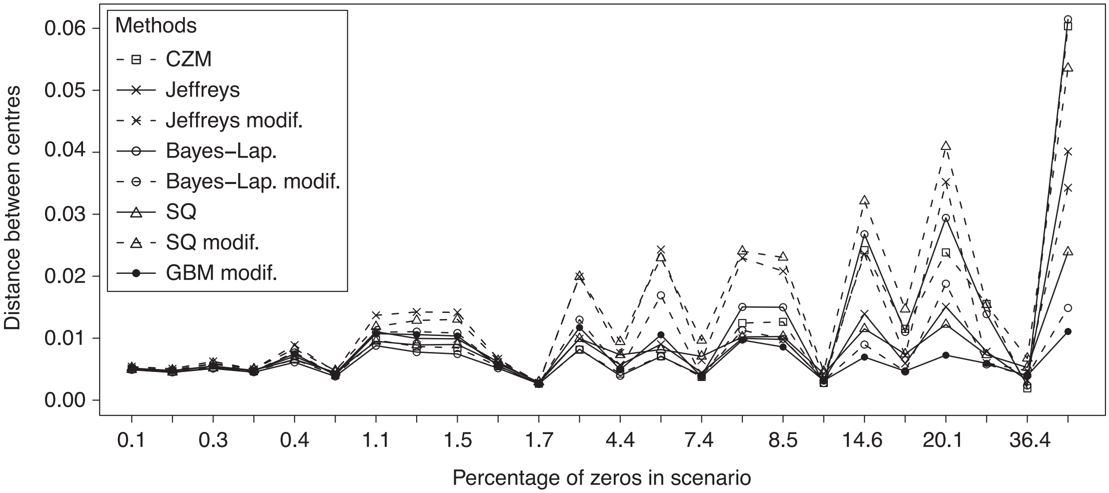

Figure 4 Performance of zero replacement methods. On the horizontal axis, the scenarios with zeros are ordered by increasing percentage of zero entries. The vertical axis accounts for differences between the centres of the imputed data sets and the reference data sets from Table 5.

Source: Authors' own.

to an increasing percentage of zeros. The different lines for each method represent differences between centres of the imputed data sets and the 'true' centres of the data sets simulated according to the parameters in Table 5. When the data sets contain only a few count zeros, the performance of all methods is very similar. However, the larger the percentage of zeros, the more the variability. Observe that only when the percentage of zeros is greater than 14%, the GBM method seems to be the best replacement strategy. In any case, modified Bayes-Laplace and SQ methods show a behaviour close to GBM. The remaining zero replacement methods provide worse performance, and the distances between the centres turn out to be large in the most extreme scenario (n_1, $\boldsymbol{\pi}_6$).

5 Final remarks and further developments

The Bayesian estimation methodology introduced here basically uses the information of the total counts (n) and a prior. Different priors provide different imputed values with their particular properties. Most of the priors were designed to be applied to only one vector rather than to a data set. An apparent advantage of the BM and CZM methods is that the replacement is possible already for one single composition. Indeed, in very particular cases this might be useful. However, from a practical point of view, a statistical analysis of a single observation is irrelevant. Thus, a natural question arises, whether there exist other alternatives that would use information from all observed values in the data set at once.

There might be even further doubts concerning the Bayesian replacement, as it is not fully compatible with the principle of scale invariance of CoDa analysis (Egozcue, 2009). For all the mentioned methods, the number of counts forming a composition plays an important role in the estimation. If the data are of compositional nature, neither the actual number of counts nor a particular representation is important, just the ratios between the parts are relevant to be considered. In practice, it is usually the analyst's decision whether exclusively the relative, rather than the absolute, structure of the parts is of primary interest. Also whether the discrete character of the compositional parts is an artefact of the data collection process or, instead, a feature that should be considered. Unfortunately, none of the Bayesian methods, neither the GBM replacement, do fully account for the scale invariance. A possible way out would be to directly use a model-based replacement procedure designed for imputation of rounded zeros. The lack of a sufficient number of counts in a composition, that results in count zeros, also enables us to consider them as a specific kind of values below the detection limit. Such a limit being, taking the original count compositions, always equal to 1 according to the discrete character of the observations, or 0.5 according to Elston *et al.* (1996) (the latter one was used also for the GBM replacement). In the vector of proportions the detection limit is thus $1/n$ and $0.5/n$, respectively, where n stands for the total counts number. In particular, the iterative imputation algorithm (Martín-Fernández *et al.*, 2012) starts from the initial imputation of rounded zeros by 65% of the detection limit. Then the ilr transformation (1.1) is applied to a compositional data set with its parts sorted in decreasing order according to the amount of zeros. So the first one, x_1, contains the highest number of zeros, x_2 the second highest, and so on. Thus, when performing a censored regression of z_1 on $z_2, \ldots, z_{D-1}$, which guarantees that the imputed values lie below the detection limit, z_1 will be only influenced by the initial values of x_1, but not by the remaining ilr variables. The idea of the procedure is thus to iteratively improve the estimation of the non-detects. After the censored regression of z_1 on $z_2, \ldots, z_{D-1}$ is performed, the results are back-transformed to the simplex, and the cells that were originally non-detects are updated with values below the detection limit, resulting from the estimation. Next we consider the part which originally has the second highest amount of rounded zeros, and the same regression procedure is applied in the ilr space. After each variable containing missing values has been processed, one can start the whole process again until the estimated values stabilize (i.e., when a proper convergence criterion is fulfilled). Finally, a robust alternative of the censored regression instead of the classical one (based on the least squares method) can be used to downgrade the influence of outlying observations on the imputation results. For the exact description of the algorithm we refer to Martín-Fernández *et al.* (2012). A version of this procedure is currently implemented in the function impRZilr in the library robCompositions (Templ *et al.*, 2011) of the statistical software R.

