Compositional data analysis in e-tourism research. What TripAdvisor reviews complain about hotels in Barcelona

Berta Ferrer-Rosell[1], University of Lleida, Lleida (Spain), berta.ferrer@aegern.udl.cat
Germà Coenders, University of Girona, Girona (Spain), germa.coenders@udg.edu
Eva Martin-Fuentes, University of Lleida, Lleida (Spain), eva.martin@udl.cat

Abstract

Compositional Data (CoDa) contain information about the relative importance of parts of a whole, which the researcher deems more interesting than overall size or volume. In web mining, for instance, the relative frequency of a term is normally given more importance than absolute frequency, which mostly tells about web size, in other words, the sheer volume of online content. Many research questions in e-tourism are either related to the distribution of a whole or relative importance: How do the most salient contents in hotel Facebook accounts relate to hotel characteristics? What are the dominant topics on TripAdvisor comments about fish freshness in seafood restaurants? How does the relative popularity of search terms in Google relate to destination market share?

In CoDa, most of the basic statistical notions, such as center, variation, association and distance are flawed unless they are re-expressed by means of logarithms of ratios. The appeal of log-ratios is that once they are computed, standard statistical methods can be used. On the other hand, since one part can only increase in relative terms if some other(s) decrease, statistics need to be multivariate.

This chapter uses an example based on TripAdvisor hotel reviews from one of the most visited cities worldwide, Barcelona, focusing on what users complain about, to illustrate the main multivariate exploratory and descriptive tools in CoDa, including imputation of zeros prior to computing the log-ratios, multivariate outlier detection, principal component analysis, cluster analysis, and multivariate data visualization tools. The use of CoDaPack, a popular CoDa freeware, is described in a step-by-step fashion.

Keywords: Compositional Data; CoDa; Content Analysis; TripAdvisor reviews; Cluster Analysis; biplot.

[1] Corresponding author

To appear in:

Zheng X, Fuchs M, Gretzel U, Höpken W (eds) Handbook of E-Tourism. Springer Nature. ISBN: 978-3-030-05324-6

1. Introduction

1.1. Compositional research questions or compositional data?

Compositional Data analysis (CoDa) is the standard statistical methodology used when data contain information about the relative importance of parts of a whole, typically with a fixed sum. The CoDa tradition started with Aitchison's seminal work (1982, 1986) on chemical and geological compositions, where only the proportion of each part or component is of interest, since absolute amounts are irrelevant and only inform about the size of the chemical or soil sample (Buccianti et al. 2006). In the last three decades, CoDa has provided a standardized toolbox for statistical analyses whose research questions concern the relative importance of magnitudes. The term compositional analysis (Barceló-Vidal and Martín-Fernández 2016) has even been coined to stress the fact that what is ultimately compositional is not the data, which may not be parts of a whole or may fail to have a fixed sum, but the research objectives or hypotheses focusing on relative rather than absolute values. Example applications of CoDa to data which do not represent parts of any whole can be found in Ortells et al. (2016) and Linares-Mustarós et al. (2018).

Many research questions in e-tourism and related fields are either related to distribution of a whole (e.g., distribution, share, allocation, etc.), or relative importance (e.g., dominance, concentration, profile, etc.), which are deemed more relevant to the research objective than absolute data. Example research questions might be: How do the dominant contents in hotel Facebook accounts relate to hotel characteristics? (Ferrer-Rosell et al. 2019). Do large and small hotels use Weibo in the same way regarding the relative weight of posts about events, facilities, promotions and menu? (Zhou et al. 2017). How does the share of TripAdvisor rating categories of a hotel relate to the distribution of reviewers by market segment? (Ferrer-Rosell et al. 2021). Which keywords concentrate the profile difference between a destination's projected image and user-generated content? (Marine-Roig and Ferrer-Rosell 2018). How does the relative popularity of search terms in Google relate to market performance? (Ortells et al. 2016). How do salient issues within the twitter communication agenda evolve along time? (Blasco-Duatis et al. 2019). Which types of content posted on social media webpages generate more fan engagement? (Kwok and Yu 2013; Russell 2014; Yoo and Lee 2017). Which topics are discussed more frequently in negative reviews than positive ones? (Hu et al 2019). In the example presented in this chapter, the research questions are similar to the latter: how the dominant topics of customer complaints about hotels in TripAdvisor are related to one another and how distinct hotel clusters can be drawn based on major complaint topics.

Accessible handbooks have contributed to extending the use of CoDa to many scientific fields (van den Boogaart and Tolosana-Delgado 2013; Filzmoser et al. 2018; Greenacre 2018; Pawlowsky-Glahn and Buccianti 2011; Pawlowsky-Glahn et al. 2015), as has dedicated user-friendly software (van den Boogaart and Tolosana-Delgado 2013; Filzmoser et al. 2018; Greenacre 2018; Palarea-Albaladejo and Martín-Fernández 2015; Thió-Henestrosa and Martín-Fernández 2005), although in many cases standard software can be used after transforming the data. Nowadays, CoDa is employed in almost all of the hard sciences and has started to be used in several social science fields and application domains. These include but are not limited to psychology (Batista-Foguet et al. 2015; van Eijnatten et al. 2015), economics (Fry 2011; Hruzová et al. 2017), accounting (Carreras-Simó and Coenders 2020; Linares-Mustarós et al. 2018), marketing (Joueid and Coenders 2018; Morais et al. 2018; Vives Mestres et al. 2016), sociology (Kogovšek et al. 2013; Di Palma and Gallo 2019) political science (Blasco-Duatis et al. 2018), geography (Godichon-Baggioni et al. 2019; Sanz-Sanz et al. 2018), e-communication (Blasco-Duatis et al. 2019; Blasco-Duatis and Coenders, 2020; Ortells et al. 2016), and tourism (Coenders and Ferrer-Rosell 2020; Coenders et al. 2017; Ferrer-Rosell and Coenders 2017; 2018; Ferrer-Rosell et al. 2015; 2016a; 2016b; Song et al. 2019; Voltes-Dorta et al. 2014). The first applications to e-tourism have just started to appear (Ferrer-Rosell et al. 2019; 2020; 2021; Ferrer-Rosell and Marine-Roig 2020; Marine-Roig and Ferrer-Rosell 2018; Zhou et al. 2017).

