Statistical Modelling, 17, 6 (2017): 494 –512.

When relative and absolute information matter. Compositional predictor with a total in generalized linear models

Germà Coenders [1], Josep A. Martín-Fernández [2], Berta Ferrer-Rosell [3]

[1] Department of Economics, University of Girona, Spain

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

[3] Department of Business Administration and Economic Management of Natural Resources, University of Lleida, Spain

Address for correspondence: Germà Coenders, Department of Economics, University of Girona, Faculty building of Economics and Business, Campus Montilivi, E-17003 Girona, Spain.

E-mail: germa.coenders@udg.edu.

Phone: (+34) 972418736.

Fax: (+34) 972418032.

Abstract: The analysis of Compositional Data (CoDa) consists in the study of the relative importance of parts of a whole rather than the size of the whole, because absolute information is either unavailable or not of interest. On the other hand, when absolute and relative information are both relevant, research hypotheses concern both. This article introduces a model including both the logratios used in CoDa and a total variable carrying absolute information, as predictors in an otherwise standard statistical model. It shows how logratios can be tailored to the researchers' hypotheses and alternative ways of computing the total. The interpretational advantages with respect to traditional approaches are presented and the equivalence and invariance properties are proven. A sequence of nested models is presented to test the relevance of relative and absolute information. The approach can be applied to dependent metric, binary, ordinal or count variables. Two illustrations are provided, the first on tourist expenditure and satisfaction and the second on solid waste management and floating population.

Key words: Compositional data; generalized linear model; isometric logratio transformation; T$-$space

1 Introduction

Analysis of Compositional Data (CoDa) is the standard method of statistical analysis when a positive vector variable carries only information about the relative size of its components (Aitchison, 1986). Typical examples are chemical and geological analyses, where only the proportion of each component is of interest, since the absolute amount is only telling about the *size* or *volume* of the analysed container or sample of rocks. Some accessible recent references are van den Boogaart and Tolosana-Delgado (2013), Pawlowsky-Glahn and Buccianti (2011), and Pawlowsky-Glahn et al. (2015a).

Let $\mathbf{x}$ be a vector in the positive orthant of the real space:

$$\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_D)\ \in R_+^D\ ,\ \text{with}\ x_j > 0\ \text{for all}\ j = 1, 2, \ldots, D, \tag{1.1}$$

where D is the number of components. The *closure* of vector $\mathbf{x}$ to a constant k sum is the compositional vector $\mathbf{z}$ which resides in a R_+^{D-1} subspace, known as the *simplex*:

$$\begin{gathered}\mathbf{z} = C(\mathbf{x}) = \left(\frac{kx_1}{\sum_{j=1}^D x_j},\ \frac{kx_2}{\sum_{j=1}^D x_j},\ \ldots,\ \frac{kx_D}{\sum_{j=1}^D x_j}\right) = (z_1,\ z_2,\ \ldots,\ z_D) \\ \text{with}\ z_j > 0\quad for\,all\, j = 1, 2, \ldots, D; \quad \sum_{j=1}^D z_j = k\,. \end{gathered} \tag{1.2}$$

It can even be the case that the absolute size of $\mathbf{x}$ is already constant. The analysed container may always be of the same volume or data may be only available in percentage units. Another common case of constant absolute size is time use research, on a 24 hour basis. Despite the change in the closure constant, and even regardless of whether closure is performed at all, the relative information carried out by the D parts should remain the same, ensuring the so-called *compositional equivalence* property. That is, both vectors $\mathbf{x}$ and $\mathbf{z}$ are elements of the same equivalence class (Barceló-Vidal and Martín-Fernández, 2016).

The most common CoDa approach is to express an original compositional vector of D components into *logratio coordinates*: logratios among components or among their geometric means (Aitchison, 1986; Egozcue et al., 2003). There are three main arguments for logratios. First, the ratios of geometric means of parts and their logarithms constitute a natural way of distilling the information about relative size of components. Second, logratios are unbounded and, once they have been computed, standard statistical analyses assuming that variables lie in the full R^{D-1} real space are appropriate. Third, logratios are compositionally equivalent, as they yield the same result regardless of whether they are computed from $\mathbf{x}$ or $\mathbf{z}$, and, in the latter case, are invariant to the closure constant k. As it is well known, logratio transformations imply that $\mathbf{x}$ may contain no zero values. If the $\mathbf{x}$ vector contains zeros, they have to be replaced prior to computing the logratios. This issue is outside the scope of the article. The interested reader may resort to Martín-Fernández et al. (2003), Martín-Fernández et al. (2011), Martín-Fernández et al. (2015), and Palarea-Albaladejo et al. (2007).

This article considers the situation in which not only the relative size of components is interesting to the research objectives, but also absolute size, provided that size is not constant. Ultimately the researchers' objective and knowledge dictate whether size and the absolute amount of each component in $\mathbf{x}$, if available, matter to the research question beyond the information carried by $\mathbf{z}$. If the answer is that size and absolute information might matter, ignoring them may result in a loss of predictive power at least or in a misspecified model at worst. For instance, in web mining research looking for the occurrence of terms in web pages or other electronic documents, one would first tend to think that term relative frequency matters more than absolute frequency (Russell, 2014). After all, long documents can have more of every term. In any case, there is room for reasonable doubt; for instance, what if certain specific behaviours encountered in long documents correlate with the variables of interest to the researcher? In this example, absolute information exists and is ready to use.

If absolute information matters one may, of course, use standard statistical analyses on the D absolute variables directly, usually after a log transformation. However, the log-absolute value of a component depends both on the overall size and on the relative importance of that component, thus making tests of theoretical hypotheses concerning only absolute size or only relative importance difficult (Ferrer-Rosell et al., 2016a). Recently, the so-called T−spaces have been developed to enable researchers to analyse the relative and absolute size of components together in the same statistical model, while not confounding effects involving the relative importance and effects involving absolute size (Pawlowsky-Glahn et al., 2015b). The approach boils down to adding some form of total to the logratio coordinates and is referred to as *CoDa with a total.* Pawlowsky-Glahn et al. (2015b) show the statistical properties of a total computed from the geometric mean of all D absolute values on the one hand, and of a total computed from the sum of all D absolute values, on the other. Ferrer-Rosell et al. (2016a) introduce the compositions and the total as dependent variables in a linear model. The authors show that, in the role of dependent variable, the researcher enjoys some freedom in tailoring the computed total to the research questions. For example, under some circumstances it is not even necessary to include all D components in the total, which may be computed from the geometric mean of a subset of components. In addition, they show that computing the total in one way or another does not modify the results of the tests involving the logratios.