The proposed model-based procedure for imputation of values below the detection limit fulfills all the requirements, mentioned above. However, given the nature of the count zero problem, the scale invariance principle is not fully taken into account also here. The number of counts in the non-zero parts is still partly relevant through the detection limit (e.g., $0.5/n$ when the original compositions are expressed in proportions) and causes the final imputed value to depend on the original data scale (as is also the case for rounded zeros in general). In addition, the imputed values should be always under the detection limit, at each iteration. The convergence guarantees that they have reached stable (ML) values. Nevertheless, in exceptional circumstances, due to numerical instabilities, the implemented model-based algorithm in the function impRZilr might fail to converge. In this case, the algorithm sets imputed values above the detection limit. Also, the requirement that all imputed values will be lower than the minimum observed proportion must not necessarily be fulfilled.

The model-based count zero replacement methods are designed to follow the multivariate structure of the compositional data set. Therefore, another point is the probability distribution that is assumed by a model-based approach, like the multivariate normal distribution in the case of Martín-Fernández *et al.* (2012). This assumption, as stated in Graffelman and Egozcue (2011) and Graffelman (2011), could have difficulties with a bias in the case of discrete compositions that recommends other strategies or distributions. Historically, the Dirichlet distribution has been applied to model proportions in the simplex. Monti *et al.* (2011) introduce a new approach based on the scaled Dirichlet distribution with the purpose to overcome the limitations of the typical Dirichlet in CoDa analysis. Nevertheless, this novelty requires further research in order to be incorporated in a model-based procedure for count zero replacement.

6 Conclusions

When applying multivariate statistical methods to compositional data sets, such as cluster analysis, multidimensional scaling, discriminant analysis or regression analysis, the values in the data matrix must be strictly positive. For count data collected from a multinomial experiment, a BM replacement combines Bayesian estimation with a multiplicative modification of non-zero values. This modification causes just minor distortion in the covariance structure and becomes an appropriate treatment for count zeros. Theoretical analysis and practical experiments showed that the prior GBM provides the most satisfactory results when compared with the other alternatives considered. Among the BM replacements, other methods showing good performance are SQ and modified Bayes-Laplace. In any case, none of the priors fulfills the scale-invariance principle. In addition, the alternative model-based approach has the same difficulty. However, model-based proposals could improve the performance of GBM replacement. Since current proposals of model-based replacement rely on the normal distribution, further developments should explore the performance of other distributions on the simplex, e.g., the extensions of the Dirichlet probability distribution.

Acknowledgements

This research was supported by the Ministerio de Economía y Competividad under the project 'METRICS' Ref. MTM2012-33236, by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424, and by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). The authors also gratefully acknowledge the support by the Operational Program Education for Competitiveness—European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic). Finally, we would like to thank the Animal Behaviour and Welfare group at Scotland's Rural College (SRUC) for kindly providing the animal behaviour data set.

References

Aebischer NJ, Robertson PA and Kenward RE (1993) Compositional analysis of habitat use from animal radio-tracking data. *Ecology*, **74**(5), 1313–25.

Agresti A (2003) *Categorical data analysis*. Wiley Series in Probability and Statistics, p. 710. 2nd edn, Hoboken: John Wiley & Sons.

Aitchison J (1986) *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability (Reprinted 2003 with additional material by The Blackburn Press). London: Chapman and Hall Ltd., p. 416.