When relative information is at hand, most of the basic statistical notions, such as center, variation, association and distance are flawed unless they are re-expressed by means of logarithms of ratios in the so-called CoDa methodology. The appeal of log-ratios is that once they are computed, standard statistical methods can be used as long as the relative character of the information is taken into account when interpreting the results. On the other hand, since one part can only increase in relative terms if some other(s) decrease, statistics needs to be multivariate. After dealing with log-ratios, center, variation, association and distance, this chapter presents the main multivariate exploratory and descriptive tools in CoDa, including imputation of zeros prior to computing the log-ratios, multivariate outlier detection, principal component analysis, biplots and cluster analysis. At a later stage, clusters and principal components can be related to external non-compositional variables in the usual manner. An application example based on customer complaints about hotels in TripAdvisor follows. For this purpose, CoDaPack, a popular menu-driven CoDa freeware, is used in a step-by-step fashion.

1.2. Composition definition

The *composition* **x** is a vector in the positive *D*-dimensional real space carrying information about the relative importance of its parts:

$$\begin{gathered}\mathbf{x}=\left(x_1, x_2, ..., x_D\right)\in R_{+}^{D}\ ,\\ \text{with}\quad x_j>0\quad \text{for all } j=1,2,...,D,\end{gathered}\qquad\text{(Eq. 1)}$$

where *D* is the number of *parts* or *components*. Individual compositions could be, for instance, hotels, and parts could be the possible reasons for complaining about them in TripAdvisor, or the content of photos posted by them in their Facebook accounts. In order to focus on the relative importance of the parts, the *closure* of **x** to a constant sum is common practice. It can also be the case that the raw data already have a fixed sum (e.g., 100% in market share data). Without loss of generality we consider the unit sum, so that after closure, **z** contains part proportions.

$$\begin{gathered}\mathbf{z}=C(\mathbf{x})=\left(\frac{x_1}{S},\frac{x_2}{S},...,\frac{x_D}{S}\right)=\left(z_1, z_2, ..., z_D\right)\\ \text{with } z_j>0\quad \text{for all } j=1,2,...,D;\quad \sum\nolimits_{j=1}^{D}x_j=S;\quad \sum\nolimits_{j=1}^{D}z_j=1.\end{gathered}\qquad\text{(Eq. 2)}$$

Regardless of whether closure is performed or not, the relative information contained by the *D* parts should remain the same, thus ensuring the so-called *compositional* *equivalence* property (Barceló-Vidal and Martín-Fernández 2016). This implies that results of a compositional analysis should be invariant to changes of scale in the data (*scale invariance principle*).

1.3. Fine-tuning the research questions

It is up to the researcher to select which *D* parts to analyze. In a *content analysis* of photos posted by hotels in their Facebook accounts one could, for instance, think of x_1=outside facilities (garden, terrace, swimming pool,...), x_2=inside facilities (gym, sauna,...), x_3=rooms, x_4=common inside spaces, x_5=menu, x_6=events, x_7=natural surroundings, x_8=urban surroundings. If the distinction between urban and natural surroundings is not of interest to the researcher (after all, a hotel in an urban environment has no other choice than to picture urban surroundings, and a hotel in a natural environment natural surroundings), both categories can be merged into one part termed "x_7+x_8=surroundings as a whole". This operation is called *amalgamation*. Due to the particularities of CoDa, amalgamated parts cannot be analyzed separately at a later stage (e.g. van den Boogart and Tolosana-Delgado 2013). In other words, amalgamated parts remain so forever, and amalgamation should take place in the problem definition stage. Following up with the same example, one could decide to study only the subset of parts x_1 to x_6 having to do with the hotel itself. This is referred to as a *subcomposition* in CoDa. In this particular example, the subcomposition would imply that the researcher is uninterested in content about surroundings. The amalgamation x_7+x_8 would imply that the researcher is interested in comparing the relative importance of content about surroundings as a whole with content about the hotel itself. Analyzing all parts x_1 to x_8 would imply that the researcher is additionally interested in comparing the relative importance of contents about urban and natural surroundings.

It is often claimed that all compositions are, in fact, subcompositions and amalgamations. After all, gym and sauna could have been treated as separate parts instead of amalgamating them within inside facilities. Additional contents could also have been added in order to have a more general composition of which the current one is only a subcomposition. What if, for instance, the researcher would have been interested in pictures about events organized in the hotel, or about guest celebrities?

1.4. Why are classical statistical techniques inappropriate?

The closed composition **z** resides in a subspace called the *simplex*, which is constrained by positiveness and unit sum, with different operations, angles and distances from the full real space. This explains why most statistical workhorses, such as mean, variance, correlation and distance, are to a greater or lesser extent meaningless when applied to **z**. Since one part can only increase if one or more of the others decrease(s), negative spurious correlations among the parts emerge (Pearson 1897). *Euclidean distances* among the individual compositions are also meaningless (Aitchison et al. 2000). Euclidean distance considers the pair of proportions 0.01 and 0.02 to be as mutually distant as 0.21 and 0.22, while in the first pair the difference is twofold and in the second it is less than 5% (Coenders and Ferrer-Rosell 2020).

