Management and Production Engineering Review
Volume 6 • Number 2 • June 2015 • pp. 63–72
DOI: 10.1515/mper-2015-0018

SOME COMMENTS ON COMPOSITIONAL ANALYSIS IN MANAGEMENT AND PRODUCTION ENGINEERING

Marina Vives-Mestres, Josep A. Martín-Fernández

Dept. Computer Science, Applied Mathematics and Statistics, University of Girona, Spain

Corresponding author:
Marina Vives-Mestres
Dept. Computer Science, Applied Mathematics and Statistics
Universitat de Girona (UdG)
Campus Montilivi, Edif. P-IV
Girona (E-17071) Spain
phone: (+34) 972-41-84-17
e-mail: marina.vives@udg.edu

Received: 31 March 2015
Accepted: 12 May 2015

Abstract
This paper introduces the most basic concepts of the compositional analysis of data with a simple but real example form the Management and Production Engineering (MPE) field. Compositional Data (CoDa) are vectors of positive elements that represent parts of a whole and are widely found in MPE, i.e. production times, resource composition, percentage utilization of work stands, waste components... The need for an analysis based on ratios of components (or better log-ratios of components) is illustrated step by step, and findings are compared to the corresponding standard methods applied to raw compositions. The paper also exposes the principles of CoDa analysis and presents two basic descriptive tools suitable for CoDa: the clr-biplot and the CoDa dendrogram. The example is a time series, from 1994 to 2013, of motor vehicle production in 8 countries and regions.

Keywords
production composition, log ratio analysis, clr-biplot, CoDa dendrogram, Ternary diagram, Simplex.

Introduction

In Management and Production Engineering (MPE) it is common to analyse multivariate data which are frequently percentages. Typical examples are data from surveys, among many others, such as the distribution of employment by major industry sector, profiles of consumer expenditure by purchasing power, proportions of labour force by industrial sectors, or distribution of world crude production.

Two typical examples found in the day-to-day operations of MPE are the process yield and the planned production time. In the former, the user could be interested in the distribution of goods in [defect-free, repaired, defective], whereas the latter considers a vector in time such as [operating time, equipment failures, process set-up and adjustments, start-ups after shifts-breaks-lunch-weekends].

In all the above cases it is important to state that, when the total sum of the vector is assumed to be irrelevant, then the focus of the analysis has to be on the relative distribution among the variables. In other words, this multivariate data may be considered as Compositional Data (CoDa) because they describe quantitatively the components of some whole. The components in CoDa are usually termed *parts*.

In the experimental field, CoDa appear as vectors of percentages, parts per unit, parts per million, or other non-closed units, like molar concentrations or absolute frequencies. The units used are irrelevant, because the total sum of the vector is not informative, i.e. the information is relative, rather than absolute, and lies in the ratios of the parts.

There is a general agreement that the sample space of CoDa is the simplex

$$S^D = \left\{ x \in R_+^D : \sum_{j=1}^{D} x_j = k \right\},$$

where the value of k is irrelevant, a popular choice is k=1. When k = 3, the composition lies in an equilat-

eral triangle in R^3 (Fig. 1 top), although it is more common to represent the data in the ternary diagram (Fig. 1 bottom), which is an equivalent representation.

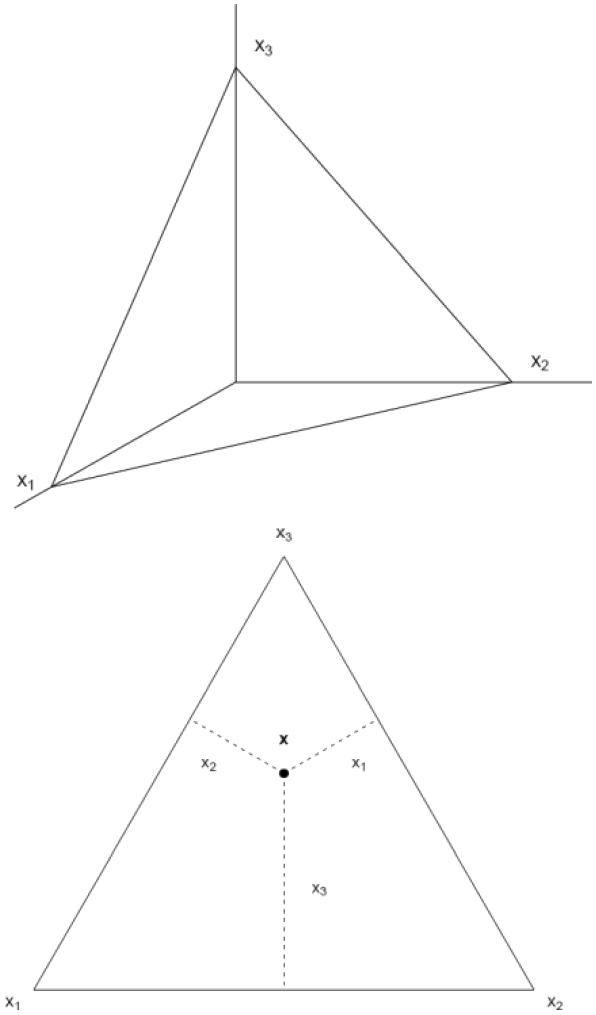

Fig. 1. Two Different but equivalent representations of the Simplex with k = 3 in (top) R^3 and (bottom) in the ternary diagram.

Standard statistical data analysis applied to CoDa may carry technical difficulties and may induce misleading conclusions due to the scale invariance property (i.e., relative information). The classic monograph by Aitchison (1986) [1] introduces the log-ratio methodology, the first consistent methodological proposal to deal with CoDa.

Nowadays there are numerous new ideas and strategies to deal with CoDa analysis. Those advances were presented at the five CoDaWork meetings (e.g. [2]) and collected in some special publications (e.g. [3]) where a review of the state of the art is provided. The main point of this new methodology is the statement of the properties of CoDa analysis.