The aim of this article is to present CoDa with a total in the case in which the composition and the total play the role of explanatory variables. The extension from a purely linear model into a generalized linear model is straightforward. For instance, if the dependent variable is a count, a Poisson regression can be specified, or if the dependent variable is ordinal or binary, an ordered or a binary logit model can be specified. Section 2 presents the concept of logcontrast, which is crucial to construct the logratio coordinates, and the model in which only the composition is the explanatory variable. Section 3 introduces the CoDa model with a total. Its properties and its relation with the classical model using logarithms are described.

Different submodels are explored in Section 4. Sections 5 and 6 illustrate the approach with two real data examples from different fields and with different types of dependent variables. Section 7 concludes and discusses the main contributions.

2 Model with compositional predictor

The CoDa methodology started when Aitchison (1986) introduced *logratio coordinates*. In the simplest case of having only D =2 components, only one logratio coordinate is needed:

$$f(\mathbf{x}) = \ln(x_1/x_2) = \ln\left(x_1^1 x_2^{-1}\right) = \ln(x_1) - \ln(x_2) \; . \tag{2.1}$$

The logratio $f(\mathbf{x})$ takes the values in the full real space, and it is symmetric in the sense that $\ln(x_1/x_2) \quad = -\ln(x_2/x_1)$. For the general case, the most interesting type of logratio coordinate is the *logcontrast*:

$$f(\mathbf{x}) = \psi_1 \ln(x_1) + \psi_2 \ln(x_2) + \ldots + \psi_D \ln(x_D) \quad \text{with} \quad \sum \psi_j = 0 \; , \tag{2.2}$$

where the restriction$\sum \psi_j = 0$ ensures the compositional equivalence property defined in Section 1. $D-1$ linearly independent logcontrasts contain all information about the relative importance of the D components (Pawlowsky-Glahn et al., 2015a).

As a general guideline to find $D-1$ interpretable logcontrasts, Egozcue et al. (2003) propose the so-called *isometric logratio (ilr) coordinates*. Key advantages of ilr coordinates are first, that they can be used as variables in all standard statistical methods, both as dependent and as explanatory variables (Di Marzio et al., 2015); second, that they preserve key properties of the original data (Euclidean distances computed from ilr coordinates are interpretable and equivalent to Aitchison's compositional distances); and third, that they are flexible and can be tailored to the research questions of interest. Standard statistical methods can thus be directly applied on ilr coordinates, which is a common practice referred to as *working on coordinates* in the CoDa literature (Mateu-Figueras et al., 2011).

Ilr coordinates can be easily formed from a *sequential binary partition* (SBP) of components and are then called *balances* (Egozcue and Pawlowsky-Glahn, 2005). A SBP consists in selecting which parts contribute to the logratio and deciding if these will appear in the numerator or in the denominator. To create the first balance, the complete composition $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_D)$ is split into two subsets of components: one for the numerator and the other for the denominator. In the following step, one of the two subsets is further split into two new subsets to create the second ilr coordinate. In step k when the y_k balance is created, a subset containing $r_k + s_k$ parts is split into two: the r_k parts $(x_{n1,}\ldots,\ x_{nr})$ in the first subset are placed in the numerator, and the s_k parts $(x_{d1,}\ldots,x_{ds})$ in the second subset appear in the denominator. The obtained balance is a normalised logratio of the geometric means of each subset of parts (Egozcue et al., 2003)

$$y_k = \sqrt{\tfrac{r_k s_k}{r_k+s_k}}\ln\frac{(x_{n1}\cdots x_{nr})^{1/r_k}}{(x_{d1}\cdots x_{ds})^{1/s_k}} = \sqrt{\tfrac{r_k s_k}{r_k+s_k}}\left(\frac{\ln(x_{n1})+\cdots+\ln(x_{nr})}{r_k} - \frac{\ln(x_{d1})+\cdots+\ln(x_{ds})}{s_k}\right),\ k = 1,\cdots,D-1\ , \tag{2.3}$$

where $\sqrt{\frac{r_k s_k}{r_k+s_k}}$ is the factor that normalises coordinates. Remarkably, the coefficients in the logcontrast expression (2.2) of the kth *ilr*-coordinate (2.3) are $\psi_{jk} = \sqrt{\frac{s_k}{r_k(r_k+s_k)}}$ if the part is placed in the numerator, $\psi_{jk} = -\sqrt{\frac{r_k}{s_k(r_k+s_k)}}$ for parts appearing in the denominator, and $\psi_{jk} = 0$ for parts appearing nowhere.

Let $\ln(\mathbf{x_i})$=($\ln(x_{i1})$, …, $\ln(x_{iD})$) be a row vector containing the D log absolute compo-nents for the ith individual and let $\mathbf{y_i} = (y_{i1}, \ldots, y_{iD-1})$ be the row vector containing its corresponding $D-1$ balances. The transformation matrix $\mathbf{\Psi}$ with D rows and D-1 columns that yields $\mathbf{y_i}$=$\ln(\mathbf{x_i}) \cdot \mathbf{\Psi}$ is

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{11} & \cdots & \psi_{1D-1} \\ \vdots & & \vdots \\ \psi_{D1} & \cdots & \psi_{DD-1} \end{pmatrix}, \tag{2.4}$$

with orthonormal columns; that is, with columns having unit sums of squares and zero scalar products. The $\mathbf{\Psi}$ matrix can be interpreted as an orthonormal projection matrix from $\ln(\mathbf{x_i}) \in R^D$ to a *D-1* dimensional subspace orthogonal to the unit vector $\mathbf{1}$=(1,1,…,1), which is isometric to the simplex (Egozcue et al., 2003).

Each possible SBP leads to a different $\mathbf{\Psi}$ matrix, and ilr coordinates have to be interpreted with respect to the chosen partition. A positive relation of the ilr coordinate with an external dependent variable implies that increases in the group of parts in the numerator (or decreases in the group of parts in the denominator) tend to occur together with increases in the external variable. Parts can be partitioned in such a way that the relationships between the balances and external variables are related to hypotheses or research questions of interest. Once balances have been computed, a linear statistical model to relate an external metric dependent variable w to a composition acting as explanatory variable for the ith individual is

$$w_i = \alpha_0 + \alpha_1 y_{i1} + \ldots + \alpha_{D-1} y_{iD-1} + u_i = \alpha_0 + \mathbf{y_i} \cdot \boldsymbol{\alpha} + u_i, \tag{2.5}$$

where $\alpha_0$is the constant term, u_i is the disturbance term and $\boldsymbol{\alpha}$ are the regression coefficients of the balances arranged in a column vector.