In addition, statistical modeling with unbounded distributions such as the normal distribution is not feasible, as it results in values larger than 1 or lower than 0 having a positive probability of occurrence. The statistical and distributional assumptions of most classical statistical models are to a greater or lesser extent violated in **z** (Aitchison 2001; Pawlowsky-Glahn et al. 2015). Uncritical use of standard statistical models on raw untransformed compositional data is thus generally inappropriate.

Finally, the fact that one part can only increase in relative terms if some other(s) decrease(s), makes interpretation of the results dependent on which parts are made to decrease. Interpretation around one single part are thus bound to be misleading, which means that CoDa necessarily uses multivariate statistical methods.

2. Compositional data analysis in practice

2.1. Log-ratio transformations

In order to solve the aforementioned drawbacks, the most common CoDa approach is to express an original compositional vector of *D* parts in logarithms of ratios among parts (Aitchison 1986; Egozcue et al. 2003). *Log-ratios* are unbounded and thus have a chance to meet the distributional assumptions of classical statistical models (Pawlowsky-Glahn et al. 2015). In addition, they constitute a natural way of distilling the information about the relative size of parts and form the basis for defining association, variance and distance in a meaningful way. Finally, it must be noted that they yield the same result regardless of whether they are computed from **x** or **z** thus adhering to the scale invariance principle. In some instances, compositional data are even defined as those data for which the relevant information is carried by ratios (Egozcue and Pawlowsky-Glahn 2019).

Log-ratios may, for instance, be computed among all possible pairs of parts in the so-called *pairwise log-ratios*:

$$\ln\left(\frac{z_j}{z_k}\right) = \ln\left(\frac{x_j}{x_k}\right) \qquad \text{(Eq. 3)}$$
$$\text{with } j < k;\ k = 2,3,...,D;\ j = 1,2,...,k-1,$$

or between each part and the geometric mean of all parts including itself, in the so-called *centered log-ratios*. From now on we present only the formulation using the closed *z* parts:

$$\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 ... z_D}}\right) \qquad \text{(Eq. 4)}$$
$$\text{with } j = 1,2,...,D.$$

There are alternative interpretations and expressions of centered log-ratios (Filzmoser et al. 2018; Pawlowsky-Glahn et al. 2015). They can also be understood as the balance between one part and the geometric mean of the rest. The corresponding expression, which is equivalent to (Eq. 4) is:

$$\frac{D-1}{D}\ln\left(\frac{z_j}{\sqrt[D-1]{z_1 z_2 \ldots z_{j-1} z_{j+1} \ldots z_D}}\right)$$
$$\text{with } j = 1, 2, \ldots, D. \qquad \text{(Eq. 5)}$$

Equation (Eq. 5) stresses the fact that the value and the interpretation of centered log-ratios are subject to the definition of the research problem and the research questions. They will change when adding parts or when defining an amalgamation or a subcomposition. It also stresses the fact that both problem definition and data analysis must be mutually coherent and multivariate. Once the compositional research problem has been defined precisely, the greater the centered log-ratio, the greater the importance of the part, compared to the geometric mean of the rest of the parts included in the research problem.

One attractive feature of CoDa is that once the raw composition has been transformed into centered log-ratios, classical statistical techniques for unbounded data can be applied in the usual way, and even with standard software. Log-ratio transformations thus constitute the easy way out in compositional problems. The applied researcher can concentrate his or her effort in interpreting the results taking the compositional nature of the data and the research questions into account: what does increase at the expense of decreasing what?

To this end it must be taken into account that the *D* centered log-ratios have zero sum for any individual. This is a reflection of the sheer fact that, in relative terms, one part can only increase if some others decrease. Statistically speaking, the covariance matrix among the *D* centered log-ratios is singular and non-invertible. Among the methods described in this chapter, singularity only affects outlier detection, but the researcher must have in mind the fact that centered log-ratios cannot be applied for statistical techniques which require inverting the covariance matrix without taking extra precautions. Alternative transformations which lead to invertible covariances and can be readily used in more advanced statistical methods are described in van den Boogaart and Tolosana-Delgado (2013), Egozcue et al. (2003), and Pawlowsky-Glahn et al. (2015).

2.2. Basic statistical concepts

2.2.1. Center

In order to assess the overall relative importance of each part for all individual compositions, the composition center can be described from the arithmetic means of the centered log- ratios. For ease of interpretation, the researchers may wish to exponentiate these means (which then become geometric means), and close them to the original unit sum of the composition, in order to express them in the original scale.

2.2.2. Association

Proportionality between pairs of parts is a valid alternative to correlation (Lovell et al. 2015). The same pairwise log-ratios (Eq. 3) and their variances are computed as:

$$Var\left(\ln\left(\frac{z_j}{z_k}\right)\right)$$
$$\text{with } j<k;\, k=2,3,...,D;\; j=1,2,...,k-1. \qquad \text{(Eq. 6)}$$

These variances can be arranged in a symmetric matrix with parts defining both D rows and D columns, with the same layout as a correlation matrix. This is the so-called *variation matrix*. Variance (Eq. 6) is zero when z_j and z_k behave perfectly proportionally (compositions with twice the amount of part j also have twice the amount of part k), corresponding to perfect positive association. It goes without saying that a part is proportional to itself, hence the zeros in the matrix diagonal. The further variance (Eq. 6) is from zero, the lower the association. There is neither a clearly defined threshold representing no association, nor is there an upper bound representing perfect negative association, so values in the variation matrix can only be assessed comparatively.