According to [3], log-ratio analysis can be reduced to three steps, termed principles of working in coordinates [4]: 1. represent CoDa in log-ratio type coordinates; 2. apply standard statistical analysis to the coordinates as real random variables; and 3. interpret results in coordinates and/or in terms of the original components.

In the engineering field, a control scheme suitable for CoDa has been proposed in [5]. It is based on the Mahalanobis distance, T^2, and interpretation of the CoDa out of control signals is discussed for the three part case in [6] and for the general case in "unpublished" [7].

The data analyses discussed in this work have been conducted using our own R routines [8] and the open-source CoDaPack [9]. Other useful R packages to perform CoDa analysis are "compositions", "robCompositions" and "zCompositions". Computer routines implementing these methods, as well as other related compositional techniques, can be obtained from the website http://www.compositionaldata.com.

The rest of this paper is organized as follows. The difficulties of typical statistical methods when applied to raw CoDa are illustrated in the following section. Next, we introduce the basic properties and geometric settings of CoDa analysis. Afterwards, a typical MPE dataset is analysed using descriptive elements and techniques of log-ratio analysis. Last section concludes with some final remarks.

Usual Statistical Methods Applied to Raw CoDa

Following [10], we consider the world motor vehicle production where "*interest focuses on the proportions rather than the amounts*". The data analysed (Table 3 in the Appendix) is available at http://www.rita.dot.gov/ of the Bureau of Transportation Statistics (U.S. Department of Transportation). Hereafter we will call WMVP to this data set. To illustrate usual and log-ratio methods we consider the motor vehicle production from 1994 to 2013 (N = 20) distributed in D = 8 parts representing different countries or regions: China, United States, Japan, Germany, Other Asia, Other America, Other Europe and Rest. When the total sum of the vector equals k (e.g., 1) the data will be referred to as full-composition. A subset of the parts is called a sub-composition.

We follow by applying an example of standard statistical method to the WMVP datasets. Figure 2 shows the ternary diagram for the three part sub-compositional dataset involving [China, United

States, Japan] = [X, Y, Z]. It can be seen clearly that the large variability appears in the production from China (X). The label of the first year (1994) and the last year (2013) suggest the trend of this sub-composition: China increases its relative percentage, whereas United States and Japan decrease both in approximately the same proportion. Dashed line in Fig. 2 represents the projection from the sub-composition [X, Y, Z] in the simplex S^3 to the sub-composition [s_Y, s_Z] in the simplex S^2, which shows small variability. Formally the sub-composition [s_Y, s_Z] is obtained applying the closure operation $\mathbb{C}$ to the raw parts [Y, Z], i.e. [s_Y, s_Z]=$\mathbb{C}$([Y, Z]) = [Y/(Y + Z), Z/(Y + Z)]. Similarly to any projection in a Euclidean space, the sub-composition operation generally produces a loss of information.

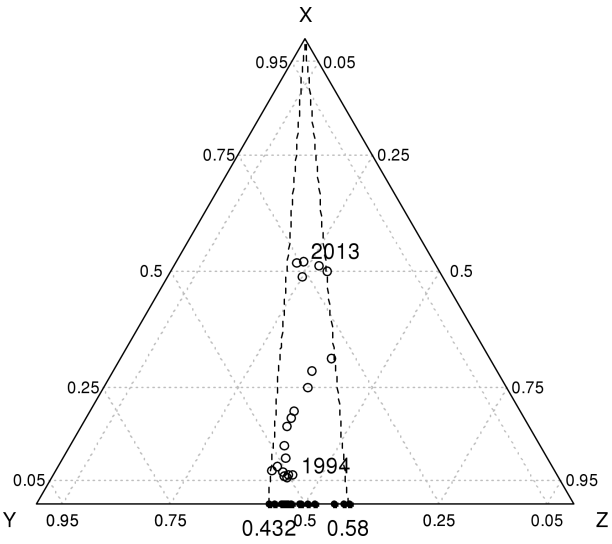

Fig. 2. Sub-composition [China, United States, Japan] = [X, Y, Z] (○) from the WMVP dataset in the ternary diagram. Labels 1994 and 2013 show the first and last year respectively. Vertical dashed lines are the projection to the sub-composition [s_Y, s_Z] = [United States, Japan] (●), into the YZ edge.

This effect could be crucial in statistical analysis, for example, cluster groups in the full-composition may collapse in a sub-composition.

Observing the typical scatterplot of the raw parts [Y, Z] (Fig. 3) we get to a misleading interpretation indicating large variability in these parts. Empty circles in the ternary diagram are the sub-composition [China, United States, Japan] = [X, Y, Z]. Vertical dotted lines represent its projection to the plane of the raw parts [Y, Z]. The sub-composition [s_Y, s_Z] is represented in the edge of parts Y and Z. Note that, due to the fact that X + Y + Z = 1, when the raw parts [Y, Z] are plotted, actually the information of the production of China (X) is also included in the relationship.

Note that proportions Y and Z are obtained from the closure of quantities W_Y and W_Z respectively, against a total T production of the three countries, i.e. $Y = W_Y/T$ and $Z = W_Z/T$. In our dataset, W_Y and W_Z are the number of motor vehicles produced by United States and Japan, respectively. Therefore, the part X could be considered as a residual part, X =1 – Y – Z, and $W_X = T - W_Y - W_Z$, where $T = W_X + W_Y + W_Z$. Note that the raw parts [Y, Z] = [W_Y/T, W_Z/T] include a spurious relationship between Y and Z, a misleading effect known from [11].