3 Compositions with a total as explanatory variables

3.1 Classical approach

The classical linear model to predict a metric variable w from absolute component size normally uses the logarithms of parts. The log transform is favoured in many scientific fields, in order to restore positive variables into the full real range (Pawlowsky-Glahn et al., 2015b), correct positive skewness, and bring large outliers closer to the centre of the distribution. In economics it is also favoured because it is aligned with the economic thinking in terms of elasticities (e.g. Thrane, 2014). The model can be written as

$$w_i = \gamma_0 + \gamma_1 ln(x_{i1}) + \ldots + \gamma_D ln(x_{iD}) + u_i = \gamma_0 + ln(\mathbf{x_i}) \cdot \boldsymbol{\gamma} + u_i. \tag{3.1}$$

This model has an interpretational drawback. γ_1 to γ_D refer to the effect of increasing the logarithm of one component while keeping the remaining components constant, which is the combined effect of increasing the relative importance of that component and increasing absolute size, and thus leads to confounding relative and absolute information. In other words, all parameters are related to both absolute and relative importance.

3.2 Approach by the T−space

Pawlowsky-Glahn et al. (2015b) study the properties of the T−space defined by a composition and a total. They state that $D-1$ ilr coordinates $\mathbf{y}_\mathrm{i}$ together with a total t_i computed as $\sqrt{D}$ times the logarithm of the geometric mean of all absolute values

$$t_i = \sqrt{D}\ln\left(\sqrt[D]{x_{i1}x_{i2}\cdots x_{iD}}\right) = \frac{1}{\sqrt{D}}\left(\ln(x_{i1}) + \ln(x_{i2}) + \cdots + \ln(x_{iD})\right) \ , \tag{3.2}$$

lead to the same distances among individuals as in the space of the logarithms of absolute values. Note that t_i can be interpreted as the projection of $\ln(\mathbf{x}_\mathrm{i})$ to the unit normalized vector $(1/\sqrt{D})\mathbf{1}$. Using the ilr coordinates and this total as predictors leads to the model

$$w_i = \beta_0 + \beta_1 y_{i1} + \ldots + \beta_{D-1} y_{iD-1} + \beta_D t_i + u_i = \beta_0 + (\mathbf{y_i}\ t_i)\cdot\boldsymbol{\beta} + u_i. \tag{3.3}$$

The $(\mathbf{y_i}\, t_i)$ vector is formed by the balances $\mathbf{y_i}$ augmented with the total t_i. Compared to the model (2.5), in (3.3) $\boldsymbol{\beta}$ has one more coefficient.

The global F statistic in this model tests the hypothesis that the explanatory vector, all things considered (relative and absolute information) has no effect on w. Individual tests of β_1 to β_{D-1} refer to the relative importance of components and are interpreted with respect to the particular balances, as the effect of increasing one balance, while keeping the remaining balances and total constant. The fact that the remaining balances are held constant implies that parts in the numerator of the balance increase all in the same proportion and parts in the denominator decrease all in the same proportion. The fact that the total is kept constant implies that the increase in the numerator is exactly offset by the decrease in the denominator. The tests of β_1 to β_{D-1} depend on the SBP constructed and can be used to answer the research questions which guided balance construction. It is not recommended to use these tests to remove non-significant balances, because it can affect the interpretation of the kept balances. If all $D-1$ balances are in the model, then the test of β_D is interpreted with respect to increasing overall size while leaving relative importance of components constant, i.e. increasing all absolute component sizes in the same proportion.

Some properties of the CoDa model with a total (3.3) are:

- The global F statistic gives the same result as (3.1) and is SBP invariant; that is, invariant to the choice of ilr coordinates (balances).
- The overall goodness of fit of the model (e.g. the R$-$squared value, and the residual standard error) is the same as in (3.1) and is SBP invariant.
- The β_D total effect is SBP invariant.

Indeed, we define the transformation matrix $\mathbf{U}$ as

$$(\mathbf{y_i}\ t_i) = \ln(\mathbf{x_i}) \cdot \mathbf{U} \quad \text{and} \quad \ln(\mathbf{x_i}) = (\mathbf{y_i}\ t_i) \cdot \mathbf{U^{-1}}, \tag{3.4}$$

where

$$\mathbf{U} = \begin{pmatrix} & 1\big/\sqrt{D} \\ \mathbf{\Psi} & \vdots \\ & 1\big/\sqrt{D} \end{pmatrix} = \begin{pmatrix} \psi_{11} & \cdots & \psi_{1D-1} & 1\big/\sqrt{D} \\ \vdots & & \vdots & \vdots \\ \psi_{D1} & \cdots & \psi_{DD-1} & 1\big/\sqrt{D} \end{pmatrix}. \tag{3.5}$$

It must be noted that $\mathbf{U^T U}$=$\mathbf{I}$. The $\mathbf{U}$ matrix is an orthonormal change of basis matrix from $\ln(\mathbf{x_i})$ into the R^D space. From this it follows that $\mathbf{U^{-1}} = \mathbf{U^T}$ and

$$w_i = \ \gamma_0 + \ln(\mathbf{x_i}) \cdot \boldsymbol{\gamma} + u_i = \ \gamma_0 + (\mathbf{y_i}\ t_i) \cdot \mathbf{U^T} \cdot \boldsymbol{\gamma} + u_i = \beta_0 + (\mathbf{y_i}\ t_i) \cdot \boldsymbol{\beta} + u_i, \quad (3.6)$$

$$w_i = \beta_0 + (\mathbf{y_i}\ t_i) \cdot \boldsymbol{\beta} + u_i = \ \beta_0 + \ln(\mathbf{x_i}) \cdot \mathbf{U} \cdot \boldsymbol{\beta} + u_i = \ \gamma_0 + \ln(\mathbf{x_i}) \cdot \boldsymbol{\gamma} + u_i, \quad (3.7)$$

from which we conclude that $\gamma_0 = \beta_0$; $\boldsymbol{\gamma} = \mathbf{U}\cdot\boldsymbol{\beta}$; $\boldsymbol{\beta}$ =$\mathbf{U^T} \cdot \boldsymbol{\gamma}$ and that both models are equivalent, yielding identical predicted values and disturbances. Any global test or goodness of fit measure which is a function of predicted values, observed values and disturbances will yield the same results in the classical model (3.1) and the model (3.3) using the total (3.2). Any estimation method optimising a function of predicted values, observed values and disturbances will yield the same optimum, and estimates will be equivalent following the relationships (3.4), (3.6) and (3.7).