This comparative assessment may be carried out relatively to the mean log-ratio variance (Pawlowsky-Glahn et al. 2015). There are $D(D–1)/2$ distinct elements in the variation matrix:

$$\frac{1}{D\left(D-1\right)/2}\sum_{k=2}^{D}\sum_{j=1}^{k-1}Var\left(\ln\left(\frac{z_j}{z_k}\right)\right). \qquad \text{(Eq. 7)}$$

Log-ratio variances larger than (Eq. 7) show pairs of parts contributing to a larger share of the variation matrix than the average log-ratio, and variances lower than (Eq. 7) show pairs of parts with a small contribution, in other words, with positive association. Strong association can be inferred, for instance when the ratio of (Eq. 6) over (Eq. 7) is lower than 0.2 (Egozcue and Pawlowsky-Glahn 2019).

2.2.3. Total variance

Total variance in a compositional data set can be computed in two alternative equivalent manners. Firstly as the sum of variances of the D centered log-ratios:

$$\sum_{j=1}^{D}Var\left(\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 ... z_D}}\right)\right), \qquad \text{(Eq. 8)}$$

and secondly from the sum of the distinct elements in the variation matrix:

$$\frac{1}{D}\sum_{k=2}^{D}\sum_{j=1}^{k-1}Var\left(\ln\left(\frac{z_j}{z_k}\right)\right). \qquad \text{(Eq. 9)}$$

2.2.4. Distance

Aitchison’s distance (Aitchison 1983; Aitchison et al. 2000) between two individual compositions $\mathbf{z}$ and $\mathbf{z}^*$ considers that pairwise log-ratios (Eq. 3) carry all the required information about the difference between them:

$$\mathrm{d}\left(\mathbf{z},\mathbf{z}^*\right)=\sqrt{\frac{1}{D}\sum_{k=2}^{D}\sum_{j=1}^{k-1}\left(\ln\frac{z_j}{z_k}-\ln\frac{z_j^*}{z_k^*}\right)^2}\,. \qquad \text{(Eq. 10)}$$

Two compositions at zero distance have identical part proportions. When there is a larger difference between the log-ratios of two compositions, their distance is likewise larger. Aitchison’s distances can also be expressed in terms of centered log-ratios (Eq. 4) as:

$$\mathrm{d}\left(\mathbf{z},\mathbf{z}^*\right)=\sqrt{\sum_{j=1}^{D}\left(\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 \ldots z_D}}\right)-\ln\left(\frac{z_j^*}{\sqrt[D]{z_1^* z_2^* \ldots z_D^*}}\right)\right)^2}\,. \qquad \text{(Eq. 11)}$$

Expression (Eq. 11) has the attractive feature that it equals Euclidean distance computed from data transformed as centered log-ratios (Eq. 4). Computing centered log-ratios from equation (Eq. 5) makes no difference.

2.3. Data preprocessing

2.3.1. Zero replacement

As is well known, computing log-ratios implies that **x** and **z** may contain no zero values. If the **x** and **z** vectors contain zeros, they must be replaced beforehand (Martín-Fernández et al. 2011). Treatment of zeros in CoDa depends on the assumed reason for their occurrence, which is deemed more important than their sheer existence.

On the one hand, there are *absolute zeros, essential zeros*, or *structural zeros*, which represent values that can only be zero given certain characteristics of the individual compositions (e.g., nature pictures in a hotel located in an urban environment, tobacco consumption in a non-smoking home). The presence of this kind of zeros may lead to different variance structures of the parts of interest, and usually indicates that the choice of parts to be analyzed is not meaningful to a certain subpopulation. Thus, data with absolute zeros should be considered as distinct subpopulations (Bacon-Shone 2003) and either be excluded (e.g., by analyzing only hotels in a natural environment) or analyzed separately. Amalgamation of problematic parts can constitute an alternative, if the researcher is happy with its implications for the definition of the research questions.

On the other hand, so-called *rounded zeros, trace zeros*, or *zeros below detection limit* constitute parts which are believed to be present, but are not observed due to randomness or limitations of measurement. Consider a study about dollar spending in e-shops by product categories (parts: apparel, books, music, travel, hobbies and other). Certain consumption values may be zero in a short reference period, but might not be if observed over a longer period. They are, thus, analogous to missing data with the added information that they have to be below a detection limit. If there is no external or theoretical indication on what the detection limit should be, it can be set as the minimum observed value of each part. The situation is therefore analogous to missing value imputation and zeros can be replaced with a value below the detection limit following certain criteria. Palarea-Albaladejo and Martín-Fernández (2008; 2015) modified the well-known *EM imputation method* to the compositional case by introducing the restriction that imputed values are below the detection limit. At least one part must be complete for all individuals. If this is not the case the researcher can take the part with fewest zeros and previously replace zeros with a small amount around two thirds of the detection limit.

Finally, the **x** data can also be counts of phenomena, whose sum for the *i*th individual is S_i. For instance, an individual's total count of S_i tweets can be classified into D content categories. Our hotel photo and complaint examples also constitute count data. The counts of the *i*th individual can be considered to be a realization of a multinomial distribution with θ_{i1}, θ_{i2},…, θ_{iD} unobserved non-zero probabilities. Even if these probabilities are non-zero, a combination of a small probability and a small S_i may result in certain x values being zero, referred to as *count zeros*. This opens up the possibility of using the Bayesian methods described in Martín-Fernández et al. (2015). An alternative approach is to treat count zeros as rounding zeros, which is considered appropriate if the total counts S_i are large (Filzmoser et al. 2018). In this case, detection limits are straightforward. Since the minimum observable count is 1, the detection limit in terms of the closed composition can be set for each individual at $1/S_i$.