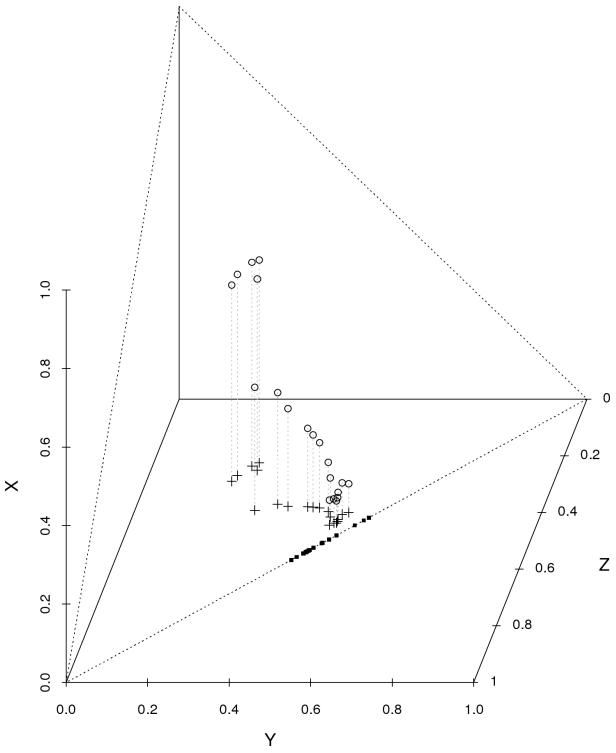

Fig. 3. Sub-composition [China, United States, Japan]= [X, Y, Z] (○) from the WMVP dataset in the ternary diagram. Vertical lines represent a typical projection on the plane of the raw parts [United States, Japan]= [Y, Z] (+). The sub-composition [s_Y, s_Z] (●) are represented in the YZ edge.

In other words, the typical correlation coefficient r_{YZ}= 0.8867 that measures the relationship between the production in United States and Japan includes the effect of the residual part X, the production in China.

On the other hand, the sub-composition $[s_Y, s_Z] = \left[\frac{Y}{Y+Z}, \frac{Z}{Y+Z}\right] = \left[\frac{W_Y/T}{W_Y/T+W_Z/T}, \frac{W_Z/T}{W_Y/T+W_Z/T}\right] = \left[\frac{W_Y}{W_Y+W_Z}, \frac{W_Z}{W_Y+W_Z}\right]$ is completely free of the spurious influence of the residual part X. However, again

because of the relationship induced by the closure, it makes no sense to calculate the typical correlation coefficient between both parts ($r_{s_X s_Y} = -1$; Fig. 3).

Standard variability statistics can be expressed in terms of Euclidean distances. The Euclidean distance is based on the subtraction of variables: Y – Z. The ratio $Y/Z = s_Y/s_Z = W_Y/W_Z$ provides the same information in the full and in any sub-composition unlike the differences Y – Z and s_Y–s_Z. Note that the closure $\mathbb{C}$ is not a necessary operation when analyzing a ratio and it suggests a way to analyze CoDa avoiding spurious correlation.

As we stated before, the total sum of a CoDa vector is irrelevant, thus the information carried in any D-composition $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ is the same as in $c\cdot\mathbf{x}=[c\cdot x_1, c\cdot x_2, \ldots, c\cdot x_D]$ $\forall c >0$. In particular, when $c = 1/x_D$, a composition $\mathbf{X}$ is completely determined by the vector of $D-1$ ratios $[x_1/x_D, x_2/x_D, \ldots, x_{(D-1)}/x_D, 1] \equiv [x_1/x_D, x_2/x_D, \ldots, x_{(D-1)}/x_D]$ [1]. In the example, the full-composition [X, Y, Z] is completely determined by the vector of ratios [Y/X, Z/X] and the ratio Y/Z determines the sub-composition [s_Y, s_Z]. Figure 4 shows the scatterplot of ratios [Y/X, Z/X], where the cloud is very similar to the cloud in the ternary diagram (Fig. 2). The boxplots of the ratios are also plotted in the margins; its slight skewness suggests taking logarithms to improve the symmetry of the data.

Basic properties and geometric settings of CoDa analysis

The analysis of CoDa introduced in [1] has two main properties: *scale invariance* and *sub-compositional coherence*. As we stated above, scale invariance means that vectors with proportional positive components represent the same composition. Therefore, the vector $\mathbb{C}(\mathbf{x}) = \left[\frac{x_1}{\sum x_j}, \frac{x_2}{\sum x_j} \ldots, \frac{x_D}{\sum x_j}\right]$ or the vector $[x_1/x_D, x_2/x_D, \ldots, x_{(D-1)}/x_D]$ can be selected as representatives of any composition $\mathbf{x}$. According to [1], it is stated that "all meaningful function of a composition can be expressed in terms of a set of component ratios". This property is crucial when defining a distance function or a probability density function on the simplex.

The property of sub-compositional coherence means that the interpretation and results provided by any analysis of a subset of parts does not depend on the rest of parts. Typical statistical analysis based on Euclidean distance or multivariate normal distribution does not fulfil these properties when applied to raw CoDa, e.g., [12, 13]. It is explained because typical statistical techniques are based on measuring absolute differences using a subtraction: $\mathbf{x}$–$\mathbf{y}$, while, on the other hand, working with differences based on ratios $\mathbf{x}/\mathbf{y}$ automatically involves scale invariance and sub-compositional coherence.

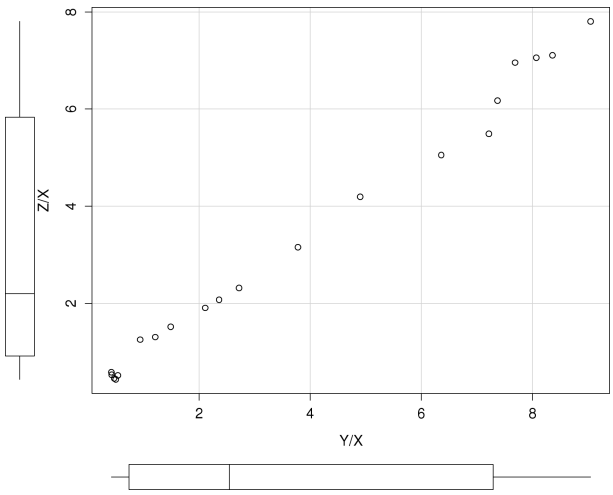

Fig. 4. Scatterplot of vector of ratios [Y/X, Z/X] for the sub-composition [X, Y, Z] = [China, United States, Japan] from the WMVP dataset.