By construction of the last row of $\mathbf{U^T}$, the total effect β_D is

$$\beta_D = \frac{1}{\sqrt{D}}\gamma_1 + \cdots + \frac{1}{\sqrt{D}}\gamma_D \ . \quad (3.8)$$

This holds regardless of the $\boldsymbol{\Psi}$ matrix and SBP constructed. From the expression (3.8) it follows that the total (3.2) will have greater predictive power the further the sum of the coefficients in $\boldsymbol{\gamma}$ is from zero. On the contrary, the sum of the $\boldsymbol{\gamma}$ vector equal to zero indicates that the total is not informative.

Model (3.1) and model (3.3) are thus equivalent. Both models are also subject to the same well-known statistical and distributional assumptions. The choice for one or the other depends only on the ease of interpretation. The model (3.3) makes it easier to test hypotheses on absolute and relative importance separately. Besides, the hypotheses about relative importance can be chosen by the researcher in the SBP.

The CoDa model with a total (3.3) can be extended to a generalized lineal model to deal with a count w variable (e.g. Poisson or negative binomial regression), a binary w variable (e.g. logit or probit model), or an ordinal w variable (e.g. ordered logit or probit model). The total (3.2) recommended by Pawlowsky-Glahn et al. (2015b) best separates absolute and relative information and we advise researchers to use it. Below we show how alteration or omission of the total affect the model properties and interpretation.

4 Different submodels

4.1 Composition-only model: consequences of omitting the total

Standard CoDa is equivalent to fitting (2.5), which we refer to as composition-only model. It can thus be understood as a particular case of CoDa with a total when $\beta_D=0$, or, following (3.8), as a particular case of the classical approach (3.1) in which the sum of the coefficients in $\boldsymbol{\gamma}$ is zero.

All known results about the removal of variables in a generalized linear model can thus be applied here. If the total has a non-zero β_D effect on the dependent variable in the population, omitting it constitutes a model misspecification which can bias the effect estimates involving balances, more so if balances and total are correlated.

If data are purely compositional (i.e. if absolute information is truly arbitrary and irrelevant) then the total is expected both to be uncorrelated with the balances and with any variable of interest.

Omitting the total when relevant affects parameter interpretation. The α_j param-eter relates to the effect of increasing all components in the numerator by a given proportion, while decreasing all components in the denominator by another given proportion. However, the interpretation is ambiguous: it is not known what happens with the total, which will depend on the unknown relationship between y_j and t. If this relationship is zero, the problem disappears and a purely compositional analysis is correct. If this relationship is positive, the parts in the numerator are inadvertently increasing to a greater extent than the parts in the denominator are being reduced; if negative, it is the other way around.

Actually, the statistical significance of the β_D effect of the total can be understood as evidence that the absolute value of components does matter to the research question. Only if the total is not significant can data be held as purely compositional and is the standard CoDa model (2.5) reasonable.

4.2 Total-only model

One could also consider a model in which relative information does not matter, which we term total-only model

$$w_i = \beta_0 + \beta_D t_i + u_i. \tag{4.1}$$

This model is a legitimate option if the distribution of the total among parts has no effect on the dependent variable. However, the interpretation of the coefficient is once more unclear. It is related to the effect of increasing the product of components as a whole, by unknown, equal or unequal factors.

Analogously to the composition-only model, the model (4.1) can also be related to the classical model (3.1). Since by construction of $\mathbf{U}^{\mathbf{T}}$ all rows from 1 to D-1 have zero sum, when all γ coefficients in (3.1) are equal it follows that $\beta_1 = ... = \beta_{D-1} =0$.

4.3 Nested model hierarchy

The models (2.5) and (4.1) are nested into the model (3.3) and can thus be compared to it in order to test for the relevance of relative and absolute information when predicting w.

The test of the full model (3.3) against the composition-only model (2.5) refers to the null hypothesis that total does not matter (H$_0$: β_D =0). The test of the full model (3.3) against the total-only model (4.1) refers to the null hypothesis that composition does not matter (H$_0$: $\beta_1 = \beta_2 = \dots = \beta_{D-1} =0$).

We suggest using the model hierarchy for hypothesis testing rather than for model selection. In other words, we suggest keeping the full model, whose interpretation is always clear, regardless of the test outcomes. Even if not statistically rejected, the composition-only or total-only models may be misspecified to some extent in the population.

Of course, we recommend users not to base their conclusions on tests alone. Some goodness of fit measure of the three models should tell to what extent size and composition contribute predictive power. Ideally, such fit measure should take parsimony into account (e.g. AIC, BIC, and adjusted R-squared).

4.4 Consequences of including the total as a sum

Pawlowsky-Glahn et al. (2015b) discussed the possibility of using a total based on its usual understanding as a sum of parts, rather than the product-based total (3.2)

$$t_i' = \ln(x_{i1} + x_{i2} + \cdots + x_{iD}) \ . \tag{4.2}$$

It must be noted that this is the logarithm of the sum used in the closure operation (1.2).

When compositions with a total are used as dependent variables, the particular definition of the total does not modify equations relating to balances (Ferrer-Rosell et al., 2016a). However, when used as explanatory, the model (2.5) combined with the total (4.2) does not yield the same results as the model (3.3). Even if there is nothing inherently wrong with (4.2) and at first sight a total as a sum looks intuitively appealing, parameter estimates and interpretation do change and many readers may find the interpretation of the model with the total (3.2) simpler. In a similar way as logarithms in (2.1), (2.3), (3.1) and (3.2), the logarithm in (4.2) has the twofold objective of transforming positive values into the whole real space and shifting the focus from absolute to relative changes. In this context, the sum is a hard-to-interpret operation, while products and geometric averages are natural in a log scale.

When the total sum (4.2) is used, the tests of β_1 to β_{D-1} refer to the effect of in-creasing all components in the numerator by a given proportion, while decreasing all components in the denominator by another given proportion. However, the interpretation is once more ambiguous. Are the components in the numerator increasing to a greater or lesser extent, proportionally speaking, than the components in the denominator are being reduced? Once more, absolute and relative information are not well separated.

In addition, the model with the total (4.2) is no longer equivalent to the model (3.1). If the researcher finds (3.1) to be a good representation of the reality to be modelled, then fitting the model with (3.2) and (3.3) is the natural thing to do and is subject to the same distributional assumptions.