The references in this section acknowledge the fact that zero imputation can introduce distortion when the proportion of zero values to be imputed in the data set is large. What constitutes a large proportion may depend on many circumstances, but in many cases sizeable distortion starts occurring when around 15% or 20% of data are zeros. In this case, dropping parts with many zeros by means of a subcomposition analysis, or amalgamating them together with other parts can mitigate the distortion, although it goes without saying that it affects the definition of the compositional research questions.

2.3.2. Multivariate outlier detection

Zero replacement is usually the first step in CoDa, and some sort of *outlier* diagnostics the second. CoDa has implications for outlier detection. Given the fact that parts cannot be considered in isolation, multivariate outlier detection methods are called for (Aitchison 1986). Once compositions have been transformed into centered log-ratios, squared *Mahalanobis distances* between each composition and the overall mean can be computed (Filzmoser and Hron 2008). Mahalanobis distances measure how far away each individual is from the center, taking into account the variances and covariances among log-ratios. It must be taken into account that Mahalanobis distances require inverting the covariance matrix, and thus they cannot be applied on the whole D centered log-ratios. In this case the situation can be solved by just leaving one of the centered log-ratios out. Fortunately, results are invariant to the decision on which one is left out.

Under multivariate normality, these squared Mahalanobis distances follow a χ^2 distribution with D–1 degrees of freedom. An appropriate percentile for this distribution can be used as cut-off criterion for outlier detection. This percentile should not be uncritically set to the usual 0.95 cut-off criterion but should take sample size into account. For instance, if the sample size n=1000 and one would use the 0.95 cut-off, around 50 cases would appear as outliers even if no true outlier was present at all. To set the cut-off, for instance, at the 0.999 percentile would be far more reasonable. An exact percentile which adapts the common 0.95 practice to the existing sample size can be obtained as $0.95^{(1/n)}$. Since Mahalanobis distances are themselves affected by outliers, an alternative is to compute robust Mahalanobis distances (Filzmoser et al. 2005; 2018).

2.4. Compositional principal component analysis and the CoDa biplot

Like standard data, compositional data require visualization tools to help researchers interpret large data tables with many individuals and parts. To this end, Aitchison (1983) extended the well-known *principal component analysis* procedure to the compositional case. This method belongs to the family of *multivariate statistical analysis* and the extension boils down to submitting centered log-ratios (Eq. 4) to an otherwise standard principal component analysis based on the covariance matrix. Together with Gabriel's (1971) *biplot*, which jointly represents cases (i.e., individual compositions) and variables (i.e., parts) in a principal component analysis, this served as the basis for Aitchison and Greenacre (2002) developing CoDa biplots.

A compositional principal component analysis computes uncorrelated linear combinations of the centered log-ratios which explain the highest possible portion of total variance (Eq. 8), called dimensions. The two first dimensions are represented in the CoDa biplot, which can be understood as the most accurate graphical representation of a compositional data set in two dimensions (or three dimensions). As in standard principal component analysis, overall biplot accuracy can be assessed from the percentage of the total variance (Eq. 8) explained by the first two dimensions. The accuracy of the representation of each part can likewise be computed from the percentage of variance of each centered log-ratio explained by the first two dimensions (Daunis i Estadella et al. 2011).

In particular, the so-called *covariance biplot* is the most commonly drawn type in CoDa. It optimizes the representation of the variation matrix among parts. Proximity among individuals is not interpretable in this type of biplot. Parts appear as rays emanating from a common origin and individual compositions appear as points. The origin of coordinates represents the composition center. The interpretation is as follows (see Aitchison and Greenacre 2002; Blasco-Duatis et al 2019; van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015 for further details):

1. Distances between the vertices of the rays of two parts are approximately proportional to the square root of the variance of their corresponding pairwise log-ratio (Eq. 6). Parts that behave proportionally for all individuals appear close together. It must be noted that unlike the general principal component analysis case, in the CoDa biplot angles between rays play no interpretational role.

2. The orthogonal projection of all individuals in the direction defined by a ray shows an approximate ordering of the importance of that part for all individuals, in relative terms, compared to the geometric average of the remaining parts in the composition.

Compared to standard principal component analysis, in CoDa parts can never have all coordinates of the same sign on any dimension, stressing the fact that along any dimension some parts increase and others decrease, in relative terms.

Like standard principal component analysis, compositional principal component analysis is not only a visualization tool, but also a data reduction tool. The first few dimensions contain a summary of the compositional information and can be used as numeric variables in further statistical analyses, provided that they can be interpreted. The composition can thus be related to external non-compositional variables, by means of correlations if the external variable is numeric, or by comparing the dimension means by a categorical external variable.

2.5. Compositional cluster analysis

Like standard data, compositional data can benefit from classifying individual compositions into groups of compositions, called clusters, which are mutually similar. In other words, pairs of compositions within the same cluster have lower Aitchison's distances than pairs of compositions belonging to different clusters. Yet, in other words, the sum of centered log-ratio variances within clusters is as small as possible. *Cluster analysis* is the typical multivariate statistical analysis method for this purpose and many alternative *clustering* methods are available. An attractive feature of *compositional cluster analysis* is that once centered log-ratios have been computed any standard cluster analysis method supporting Euclidean distances can be used (Ferrer-Rosell and Coenders 2018; Godichon-Baggioni et al. 2019; Martín-Fernández et al. 1998). This includes, among others, *Wards' method* and the *k-means* method. Any such method can be applied with standard software on the centered log-ratios, and will provide equivalent results to clustering based on Aitchison's distances.