A first geometric setting is that the natural movement from a composition $\mathbf{x}$ to another composition $\mathbf{y}$ should be based on a component wise product. Aitchison [1] introduced this operation called perturbation, and defined it as $\mathbf{x} \oplus \mathbf{y} = \mathbb{C}\ [x_1\cdot y_1, \ldots, x_D\cdot y_D]$; the perturbation difference is $\mathbf{x} \ominus \mathbf{y} = \mathbb{C}\ [x_1/y_1, \ldots, x_D/y_D]$, and the neutral element is the composition $\mathbf{n}$=[1,1,…,1]. The second operation that forms the Aitchison geometry for CoDa is the powering of a composition $\mathbf{x}$ by a real number α defined as $\alpha\odot\mathbf{x} = \mathbb{C}\ [x_1^\alpha, \ldots, x_D^\alpha]$. Perturbation and powering define a vector space structure of dimension D–1 [14]. The role played by both operations in most common statistical models and its usefulness in many applications from different fields are illustrated in [3].

In MPE the change of the production distribution is a problem that can be modelled by a perturbation process. For example, let [X_{1994}, Y_{1994}, Z_{1994}] = [5.60, 50.69, 43.71] and [X_{2013}, Y_{2013}, Z_{2013}] = [51.66, 25.85, 22.49] be, respectively, the first and last compositions in the WMVP dataset. In a simple exercise, consider the last compositions as the result of a continuous alteration of the initial composition. In this case the non normalised/closed perturbation difference vector is equal to [9.22, 0.51, 0.51], whose interpretation is that United States and Japan reduced 50% its relative importance and China increased 9 times its relative weight as regards to the production of the three countries.

By taking logarithm of a ratio (log-ratio), some features are improved. Once a ratio x/y between two positive values is calculated, the result is in the interval $[0, +\infty)$: the interval $(0, 1)$ corresponds to the "$x < y$" case and the interval $(1, +\infty)$ to the opposite "$x > y$". This asymmetry on the length of the domains recommends taking logarithms to extend the domain to the full Real space. In addition, dealing with log-ratios is more easy than with standard ratios because a "ratio" means a multiplicative way of thinking, while a "log-ratio" consists of the typical additive way of computation: $\ln(x/y) = \ln(x) - \ln(y)$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ be a D-composition. In general, a log-ratio is defined as

$$\ln\left(\prod_{\mathrm{j}=1}^{\mathrm{D}} \mathrm{x_j}^{\alpha_\mathrm{j}}\right) = \sum_{\mathrm{j}=1}^{\mathrm{D}} \alpha_\mathrm{j} \cdot \ln(\mathrm{x_j}), \tag{1}$$

where $\sum \alpha_\mathrm{j} = 0$ to verify the scale invariant property. The expression (1) is known as a log-contrast [1]. In the literature, the most famous log-contrast is the i-th centered log-ratio (clr_i) whose expression applied to a composition $\mathbf{x}$ is

$$\begin{aligned}\mathrm{clr_i}(\mathbf{x}) = &\frac{\ln(x_1)}{D} + \ldots + \frac{\ln(x_{\mathrm{i}-1})}{D} + \frac{(1-\mathrm{D})\cdot\ln(x_\mathrm{i})}{D} \\ &+ \frac{\ln(x_{\mathrm{i}+1})}{D} + \ldots + \frac{\ln(x_D)}{D} = \ln\left(\frac{x_\mathrm{i}}{\mathrm{g_m}(\mathbf{x})}\right),\end{aligned} \tag{2}$$

where $\mathrm{g_m}(\mathbf{x})$ is the geometric mean of $\mathbf{x}$. When $\mathrm{clr_i}$ log-contrast is applied to all parts, the vector of clr-coefficients [1] is obtained: $\mathrm{clr}(\mathbf{x}) = (\mathrm{clr}_1(\mathbf{x}), \ldots, \mathrm{clr}_D(\mathbf{x}))$. The expression $\mathbb{C}(\exp(\mathrm{clr}(\mathbf{x})))$ is its inverse transformation, which gives the unitary representative of $\mathbf{x}$.

The Aitchison distance, d_a, is defined as $\mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}))$, where $\mathrm{d}(\cdot,\cdot)$ denotes the Euclidean distance in R^D. The norm and inner product, consistent with the Aitchison distance, are respectively defined by $\|\mathbf{x}\|_\mathrm{a} = \mathrm{d}_a(\mathbf{x},\mathbf{n})$ and $< \mathbf{x}, \mathbf{y} >_\mathrm{a} = \sum_{\mathrm{j}=1}^{\mathrm{D}} \mathrm{clr_j}(\mathbf{x}) \cdot \mathrm{clr_j}(\mathbf{y})$, leading to a Euclidean space structure [14]. With these metric elements on hands, one can exploit the well-known properties of Euclidean spaces: orthonormal basis, orthogonal projections, angles, ellipses, etc. In other words, one can apply multivariate statistical techniques, like cluster analysis, principal component analysis, linear regression or discriminant analysis. A state-of-the-art of these techniques, including basic elements of simplicial linear algebra and geometry, differential calculus and statistical modelling, are presented in [3].