Bernard JM (2005) An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, **39**(2–3), 123–50.

Butler A and Glasbey C (2008) A latent Gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society Series C-Applied Statistics*, **57**, 505–20.

Davis CS (1993) The computer generation of the multinomial random variates. *Computational Statistics & Data Analysis*, **16**, 205–17.

Eaton ML (1983) *Multivariate statistics. A vector space approach*. New York: John Wiley & Sons, p. 512.

Egozcue JJ (2009) Reply to 'On the Harker variation diagrams; ...' by J.A. Cortés. *Mathematical Geosciences*, **41**, 829–34.

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

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

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

Elston DA, Illius AW and Gordon IJ (1996) Assessment of preference among a range of options using log ratio analysis. *Ecology*, **77**, 2538–48.

Filzmoser P, Hron K and Templ M (2012) Discriminant analysis for compositional data and robust parameter estimation. *Computational Statistics*, **27**(4), 585–604.

Friedman J and Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, **8**(9), e1002687. doi:10.1371/journal.pcbi.1002687.

Graffelman J (2011) Statistical inference for Hardy-Weinberg equilibrium using logratio coordinates. In Egozcue J.J., Tolosana-Delgado R. and Ortego M.I. (Eds), *Proceedings of the 4th International Workshop on Compositional Data Analysis*, p. 5.

Graffelman J and Egozcue JJ (2011) Hardy-Weinberg equilibrium: A nonparametric compositional approach, Ch. 15. In Pawlowsky-Glahn V. and Buccianti, A. (Eds), *Compositional Data Analysis: Theory and Applications*, pp. 208–17. Chichester, UK: John Wiley & Sons, Ltd.

Hron K, Templ M and Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics & Data Analysis*, **54**(12), 3095–107.

Martín-Fernández JA, Barceló-Vidal C and Pawlowsky-Glahn V (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, **35**(3), 253–78.

Martín-Fernández JA, Palarea-Albaladejo J and Olea RA (2011) Dealing with zeros, Ch. 4. In Pawlowsky-Glahn V. and Buccianti A. (Eds), *Compositional Data Analysis: Theory and Applications*, pp. 47–62. Chichester, UK: John Wiley & Sons, Ltd.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P and Palarea-Albaladejo J (2012) Model-based replacement of rounded zeros in compositional data: Classical and robust approach. *Computational Statistics & Data Analysis*, **56**(3), 2688–704.

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

Monti GS, Mateu-Figueras G and Pawlowsky-Glahn V (2011) Notes on the scaled Dirichlet distribution. In Pawlowsky-Glahn V. and Buccianti A. (Eds), *Compositional Data Analysis: Theory and Applications*, pp. 128–38. Chichester, UK: John Wiley & Sons, Ltd.

Palarea-Albaladejo J, Martín-Fernández JA and Gómez-García J (2007) A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, **39**, 625–45.

Palarea-Albaladejo J and Martín-Fernández JA (2008) A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, **34**(8), 902–17.

Palarea-Albaladejo J, Martín-Fernández JA and Soto JA (2012) Dealing with distances and transformations for fuzzy c-Means clustering of compositional data. *Journal of Classification*, **29**(2), 144–69.

Palarea-Albaladejo J and Martín-Fernández JA (2013) Values below detection limit in compositional chemical data. *Analytica Chimica Acta*, **764**, 32–43.

Pawlowsky-Glahn V and Buccianti A, eds (2011) *Compositional data analysis: Theory and applications*. Chichester: John Wiley & Sons, p. 378.

Pawlowsky-Glahn V and Egozcue JJ (2002) BLU estimators and compositional data. *Mathematical Geology*, **34**(3), 259–74.

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

Pierotti MER, Martín-Fernández JA and Seehausen O (2009) A mapping individual variation in male mating preference space: Multiple choice in a colour polymorphic cichlid fish. *Evolution*, **63**(9), 2372–88.

R development core team (2012) R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.

Richardson D (1997) How to recognize zero. *Journal of Symbolic Computation*, **24**(6), 627–45.

Rodrigues PC and Lima AT (2009) Analysis of an European union election using principal component analysis. *Statistical Papers*, **50**, 895–904.

Stewart C and Field C (2010) Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**(1), 45–69.

Templ M, Hron K and Filzmoser P (2011) robCompositions: An R-package for robust statistical analysis of compositional data, Ch. 25. In Pawlowsky-Glahn V. and Buccianti A. (Eds), *Compositional Data Analysis: Theory and Applications*, pp. 341–55. Chichester, UK: John Wiley & Sons, Ltd.

Walley P (1996) Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society Series B (Methodological)*, **58**(1), 3–57.