On the other hand, if the researcher believes the sum of parts to be a relevant characteristic of individuals, then (4.2) may be preferred to (3.2). A very pragmatic approach is also possible: if the researcher's objective is merely to predict the dependent variable as accurately as possible, he or she may simply choose between (4.2) and (3.2) on the basis of explanatory power alone.

4.5 Consequences of not including all absolute information in the total

Ferrer-Rosell et al. (2016a) show that in a dependent role, the geometric mean of any number of absolute values (even just one absolute value) can be used as a total. When compositions and a total are used as explanatory, any modification in the total changes what is held constant when interpreting the balances and the interpretation of the coefficients is once more rather counterintuitive. Let us assume without loss of generality that the total includes the geometric mean of the absolute values of the first s parts. In spite of defining

$$t_i'' = \sqrt{s}\ln\left(\sqrt[s]{x_{i1}x_{i2}\cdots x_{is}}\right) = \frac{1}{\sqrt{s}}\left(\ln(x_{i1}) + \ln(x_{i2}) + \cdots + \ln(x_{is})\right) \ , \qquad (4.3)$$

β_D is not related to the first s parts but to all of them. Multiplying all parts simultaneously by the same constant is the only way in which t''_i can be increased while leaving all balances constant. Besides, the last column in (3.5) is no longer orthogonal with respect to the first D-1 and hypotheses on balances are once more not well separated from the hypotheses on the total.

Interestingly, the goodness of fit of the model and the overall significance test are not modified by the inclusion of only some parts in the total.

5 An example in tourism economics

5.1 Background

This example deals with tourist expenditure on three components which constitute the three major parts of a trip budget: transportation, accommodation & food, and activities & shopping. It constitutes a clear case in which both total expenditure and the way it is distributed is interesting to tourism scholars and managers (Ferrer-Rosell et al., 2016a). This notwithstanding, tourist expenditure research has largely ignored budget distribution, ignored budget total, or confounded both. The vast majority of tourist expenditure studies take into account one single expenditure variable (Brida and Scuderi, 2013) and thus ignore budget distribution. Some studies focus on trip budget distribution among parts, thus ignoring trip budget total, by means of CoDa (e.g. Ferrer-Rosell et al., 2016b) or other analysis methods (e.g. almost ideal demand systems, see Lee et al., 2015). Finally, some studies focus on absolute trip expenditure per trip budget part. Since part expenditure in absolute terms belongs to total expenditure, budget distribution and budget total are confounded: a large absolute expenditure on a given trip budget part may correspond either to a tourist with a large overall budget or to a tourist who particularly tends to spend on that part. A repeated finding in this type of analysis is that some variables are related to all budget parts in the same direction.

In this illustration we relate tourist expenditure allocation among budget parts and total expenditure to trip satisfaction. In the literature, trip satisfaction has been used both as a predictor (Brida and Scuderi, 2013) and as an outcome (Ferrer-Rosell et al., 2017) of spending behaviour. In the latter case, to the best of our knowledge only total expenditure has been considered, and the reported effect has been positive.

We use an ordinal dependent variable: a question on overall trip satisfaction rated from 0 to 10. Accordingly, we fit an ordered logistic regression (logit) model.

5.2 Data and balances

We use official statistics microdata from the *EGATUR* tourist expenditure survey conducted by the Spanish Ministry of Industry, Energy and Tourism (ITE, 2014). We consider European leisure visitors arriving to Spain by air in 2012 and spending between 1 and 120 nights in a single destination in that country. As Ferrer-Rosell et al. (2015, 2016a,b) we exclude tourists for whom expenditure distribution among budget components is partly or completely unobserved (mainly those staying with friends/relatives or in an owned apartment, and package tourists). The sample size is 19142.

We focus on D= 3 expenditure components:

- x_1 Euro spent on transportation,
- x_2 Euro spent on accommodation and food at destination,
- x_3 Euro spent on activities and shopping at destination.

An interpretable ilr transformation is easy to compute whenever there is an interpretable SBP of components according to the researchers' questions. These partitions are best understood with a dendrogram (Pawlowsky-Glahn and Egozcue, 2011). The dendrogram (Figure 1) we use is related to research questions concerning:

- The effect on satisfaction of how tourists distribute total expenditure between transportation and at-destination expenditure.
- The effect on satisfaction of how tourists distribute at-destination expenditure into accommodation and food versus activities and shopping.

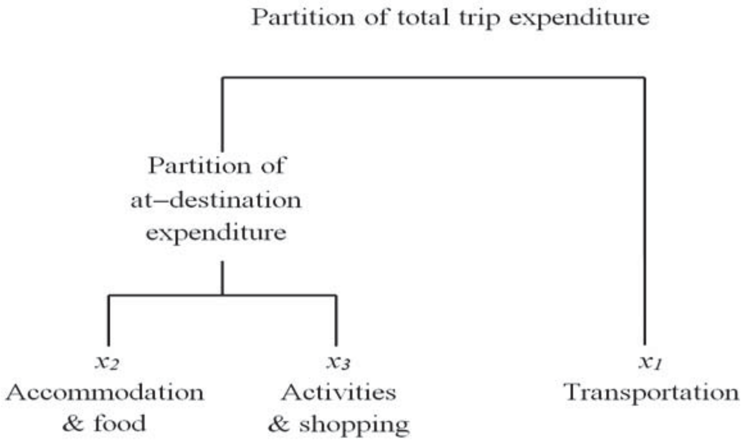
Partition of total trip expenditure
Partition of at–destination expenditure
x2
Accommodation & food
x3
Activities & shopping
x1
Transportation

Figure 1 SBP of trip budget parts.
Source: Authors' own.

The implied ilr coordinates (balances) are

$$\begin{aligned} y_1 &= \sqrt{\tfrac{2}{3}} \ln\left(\tfrac{x_1}{\sqrt{x_2 x_3}}\right) = \sqrt{\tfrac{2}{3}} \ln(x_1) - \tfrac{1}{2}\sqrt{\tfrac{2}{3}} \ln(x_2) - \tfrac{1}{2}\sqrt{\tfrac{2}{3}} \ln(x_3) \\ y_2 &= \sqrt{\tfrac{1}{2}} \ln\left(\tfrac{x_2}{x_3}\right) = \sqrt{\tfrac{1}{2}} \ln(x_2) - \sqrt{\tfrac{1}{2}} \ln(x_3) \; . \end{aligned} \qquad (5.1)$$

Adding the total (3.2) to (5.1) leads to the following transformation matrix $\mathbf{U}$ (3.5)

$$\begin{pmatrix} y_{i1} & y_{i2} & t_i \end{pmatrix} = \begin{pmatrix} \ln(x_{i1}) & \ln(x_{i2}) & \ln(x_{i3}) \end{pmatrix} \begin{pmatrix} \sqrt{\frac{2}{3}} & 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{2}\sqrt{\frac{2}{3}} & \sqrt{\frac{1}{2}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{2}\sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{2}} & \frac{1}{\sqrt{3}} \end{pmatrix} . \qquad (5.2)$$

Model parameters have to be interpreted in the context of each application, taking some care in understanding what the other terms in the model are controlling for (i.e. keeping constant when interpreting the parameter). In our model (3.3) using total (3.2) and balances (5.1):

- β_1 is associated to the effect of multiplying x_1 by a constant $a>1$, while multiplying x_2 and x_3 simultaneously by $1/\sqrt{a}$. The reader will notice that this is the only way in which y_2 and t can be held constant while varying y_1.