The statistical log-ratio analysis is based on the principle of working in coordinates [4]. The first step consists in representing a composition $\mathbf{x}$ in log-ratio type coordinates by the use of an orthonormal basis in the simplex. Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ be an orthonormal basis in S^D. The orthonormal coordinates of a composition $\mathbf{x}$ are obtained using the *isometric log-ratio* function $\mathrm{ilr}(\mathbf{x}) = [<\mathbf{x},\mathbf{e}_1>_a, \ldots, <\mathbf{x},\mathbf{e}_{D-1}>_a]$. These ilr coordinates are log-contrasts (1) and isometric, $\mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}(\mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y}))$. In practice, it is recommended to build the log-contrasts such that they have an easy interpretation on the problem studied.

A Sequential Binary Partition (SBP) [15] provides an orthonormal and its respective ilr coordinates such that enhances the interpretability of the representation of the parts of a composition. Authors in [16] provide a practical implementation and a representation in a dendrogram like structure. A SBP consists of D-1 steps, where an orthonormal coordinate, now called balance, is built in each step of the partition. In a first step, the composition x is split into two groups, which are indicated by $+1$ and -1. In consecutive steps, each previously created group of parts is split again into two groups. The partition ends when the groups are made up of a unique part. The coordinates created by a SBP are called balances because they have a very peculiar log-contrast expression. In the j-th step of a SBP, let x+ be the group of r parts marked with a $+1$ that are in the numerator and by x− the group of s parts in the denominator, marked with a -1. The corresponding balance, $\mathrm{b_j}$, is

$$\mathrm{b_j} = \sqrt{\frac{\mathrm{r}\cdot\mathrm{s}}{\mathrm{r}+\mathrm{s}}} \ln\left(\frac{\mathrm{g_m}(\mathbf{x}+)}{\mathrm{g_m}(\mathbf{x}-)}\right), \tag{3}$$

where $\mathrm{g}_m(\cdot)$ is the geometrical mean of involved parts of $\mathbf{x}$. The balances are log-ratios of geometric means of groups of parts and they have an easy interpretation. Table 2 shows the SBP used in the analysis of the full composition in the WMVP dataset. Remember that the dataset provides the number of vehicle production units in a composition of 8 countries and regions for 20 years, from 1994 to 2013. The first step in the SBP consists in separating the production in China ($+1$) from the rest (-1). In other words, the balance b_1 is

$$\mathrm{b}_1 = \sqrt{\frac{7}{8}} \ln\left(\frac{\mathrm{China}}{\sqrt[7]{a^*}}\right), \tag{4}$$

where

$$a^* = \mathrm{United\ States} \cdot \mathrm{Japan} \cdot \mathrm{Germany} \cdot \mathrm{Other\ Asia} \cdot \mathrm{Other\ America} \cdot \mathrm{Other\ Europa} \cdot \mathrm{Rest}.$$

This balance informs about the relationship between the production in China and the production

in the rest of countries and regions. In the next step, the parts that are in the denominator of balance b_1 (4), are split into the groups [United States] and [Japan, Germany, Other Asia, Other America, Other Europe, Rest] to define balance b_2

$$b_2 = \sqrt{\frac{6}{7}} \ln\left(\frac{\text{UnitedStates}}{\sqrt[6]{b^*}}\right),$$

where

$$b^* = \text{Japan} \cdot \text{Germany} \cdot \text{Other Asia} \cdot \text{Other America} \cdot \text{Other Europa} \cdot \text{Rest}.$$

The analyst is completely free to decide how to split the group variables and define the SBP based on his or her expertise and on a previous exploratory analysis of the dataset.

Basic exploratory log-ratio analysis

The basic elements of an exploratory analysis are the mean and the variability. The variability of a random composition $\mathbf{X}$ with respect to a composition $\mathbf{x}$ is $\text{Var}(\mathbf{X},\mathbf{x}) = \text{E}(\text{d}_a^2(\mathbf{X},\mathbf{x}))$, where Var and E are the typical variance and expectation in Real space [17].

Furthermore, the expectation or centre of $\mathbf{X}$ is $\text{Cen}(\mathbf{X}) = \min_{x \in S^D} \text{d}_\text{a}^2(\mathbf{X}, \mathbf{x})$, and the total variance of $\mathbf{X}$ is $\text{totVar}(\mathbf{X}) = \text{E}(\text{d}_a^2(\mathbf{X}, \text{Cen}(\mathbf{X})))$. Centre and total variance are calculated using the expressions

$$\begin{aligned}\text{Cen}(\mathbf{X}) &= \text{ilr}^{-1}(\text{E}(\text{ilr}(\mathbf{X}))) \\ &= \mathbb{C}[\exp(\text{E}(\ln X_1)), \ldots, \exp(\text{E}(\ln X_D))], \\ \text{totVar}(\mathbf{X}) &= \frac{1}{D}\sum_{\text{i}=1}^{D-1}\sum_{\text{j}=\text{i}+1}^{D} \text{Var}\left(\ln\frac{X_\text{i}}{X_\text{j}}\right) \\ &= \sum_{\text{j}=1}^{D-1} \text{Var}(\text{ilr}(X_\text{j})) = \sum_{\text{j}=1}^{D} \text{Var}(\text{clr}(X_\text{j})),\end{aligned}$$

where an estimate of $\exp(\text{E}(\ln(\cdot))$ is the geometric mean of a part $\text{g}(\cdot)$. Therefore, $\text{Cen}(\mathbf{X})$ is the unitary representative of the geometric mean of $\mathbf{X}$.