In particular, the k-means method minimizes the sum of variances of all centered log-ratios within clusters, as a measure of intra cluster similarity. For a classification into *k* groups, *k* initial cluster centers are selected randomly. Each individual composition is assigned to the nearest center and centers are iteratively updated according to the assigned individuals until no individual changes membership. Since the procedure may fall into a local minimum of within cluster variance, the procedure may be replicated a large number of times with different sets of random initial cluster centers.

As regards the cluster interpretation, cluster profiles can be described by means of within-cluster means of the centered log-ratios, if necessary exponentiated and closed back to the original composition unit sum. A standard graphical representation of cluster analysis results in CoDa is the *geometric mean barplot*. This plot depicts the log-ratios of the closed cluster means of each part over the closed mean of that part for the overall sample. Positive bars show above average parts for that particular cluster and negative bars below average parts. Since CoDa focuses on relative information, no cluster will ever have the highest or the lowest means on all parts. A well-known terminology distinguishes between clustering based on *size* and clustering based on *shape*, CoDa belonging to the latter category (Greenacre 2017).

The main procedural difference compared to standard cluster analysis is that standardization of centered log-ratios is not desirable because it modifies distances and would thus make Euclidean distances no longer equivalent to Aitchison's distances. In most other respects the analysis is carried out like a cluster analysis on any numeric data set.

Decisions on the number of clusters (*k*) are made as usual. In any case these decisions involve a trade-off between accuracy and parsimony; in other words, the higher the desired similarity of individuals within the clusters, the higher the number of required clusters. A pragmatic approach can involve to start with a low number of clusters *k* and keep adding clusters as long as the profiles of the additional clusters are meaningfully different, and as long as none of the clusters is too small for practical purposes.

A usual statistical measure of the aforementioned trade-off between accuracy and parsimony is the *Calinski index*, higher values tending to show a good choice of *k:*

$$\frac{\left(\text{total sum of variances - within cluster sum of variances}\right)\Big/\left(k-1\right)}{\text{within cluster sum of variances}\Big/\left(n-k\right)}, \qquad \text{(Eq. 12)}$$

where total sum of variances is (Eq. 8). Another statistical measure is the *average silhouette width* comparing average distances of each case with all cases in its own cluster and with all cases in the second best neighboring cluster. Higher values also tend to show a good choice of *k*.

Relationships between the cluster-membership variable and external non-compositional variables are also analyzed with the usual statistical tools in any cluster analysis. Such relationships constitute a convenient way to relate the composition to other variables in further statistical analyses. The simplest methods are contingency tables when the external variable is categorical, and analysis of variance when the variable is numeric.

2.6. Limitations and extensions

The inability to work with sparse data tables with many zeros is indeed one of the most often quoted limitations of CoDa. This precludes using CoDa, for instance in web mining of short texts if single words or single word combinations are treated as parts. Alternatives such as *correspondence analysis* are recommended in these cases (Greenacre 2018).

Another often quoted limitation is that, in a log scale, parts with very small values may end up dominating the analysis results. Advanced methods for down-weighting small parts are discussed in Greenacre (2018). Amalgamation of very small parts is an alternative, as long as it is coherent with the research problem definition.

This chapter has only presented descriptive methods. Of course CoDa lends itself to statistical inference. The composition can be the dependent or the explanatory variable in statistical models ranging from simple multivariate analysis of variance or regression models to mixture models, time series models, generalized linear models, and structural equation models (Filzmoser et al. 2018; Pawlowsky-Glahn et al. 2015), with many applications in the tourism field (Coenders and Ferrer-Rosell 2020). To make these applications possible, alternative log-ratio transformations (Egozcue et al. 2003), robust methods, and methods for high dimensional data (Filzmoser et al. 2018) have been duly developed. In the particular case of text content analysis, a noteworthy variation on the theme is that by Roberts et al. (2016), a multi-step procedure including a compositional regression.

2.7. Example

2.7.1. Data

The example presented in this chapter shows that CoDa methodology serves as an important complementary tool for content analysis. In this case, the aim of the application is to analyze the hotel reviews' content, and more particularly to relate complaint topics to one another, as well as to distinguish hotel clusters based on major complaint topics. The primary unit of content analysis is the review in itself. Some of the other available and relevant variables per review are: review identification, review date, hotel identification, user identification and score given (from 1 to 5).

Hotel reviews of the city of Barcelona were downloaded on September 2016 from TripAdvisor as it is one of the leading online traveler opinion platforms (Martin-Fuentes 2016). The downloading process was done automatically with a web scraper tool developed in Phyton and the process took less than 24 hours to obtain a random selection of 31,000 reviews from hotels of all categories.

For the example, out of the total hotels included in the sample we selected those which had at least 150 reviews (n=50 hotels). Then, we randomly selected 50 reviews of each hotel. Thus, 2,500 reviews were analyzed.

The topics of complaints were deduced from the content analysis of reviews. The hotel is the unit of statistical analysis in the compositional data set and counting the topics of complaints in all reviews of each hotel constitutes count data. The hotel's total count of identified contents in the 50 reviews (S_i) was classified into D=8 topics (content categories or parts). The topics were: nothing (the review did not include any complaint or negative comment); facilities (the review included negative comments about hotel facilities in general, beds, rooms in general, bath, or common facilities); services provided (the review included negative comments about the services provided by the hotel such as the breakfast, the Wi-fi, the bar/restaurant, the pool, etc.); cleanliness (the review included negative comments about the hotel cleanliness in general, and about the rooms in particular); location (the review included negative comments related to the location of the hotel); environment (the review included negative comments about the neighborhood, external noise, etc.); staff (the review included negative comments about staff, for example, staff not being helpful or problem solving); other complaints (the review included negative comments unrelated to the former topics). Apart from the reviews, the average hotel score was also used to relate it to the biplot dimensions and to describe the clusters.