- β_2 is associated to the effect of multiplying x_2 by a constant $a>1$, while multiplying x_3 by the inverse of the same constant $1/a$. The reader will notice that this is the only way in which y_1 and t can be held constant while varying y_2.

- β_3 is associated to the effect of multiplying x_1, x_2 and x_3 simultaneously by the same constant $a>1$. The reader will notice that this is the only way in which y_1 and y_2 can be held constant while varying t.

5.3 Results

The results of the classical approach (3.1) show positive effect estimates of all log-parts, which may be the result of confounding total and allocation (model 1 in Table 1). The global χ^2 likelihood ratio test of all the coefficients in (3.3) using the total (3.2) is significant (model 2 in Table 1). This test coincides with model 1 and is telling that expenditure, all things considered, is related to trip satisfaction.

The results of the individual parameter tests in model 2 show that trip satisfaction significantly increases when relative importance of transportation expenditure decreases compared to at-destination expenditure (y_1), when relative importance of accommodation and food within at-destination expenditure increases (y_2), and when all components of the total budget increase by the same proportion (t).

Omitting the significant total as in (2.5) modifies the estimates of the balance coefficients to a substantial extent, which in this case is interpreted as omitted variable bias (composition-only model 3). Model 4 is the total-only model (4.1).

The restricted models 3 and 4 in Table 1 can be compared to the full model 2 in order to test for the relevance of absolute and relative information, respectively. When comparing the nested models 3 and 2, the likelihood ratio test computed as the χ^2 difference is 131.680−35.012=96.668 with 1 d.f. (p−value<0.0005), leads to rejecting model 3 and H$_0$: β_3 =0 and to concluding that absolute expenditure has a non-zero effect on satisfaction. When comparing the nested models 4 and 2, the χ^2 difference is 131.680−113.421=18.259 with 2 d.f. (p−value<0.0005), leads to rejecting model 4 and H$_0$: $\beta_1 = \beta_2$ =0 and to concluding that relative expenditure has a non-zero effect on satisfaction. However, the results of the tests do not tell about predictive power in practical terms. When we look at the BIC values in Table 1, the total-only model appears to be preferable, thus telling that predictive power of relative expenditure information is negligible. This notwithstanding, we interpret the total effect in the full model, because of its more precise definition in terms of what happens when all components increase by the same proportion.

Models 5 and 6 in Table 1 illustrate the implications of modifying the total. Modifying it as the sum (4.2) in model 5 changes both the estimates of the balance effects and the global goodness of fit. The interpretation of balances in model 2 does not hold anymore.

As an example of a total focusing on specific parts (4.3), Ferrer-Rosell et al. (2016a) defined $t''_i = \ln(x_{i1})$ within a study on transportation economics. If t'' is included as explanatory, then the global χ^2 likelihood ratio test is the same as in model 2 but the interpretation of some of the coefficients is rather counterintuitive (model 6). In spite of defining $t''_i = \ln(x_{i1})$, β_3 is not related to transportation (x_1) but is proportional to the effect of multiplying x_1, x_2 and x_3 simultaneously by the same constant, and the β_3 test statistic is equal to that of model 2. β_1 is associated to the effect of increasing x_2 and x_3 while keeping $ln(x_1)$ constant, and thus leads to confounding relative and absolute information.

Table 1 Tests and estimates of alternative specifications. (*) Global χ^2 test of $\beta_1 = \beta_2 = \beta_3 = 0$ or $\gamma_1 = \gamma_2 = \gamma_3 = 0$: 131.680 with 3 d.f.; p−value<0.0005. (**) Global χ^2 test of $\alpha_1 = \alpha_2 = 0$: 35.012 with 2 d.f.; p−value<0.0005. (***) χ^2 test of $\beta_3 = 0$: 113.421 with 1 d.f.; p−value<0.0005. (****) Global χ^2 test of $\beta_1 = \beta_2 = \beta_3 = 0$: 158.183 with 3 d.f.; p−value<0.0005.

			Std.	Estimate/	Test
Model and variables		Estimate	error	Std. error	p−value
1) Classical (3.1, BIC=52142) *					
	$\ln(x_1)$	0.030	0.024	1.274	0.203
	$\ln(x_2)$	0.163	0.021	7.600	0.000
	$\ln(x_3)$	0.050	0.011	4.639	0.000
2) Full (3.2 and 3.3, BIC=52142) *					
	y_1	-0.062	0.024	-2.569	0.010
	y_2	0.080	0.019	4.300	0.000
	t	0.140	0.014	9.968	0.000
3) Composition-only (2.5, BIC=52108)**					
	y_1	-0.132	0.023	-5.717	0.000
	y_2	0.033	0.018	1.826	0.068
4) Total-only (3.2 and 4.1, BIC=46052)***					
	t	0.130	0.012	10.876	0.000
5) $t' = \ln(x_1 + x_2 + x_3)$ (3.3 and 4.2, BIC=52115)****					
	y_1	-0.041	0.025	-1.681	0.093
	y_2	-0.009	0.018	-0.465	0.642
	t'	0.271	0.024	11.265	0.000
6) $t'' = \ln(x_1)$ (3.3 and 4.3, BIC=52142)*					
	y_1	-0.261	0.026	-9.918	0.000
	y_2	0.080	0.019	4.300	0.000
	t''	0.243	0.024	9.968	0.000

Source: Authors' own.

6 An example on urban solid waste and floating population

The actual population residing in a municipality is composed by the census count and the so-called floating population (tourists, seasonal visitors, hostel students, short-time employees, and the like). Floating population may be positive if the municipality is receiving more short term residents than it is sending elsewhere, or negative if the opposite holds. It is usually expressed as a percentage above (if positive) or below (if negative) the census count.