Aitchison in [1] introduced the variation array, a very informative way to present log-ratio expectations and variances. Table 1 shows the variation array for the WMVP dataset. The lower triangle of the array contains the values of the sample means of the log-ratios of the corresponding two parts (numerator by row, denominator by column). For example the value 0.83 corresponds to E(ln(UnitedStates/China)). The upper triangle of the array contains the sample variances of the same log-ratios. This array can be easily extended adding a right column collecting the values of the variances of the clr-coefficients, and a bottom row for the $\text{Cen}(\mathbf{X})$. The sum of the right column equals totVar and is 0.9261. Note that Other America and Other Asia are the regions showing the smallest relative variability, while China and United States have large variability. When the log-ratios are analysed, the Var(ln(Japan/OtherEurope))= 0.01 is the smallest, suggesting association between them and meaning that the production in these regions is approximately proportional in the dataset. On the other hand, the productions in China and United States have the largest log-ratio variance, suggesting slight or no association between them. Observe that the centre in the bottom row suggests that the production in the Rest regions is very small, whereas the production in Other Europe shows the largest percentages. We have seen that productions in Japan and Other Europe are approximately proportional and from Table 1 we see that E(ln(OtherEurope/Japan)) = 0.2879, thus one can assume that, in average, it holds Other Europe $\approx e^{0.2879}$·Japan. Observe that the sign of $\text{E}(\ln(\cdot/\cdot))$ indicates which element show higher concentrations and a value close to zero suggests that both elements are similarly present in the artefacts. For example, E(ln(Japan/UnitedStates))= –0.06 indicates that, in average, the production in United States and Japan is very similar, perhaps is slightly more in United States. Moreover, the small value of the corresponding log-ratio variance Var(ln(Japan/UnitedStates))= 0.03 suggests that this similarity between productions holds in most of years.

The relationships between pairs of countries and regions are provided by the variation array. The biplot of the clr-coordinates is an appropriate graphical tool to analyse more complex associations. The *clr-biplot* represents a bidimensional projection of the clr-log-ratio coordinates of samples in the same plot as the projection of the centred clr variables. The coordinates of samples and variables in the plot are calculated using elements provided by a Singular Value Decomposition (SVD) of the clr-coefficients data matrix. Due to the fact that the order of the vectors in the basis provided by SVD is decreasing according to the singular value, by taking the first two vectors, the proportion of variance retained by the clr-biplot can be calculated. SVD has another useful property when looking at the location of samples: the Euclidean distance between two samples in the clr-biplot is an approximation of the Aitchison distance between the corresponding compositions in the simplex.

Table 1
Variation array of the CoDaWMVPdataset. Lower triangle: log-ratio means; Upper triangle: log-ratio variances. The bottom row contains the centre of the dataset; right column the clr-variances.

	China	United States	Japan	Germany	Other Asia	Other America	Other Europe	Rest	clrVar
China		1.26	1.03	0.84	0.48	0.66	0.93	0.45	0.59
United States	0.83		0.03	0.06	0.21	0.11	0.04	0.30	0.13
Japan	0.76	−0.06		0.02	0.12	0.05	0.01	0.18	0.06
Germany	0.15	−0.68	−0.62		0.07	0.02	0.01	0.13	0.03
Other Asia	0.11	−0.72	−0.65	−0.04		0.02	0.09	0.06	0.01
Other America	0.40	−0.43	−0.37	0.25	0.29		0.03	0.08	0.005
Other Europe	1.05	0.23	0.29	0.90	0.94	0.66		0.16	0.04
Rest	−2.24	−3.06	−3.00	−2.39	−2.35	−2.64	−3.29		0.05
Cen(**X**) (%)	8.21	18.77	17.65	9.54	9.18	12.24	25.53	0.88	

To illustrate these properties and some more, the clr-biplot of the WMVP dataset is shown in Fig. 5. The first two axis of the clr-biplot retain 93.79% of variability. The squared length of a ray associated to a part is proportional to the clr-variance of the corresponding part, thus we can check that the production in China is the longest and the Other America the shortest, in agreement with the values collected in Table 1. The variance of the log-ratio of two parts (Table 1) is approximately equal to the squared length of the link between the two corresponding vertices of rays. In consequence, the closer are two vertices in the clr-biplot, the higher proportionality have the production of the countries and regions. Observe (Table 1; Fig. 5) that the closest vertices are those of United States, Japan, Germany, Other America and Other Europe. Moreover, the cosinus of the angle between two links approaches the correlation coefficient between the corresponding simple log-ratios. Therefore, orthogonality of links in the clr-biplot suggests uncorrelation of the corresponding log-ratios. For example, the link between the vertices of the rays clr(China) and clr(Japan) is approximately orthogonal to the link of clr(OtherAsia) and clr(Rest). When the linear correlation coefficient between ln(China/Japan) and ln(OtherAsia/Rest) is checked the value 0.18 is obtained.

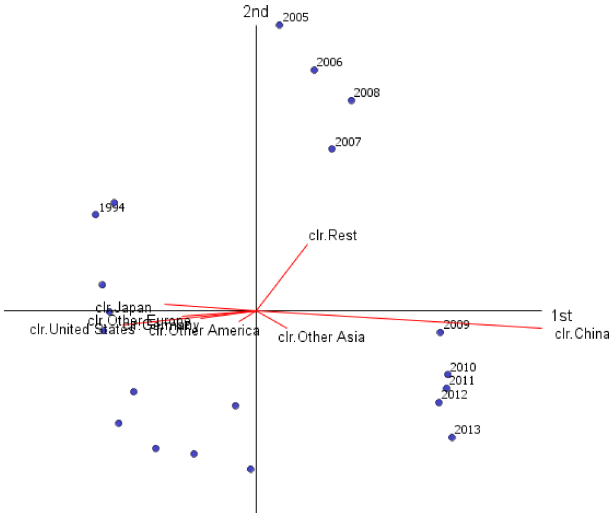

Fig. 5. Clr-biplot (1st and 2nd axis) of the WMVP dataset. Red lines are the rays of the clr-variables. The proportion of variability retained is 93.79% [colour online].

Table 2
SBP of CoDa set WMVP, represented in Fig. 6 as a CoDa-dendrogram.