2.7.2. Results

The location part has a large percentage of zeros (42.0%). Its conceptual similarity with the environment part makes amalgamation a feasible option. Both parts are not under the control of the hotel management, at least in the short term, but depend mostly on where the hotel is located. We name the amalgamated part environment, understanding that it covers both concepts.

After amalgamation, the percentage of zeros (12.57%) is deemed appropriate for replacement by the modified EM algorithm by setting the detection limits at $1/S_i$.

If the relative importance of complainers versus not complainers is outside of the example focus and the main aim is to study the distribution of the importance of complaints by topics, then a subcomposition excluding the nothing part makes sense. In the rest of the example we concentrate on the parts facilities, services, cleanliness, staff, environment and other. The outlier detection threshold is set at $0.95^{(1/50)} = 0.9990$. No outliers are found.

Table 1 shows the variation matrix and the center. The average of the variation matrix elements is 0.867. Pairs of parts with a log-ratio variance below 0.2×0.867=0.173, if any, would be considered to move proportionally. Conversely, the pairs of parts environment versus services and environment versus staff have comparatively high log-ratio variances, meaning that hotels with relatively more complaints about environment tend to have relatively fewer about services and staff. The center shows that the most often quoted reasons for complaining are facilities, environment and services, and the less often quoted reasons are staff and cleanliness.

Table 1. Center, variation matrix, and centered log-ratio variances

	Center	**Environment**	**Services**	**Other**	**Staff**	**Cleanliness**	**Clr variances**
Facilities	0.344	0.728	0.862	0.797	0.942	0.406	0.261
Environment	0.243		1.379	1.045	1.433	0.625	0.507
Services	0.157			0.817	0.828	0.788	0.418
Other	0.103				0.771	0.694	0.326
Staff	0.092					0.891	0.449
Cleanliness	0.062						0.206
							2.168

Fig. 1 shows the biplot. The distances among pairs of rays closely mirror the log-ratio variances in Table 1. Orthogonal projections along the directions defined by a ray constitute an approximate ordering of hotels according to the ratio of the frequency of a complaint topic over the geometric mean of the frequency of the remaining complaints. For instance, hotel 1 has the largest frequency of complaints about services in relative terms and hotel 24 the lowest. Hotels in the upper left quadrant stand out for having relatively more complaints on other reasons and staff and relatively fewer on cleanliness and facilities. The percentage of explained variance by the first two dimensions is deemed satisfactory at 60.5% thus arguing for a good biplot accuracy.

Fig. 1. CoDa biplot of hotel complaints

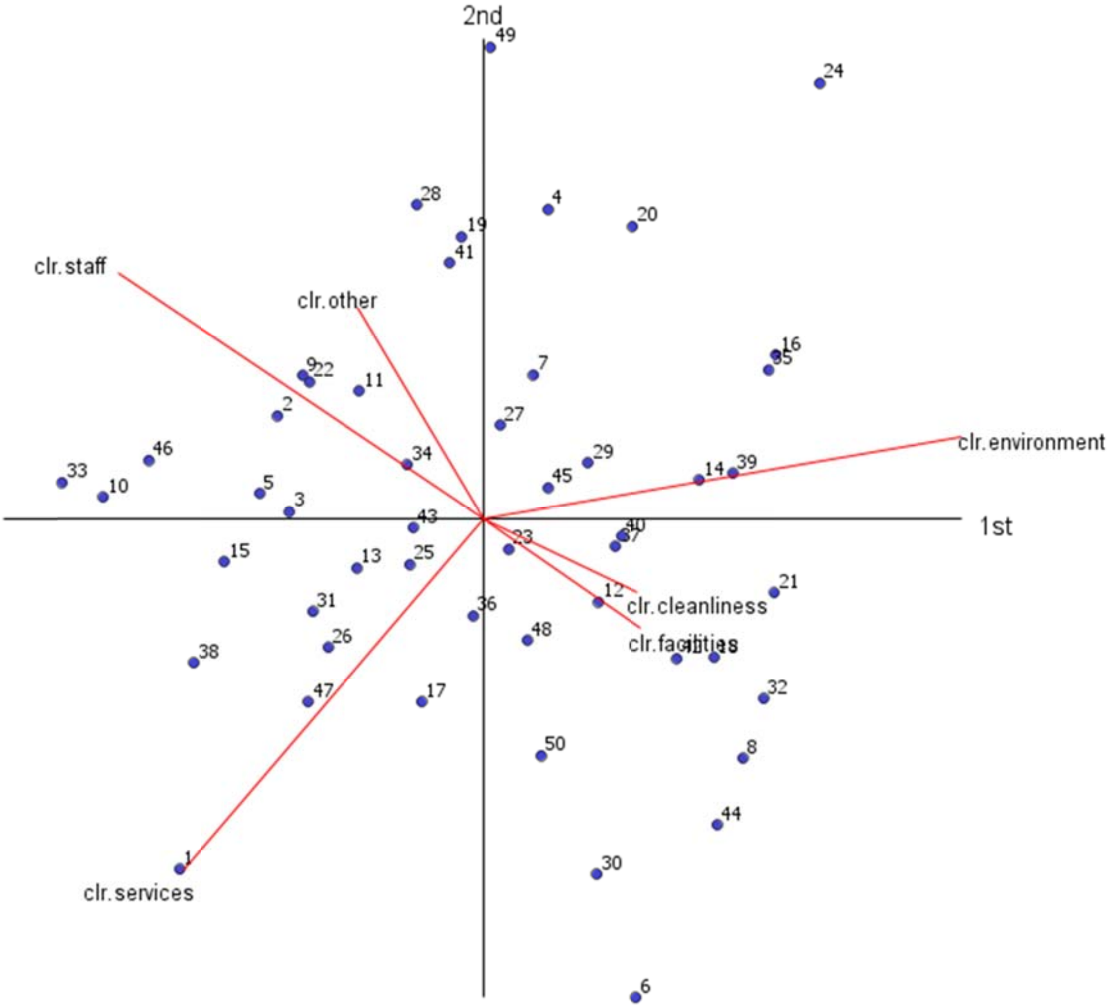


Table 2. Correlations between biplot dimensions and hotel score (mean over 50 reviews)