Floating population, including tourist population, has a large impact on solid waste generation (Mateu-Sbert et al., 2013) and thus overall solid waste can be used to predict floating population (Mateu-Sbert et al., 2013). Tourists and census population do not generate the same amount of waste (Mateu-Sbert et al., 2013) and may have different recycling patterns (Mendes et al., 2013), which calls for considering both waste total and composition.

The composition of solid waste has been studied by means of CoDa (Pivnenko et al., 2016). In this illustration we show how both absolute size (tons per census inhabitant) and composition of urban solid waste can be used to proxy floating population.

The Catalan Statistical Institute (IDESCAT) publishes floating population data for all municipalities in Catalonia (Spain) above 5000 inhabitants, together with solid waste weight classified into D =5 components: non recyclable (x_1, grey waste con-tainer in Catalonia), glass (x_2, bottles and jars of any colour −green container), light containers (x_3, plastic packaging, cans and tetra packs −yellow container), paper and cardboard (x_4, blue container) and biodegradable waste (x_5, brown container). Figure 2 shows a possible SBP tree of urban solid waste. We use data for 2014 (n =215 municipalities).

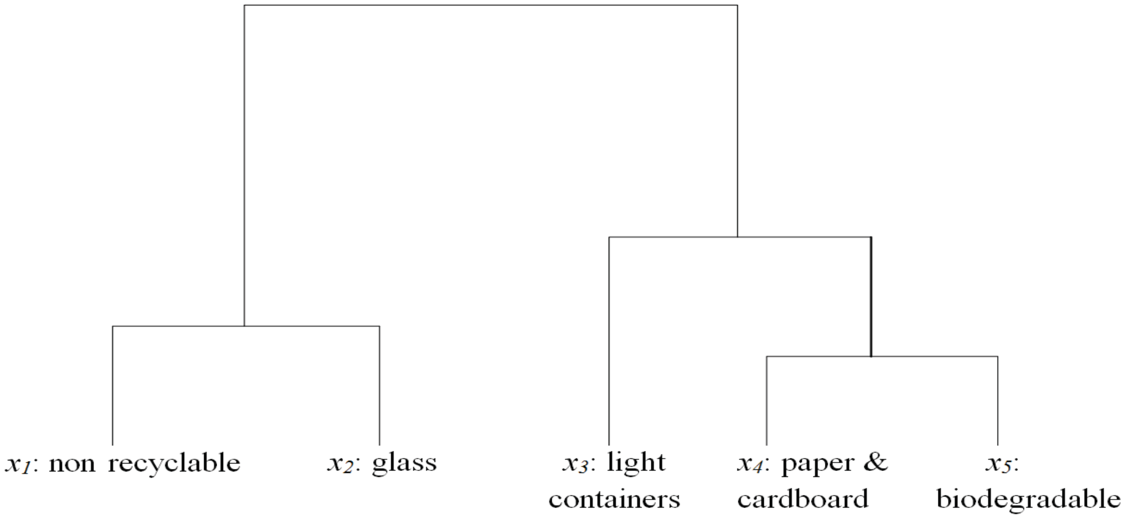

Figure 2 SBP of urban solid waste parts.
Source: Authors' own.

Since the dependent variable is numeric, a linear regression model can be appropriate. Table 2 shows the estimates of the model (3.3) with the total (3.2). The adjusted R −squared is high at 63.1%

Table 2 Test and estimates of the effects of solid waste on floating population. Global F test of $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ =0: 74.117 with 5 and 209 d.f.; p−value<0.0005.

	Estimate	Std. error	Estimate/ Std. error	Test p−value
Balance non-rec. & glass over all other (y_1)	19.244	2.312	8.324	0.000
Balance containers over bio. & paper (y_2)	-8.185	3.788	-2.161	0.032
Balance non-rec. over glass (y_3)	6.184	3.569	1.733	0.085
Balance bio. over paper (y_4)	1.993	2.105	0.947	0.345
t	23.045	1.509	15.272	0.000

Source: Authors' own.

The nested test of this full model against the total-only model leads to rejecting

H$_0$:$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (F=66.645 with 4 and 209 d.f.; p$-$value<0.0005) while the nested test of this full model against the composition-only model leads to rejecting H$_0$:$\beta_5 = 0$ (F=233.23 with 1 and 209 d.f.; p$-$value<0.0005). In this example both composition and total have substantial predictive power: the total-only model has an adjusted R$-$squared equal to 31.7% and the composition-only model 22.2%.

The fact that a higher floating population can be predicted from a higher waste total was to be expected. However, waste composition helps making substantially better predictions. A higher floating population can also be expected from a lower balance of light containers over paper, cardboard and biodegradable waste (y_2), and from a larger balance of non-recyclable and glass waste over the three types above (y_1). Floating population not only increases waste in the transient municipality, but also has different consumption or recycling patterns, compared to permanent population.

7 Final remarks

In this article we show and illustrate how relative and absolute information can be combined and used to explain other variables of interest by means of CoDa with a total. While being equivalent to modelling all log absolute values, this approach has the advantage that tests of the effect of relative importance of parts and of the effect of absolute size are separated. The approach uses, on the one hand, the logarithm of the geometric mean of all absolute values; on the other, $D-1$ balances obtained from a SBP, in other words, ilr coordinates, exactly as in standard CoDa. Nested models can be used to test the relevance of each. It is assumed that absolute size is not constant and data are available in their original $\mathbf{x}$ form prior to carrying out the closure operation. CoDa with a total is of no interest on closed data.

Parameters are easy to interpret and can be tailored to the researcher's questions and hypotheses. When D is large, it may indeed be difficult to derive the SBP from research questions. An alternative approach is to construct the SBP from the data by means of the so-called principal balances (Pawlowsky-Glahn et al., 2011).

Extensions to non-normal dependent variables are immediate by means of generalized linear models, and, once balances and total have been computed, any standard software handling generalized linear models will do the job.

Using a total other than the geometric mean of all absolute values leads to some degree of confounding between relative and absolute information, while dropping the total makes interpretation ambiguous, and even biased if the total is actually relevant. As a word of caution, including absolute information is advisable even if it is not of interest to the researcher, for the sake of a clearer and unbiased interpretation of the balance effects.