Balance	China	United States	Japan	Germany	Other Asia	Other America	Other Europe	Rest
b_1	+1	−1	−1	−1	−1	−1	−1	−1
b_2	0	+1	−1	−1	−1	−1	−1	−1
b_3	0	0	+1	−1	−1	−1	−1	−1
b_4	0	0	0	+1	−1	−1	−1	−1
b_5	0	0	0	0	+1	−1	−1	−1
b_6	0	0	0	0	0	+1	−1	−1
b_7	0	0	0	0	0	0	+1	−1

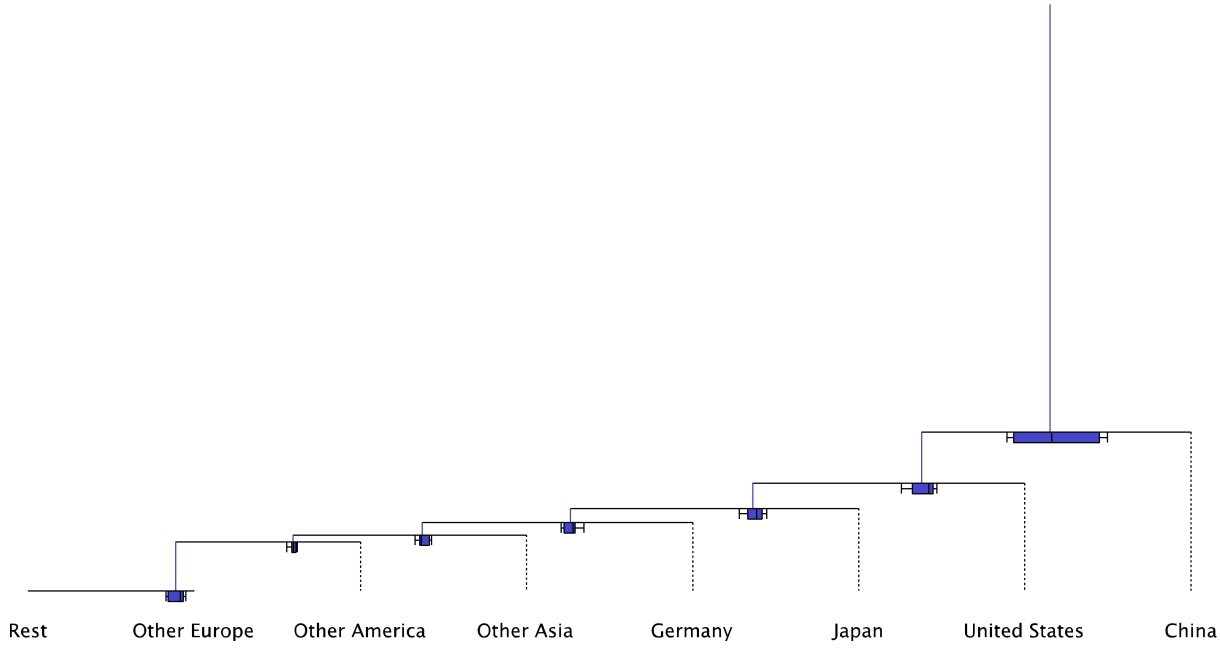

Fig. 6. CoDa-dendrogram of the WMVP dataset.

The location of samples in the clr-biplot is interpretable. In our dataset we can check that the samples corresponding to years from 2005 to 2008 are near to the ray clr(Rest), which is explained by the fact that, during those years, this region relative increased its production (Table 3). Analogously, the first years of the sample are in the negative part of the first axis and the last years in the opposite side, which is consistent with the fact that the ray associated to the Chinese production dominates this axis.

The first axis, which retains 93.23% of variability, can be expressed as the log-contrast

$$\ln\left(\frac{\text{China}^{8.3}\cdot\text{OtherAsia}^{0.9}\cdot\text{Rest}^{1.5}}{c^*}\right)$$

where

$$c^* = \text{United States}^{3.8}\cdot\text{Japan}^{2.6}\cdot\text{Germany}^{1.6}\cdot\text{Other America}^{0.5}\cdot\text{Other Europe}^{2.2}$$

which has a difficult interpretation in terms of the original parts. On the other hand, the particular SBP created in Table 2 improves the interpretation of the log-ratios. To summarize the structure of a SBP a useful tool is to represent the CoDa-dendrogram which in addition represents the ilr decomposition of the total variance and the mean and dispersion of each balance. The CoDa-dendrogram of the WMVP dataset following the SBP shown in Table 2 is represented in Fig. 6. The lengths of vertical bars, which connect two groups of parts, are proportional to the variance of the balance. Observe that, in this case, the first balance has the largest variance and is interpreted as the log-ratio of the production in China against the geometric mean of the production in the other countries and regions (4).

The point where each vertical bar joins a horizontal bar indicates the mean balance (coordinate of the sample centre). In the first balance, the joining point of the vertical bar is close to the middle of the horizontal bar which is consistent with the mean of this balance that equals –0.14. On the other hand, the joining point of the last balance

$$\text{b}_7 = \ln\left(\frac{\text{Other Europe}}{\text{Rest}}\right)$$

is on the right hand and its mean equals 2.33. Moreover, on each horizontal bar a box-plot of quantiles (0.05, 0.25, 0.50, 0.75, 0.95) of the corresponding balance is represented to visualize the ilr dispersion. For instance, the shape of the box-plot of the first balance suggests symmetry and large dispersion.

Final remarks

Frequently MPE have to face analysis of CoDa. Particular characteristics of CoDa require a coherent statistical analysis in order to avoid misleading results and conclusions. The analysis of log-ratios is the basis of a methodology free of spurious correlation. In this sense, standard statistical methods can be applied to compositions (e.g., percentages) expressed in terms of log-ratio orthonormal coordinates. The interpretations of log-ratio coordinates are easier using an appropriate SBP based on the expertise of the analyst and on a previous exploratory analysis.

In some cases, the percentages in parts are very small and are rounded to zero. In these situations, it is necessary to use imputation strategies for the "zero" values in order to be able to compute log-ratio coordinates. In literature, this topic is also known as rounded zero problem. The imputation strategies are based on completing the data matrix by replacing rounded zeros by reasonable estimates, allowing the computation of any log-ratio. Authors in [18] provide recent advances in this topic.