Applications of the method include situations in which the researcher is interested both in relative and absolute information, and situations in which the main interest lies in relative information but the relevance of absolute information (assumed to be available) cannot be ruled out. They may include research in such diverse fields as web content analysis and mining (e.g. number of postings within each content category or containing each term), bacteria or pollutants (e.g. abundance and distribution per types or species), household budgets (e.g. total expenditure and its allocation), forestry management (e.g. forest density and species distribution), marketing (e.g. market share and sales), finance (e.g. balance sheet analysis: liquidity and leveraging ratios, and total assets), quality control (e.g. defect or customer complaint count and type), ecology (e.g. abundance and distribution of resources and species), and dietetics/nutrition (e.g. fat total content by fat type), among others.

Previous research has already dealt with the case in which composition and a total act as dependent variables. Further research can firstly extend the findings in this article to statistical models in which compositions and a total are at the same time dependent and explanatory, or in which there may even be more than one composition. This includes techniques such as structural equation models and partial least squares, for which standard CoDa is already in place (Kalivodová et al., 2015; Kogovšek et al., 2013).

Secondly, further research can include developing comparable measures of effect size supplementing the information provided by the statistical tests and the goodness of fit measures, in order to better gauge the practical relevance of each model parameter. Finally it can include testing, and if necessary adapting, the zero treatment methods in CoDa when a total is present.

Acknowledgements

The authors are pleased to acknowledge the support of the Instituto de Turismo de España in providing them with the raw data for the first illustration. The authors have all responsibility for any lack of accuracy or reliability of their data analysis. The authors acknowledge the support of the Catalan Autonomous Government Consolidated Research Group Grant 2014SGR551 funding the research group "Compositional and Spatial Data Analysis (COSDA)"; the Spanish Health Ministry grant CB06/02/1002 funding the research group “CIBER of Epidemiology and Public Health (CIBERESP)”; the FEDER/Spanish Economy and Competitiveness Ministry grant MTM2015-65016-C2-1-R funding the project “COmpositional Data Analysis and RElated meThOdS (CODA-RETOS)”; and the University of Girona grants MPCUdG2016/069 and MPCUdG2016/098.

References

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, UK: Chapman and Hall Ltd.

Barceló-Vidal C and Martín-Fernández JA (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, **45**, 57–71.

Brida JG and Scuderi R (2013). Determinants of tourist expenditure: a review of microeconometric models. *Tourism Management Perspectives*, **6**, 28–40.

Di Marzio M, Panzera A and Venieri C (2015). Non-parametric regression for compositional data. *Statistical Modelling*, **15**, 113–33.

Egozcue JJ and Pawlowsky-Glahn V (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, **37**, 795–828.

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

Ferrer-Rosell B, Coenders G and Marine-Roig E (2017). Is planning through the Internet (un)related to trip satisfaction? *Information Technology & Tourism*, **17**. DOI: 10.1007/s40558-017-0082-7

Ferrer-Rosell B, Coenders G and Martínez-Garcia E (2015). Determinants in tourist expenditure composition: the role of airline types. *Tourism Economics*, **21**, 9–32.

Ferrer-Rosell B, Coenders G and Martínez-Garcia E (2016b). Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. *Tourism Analysis*, **21**, 589–602.

Ferrer-Rosell B, Coenders G, Mateu-Figueras G and Pawlowsky-Glahn V (2016a). Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics*, **22**, 269–91.

ITE (2014). Tourist Expenditure Survey. Methodology. *Madrid (ES): Instituto de Turismo de España. Downloaded 27 November 2015 from: http://estadisticas.tourspain.es/es-ES/estadisticas/egatur/metodologia/Referencia Metodolgica/Nota Metodológica Encuesta de Gasto Turístico.pdf.*

Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečcková H and Adam T (2015). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, **29**, 21–28.

Kogovšek T, Coenders G and Hlebec V (2013). Predictors and outcomes of social network compositions. A compositional structural equation modeling approach. *Social Networks*, **35**, 1–10.

Lee S, Jee W, Funk D and Jordan J (2015). Analysis of attendees' expenditure patterns to recurring annual events: examining the joint effects of repeat attendance and travel distance. *Tourism Management*, **46**, 177–86.

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**, 253–278.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P and Palarea-Albaladejo J (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, **15**, 134–58.

Martín-Fernández JA, Palarea-Albaladejo J and Olea RA (2011). Dealing with zeros. In Pawlowsky-Glahn V and Buccianti A (eds), *Compositional data analysis: Theory and applications*. Chichester, UK: John Wiley & Sons, pp. 47-62.

Mateu-Figueras G, Pawlowsky-Glahn V and Egozcue JJ (2011). The principle of working on coordinates. In Pawlowsky-Glahn V and Buccianti A (eds), *Compositional data analysis: Theory and applications*. Chichester, UK: John Wiley & Sons, pp. 31-42.

Mateu-Sbert J, Ricci-Cabello I, Villalonga-Olives E and Cabeza-Irigoyen E (2013). The impact of tourism on municipal solid waste generation: the case of Menorca Island (Spain). *Waste management*, **33**, 2589–93.

Mendes P, Santos AC, Nunes LM and Teixeira MR (2013). Evaluating municipal solid waste management performance in regions with strong seasonal variability. *Ecological Indicators*, **30**, 170–77.

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.

Pawlowsky-Glahn V and Buccianti A, eds (2011) *Compositional Data Analysis: Theory and Applications*. Chichester, UK: John Wiley & Sons.

Pawlowsky-Glahn V and Egozcue JJ (2011). Exploring compositional data with the CoDa-dendrogram. *Austrian Journal of Statistics*, **40**, 103–113.

Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R (2011). Principal balances. In Egozcue JJ, Tolosana-Delgado R and Ortego MI (eds), *Proceedings of the 4th International Workshop on Compositional Data Analysis*. Girona, ES: University of Girona.

Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R (2015a). *Modeling and Analysis of Compositional Data.* Statistics in practice. Chichester, UK: John Wiley & Sons.

Pawlowsky-Glahn V, Egozcue JJ and Lovell D (2015b). Tools for compositional data with a total. *Statistical Modelling*, **15**, 175–190.

Pivnenko K, Eriksena MK, Martín-Fernández JA, Eriksson E and Astrup TF (2016). Recycling of plastic waste: Presence of phthalates in plastics from households and industry. *Waste Management*, **54**, 44–52.

Russell MA (2014). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* Sebastopol, CA: OReilly.

Thrane C (2014). Modelling micro-level tourism expenditure: recommendations on the choice of independent variables, functional form and estimation technique. *Tourism Economics*, **20**, 51–60.

van den Boogaart KG and Tolosana-Delgado R (2013). *Analyzing Compositional Data with R.* Heidelberg, DE: Springer.