Log-ratio analysis and correspondence analysis have some similarities where dimensionality reduction of a table of positive data is concerned. A comparison of both methods is discussed in [19].

This research has been supported by the Spanish Ministry of Economy and Competitiveness under the project "METRICS" Ref. MTM2012–33236, and the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

Appendix

Table 3
World motor vehicle production (WMVP) available at http://www.rita.dot.gov/ of the Bureau of Transportation Statistics (U.S. Department of Transportation)

Year	China	United States	Japan	Germany	Other Asia	Other America	Other Europe	Rest
1994	1353000	12239288	10554000	4356138	3346564	5434920	12020772	354000
1995	1435000	11995248	10196000	4667000	3732000	5258000	12432000	331000
1996	1466000	11830157	10346000	4843000	4117000	5735000	12837000	322000
1997	1578000	12130575	10975000	5023000	4225000	6442000	12751000	349000
1998	1628000	12002663	10050000	5727000	3006000	5657000	12025000	384000
1999	1805000	13024978	9905000	5688000	4167000	6240000	13192000	311000
2000	2009000	12773714	10145000	5198000	4571000	6896000	14641000	348000
2001	2331776	11424689	9777191	5691677	4401794	6423526	14292405	319375
2002	3251225	12279582	10257690	5144714	4376229	6386168	14084912	344063
2003	4443686	12087028	10286318	5506629	4725081	6124969	14198091	413261
2004	5070527	11960354	10511518	5569954	5411449	6734992	14568009	405314
2005	5668163	11946653	10799659	5757710	5787752	7375146	14715815	905453
2006	7566233	11260277	11484233	5819614	6101086	7831734	15081328	915455
2007	8885461	10752310	11596327	6213460	6619507	8361918	15834657	866729
2008	9233290	8672141	11563629	6045730	6325283	8202788	14732133	887083
2009	13648553	5709431	7934516	5209857	6381784	6861255	11041611	597277
2010	18264667	7743093	9625940	5905985	7825544	8887209	12989554	711492
2011	18418876	8655003	8398654	6311318	8597454	8891853	13972162	751921
2012	19271808	10332626	9942711	5797471	8706707	9704425	13322556	760648
2013	22116825	11066432	9630070	5877332	8417617	10061124	13228742	756451

References

[1] Aitchison J., *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by The Blackburn Press), London (UK), 1986.

[2] Hron K., Filzmoser P., Templ M. [Eds.], Proceedings of the 5th International Workshop on Compositional Data Analysis, Codawork'13 June 3–7, Vorau, Austria, 2013, ISBN: 978-3-200-03103-6.

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

[4] Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J.J., *The principle of working on coordinates and compositional data analysis*, [in:] Compositional Data Analysis: Theory and Applications, Pawlowsky-Glahn V., Buccianti A. [Eds.], Wiley & Sons (UK), 2011.

[5] Vives-Mestres M., Daunis-i-Estadella J., Martín-Fernández J.A., *Individual T2 control chart for Compositional Data*, Journal of Quality Technology, 46 (2), 127–139, April 2014.

[6] Vives-Mestres M., Daunis-i-Estadella J., Martín-Fernández J.A., *Out-of-Control signals in 3-part Compositional T2 control chart*, Quality and Reliability Engineering International, 30 (3), 337–346, April 2014.

[7] Vives-Mestres M., Daunis-i-Estadella J., Martín-Fernández J.A., *Signal interpretation in Hotelling's T2 control chart for Compositional Data*, unpublished.

[8] R development core team 2013. R: A language and environment for statistical computing: Vienna, http://www.r-project.org.

[9] Comas-Cufí M., S. Thió-Henestrosa, CoDaPack 2.0: a stand-alone, multi-platform compositional software, [in:] Egozcue J.J., Tolosana-Delgado R., Ortego M.I. [Eds.], CoDaWork'11: 4th International Workshop on Compositional Data Analysis, Sant Feliu de Guíxols, 2011.

[10] Grunwald G.K., Raftery A.E., Guttorp P., *Time Series of Continuous Proportions. Journal of the Royal Statistical Society*, Series B (Methodological), 55 (1), 103–116, 1993.

[11] Pearson K., *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 LX, 489–502, 1897.

[12] Chacón J.E., Mateu-Figueras G., Martín-Fernández J.A., *Gaussian kernels for density estimation with compositional data*, Computer&Geosciences, 37, 702–711, 2011.

[13] Palarea-Albaladejo J., Martín-Fernández J.A., Soto J.A., *Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data*, Journal of Classification, 29 (2), 144–169, 2012.

[14] Pawlowsky-Glahn V., Egozcue J.J., *Geometric approach to statistical analysis on the simplex*, Stochastic Environmental Research and Risk Assessment (SERRA), 15 (5), 384–398, 2001.

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

[16] Thió-Henestrosa S., Egozcue J.J., Pawlowsky-Glahn V., Kovács L.O., G. Kovács, *Balance-dendrogram a new routine of CoDaPack*, Computer and Geosciences, 34 (12), 1682–1696, 2008.

[17] Pawlowsky-Glahn V., Egozcue J.J., *Exploring Compositional Data with the Coda-Dendrogram*, Austrian Journal of Statistics, (1–2), 103–113, 2011.

[18] Palarea-Albaladejo J., Martín-Fernández J.A., *zCompositions – R package for multivariate imputation of nondetects and zeros in compositional data sets*, Chemometrics and Intelligent Laboratory Systems, 143, 85–96, 2015.

[19] Greenacre M., *Compositional data and correspondence analysis*, [in:] Compositional Data Analysis: Theory and Applications, Pawlowsky-Glahn V., Buccianti A. [Eds.], Wiley & Sons (UK), 2011.