	1st dimension (horizontal axis)	**2nd dimension (vertical axis)**
Hotel score (mean)	-0.033	-0.194

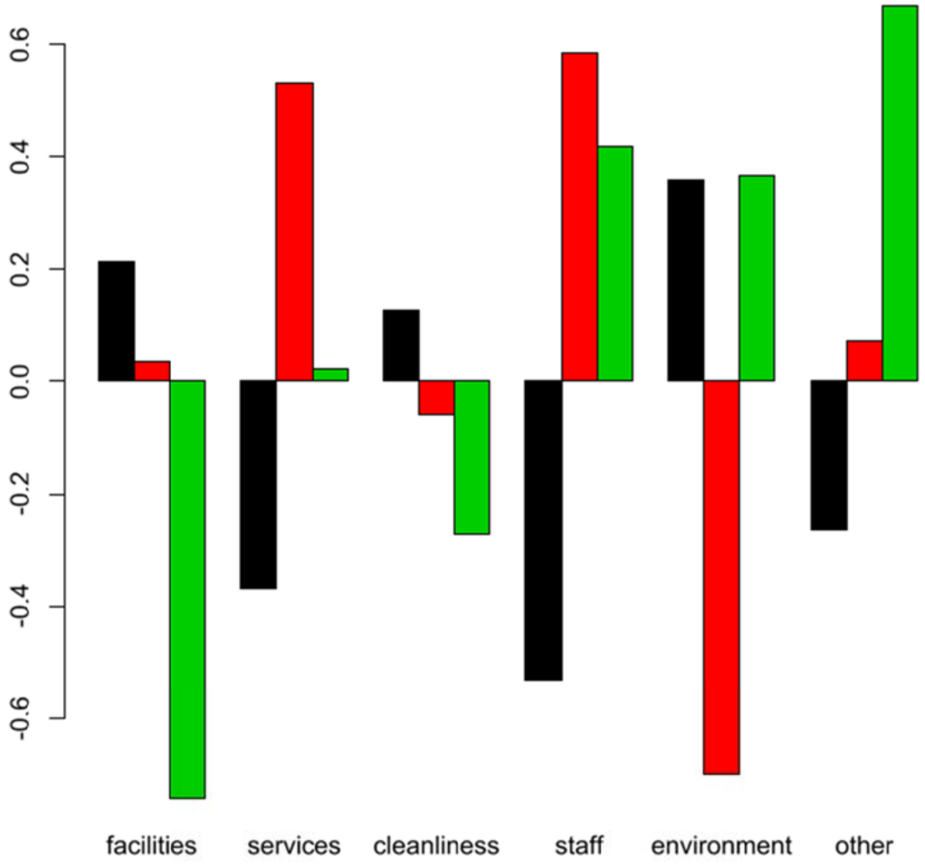
According to Table 2, hotels with the highest average scores are at the bottom of biplot, thus, satisfied reviewers tend to complain more about services. The most unsatisfied customers are those lodged at hotels located at the upper part of the biplot and tend to complain more about staff. Having said this, the correlation is admittedly low.

The results of a *k*-means clustering with three clusters are depicted in the geometric mean barplot in Fig. 2. The first cluster (black bars, *n*=25) stands out for relatively more complaints on environment, cleanliness, and facilities, and relatively fewer on staff, services and other topics, compared to the overall sample average. The second cluster (red bars, *n*=17) stands out for relatively more complaints on services and staff, and relatively fewer on environment, compared to the overall sample average. Finally, the third cluster (green bars, *n*=8) stands out for relatively more complaints on other topics, staff and environment, and relatively fewer on facilities and cleanliness. These results can be numerically observed in Table 3, which shows the centers of the parts of each cluster.

Table 3. Centers of parts in each cluster

	Facilities	Services	Cleanliness	Staff	Environment	Other
Cluster 1	0.392	0.100	0.065	0.049	0.321	0.073
Cluster 2	0.330	0.248	0.054	0.153	0.112	0.103
Cluster 3	0.154	0.151	0.044	0.132	0.330	0.189

Fig. 2. Geometric mean barplot of clusters and complaint topics



The three clusters are also depicted in Fig. 3. Half of hotels are included in the first cluster, with relatively more complaints about facilities, cleanliness and environment. According to Table 4, hotels in clusters 1 and 3 are the ones with highest average and median hotel score, while hotels of second cluster are worst evaluated (lowest hotel score mean and median), although differences are admittedly minimal. Fig. 4 shows the corresponding boxplots.

Table 4. Statistics of hotel score per cluster. Mean, standard deviation and percentiles.

	Mean	Std. Dev.	0	25	50	75	100
Cluster 1	4.115	0.443	2.840	3.940	4.200	4.380	4.960
Cluster 2	4.007	0.417	3.400	3.640	4.120	4.300	4.660
Cluster 3	4.168	0.395	3.560	4.000	4.240	4.460	4.880

Fig. 3 CoDa biplot of hotel complaints per cluster

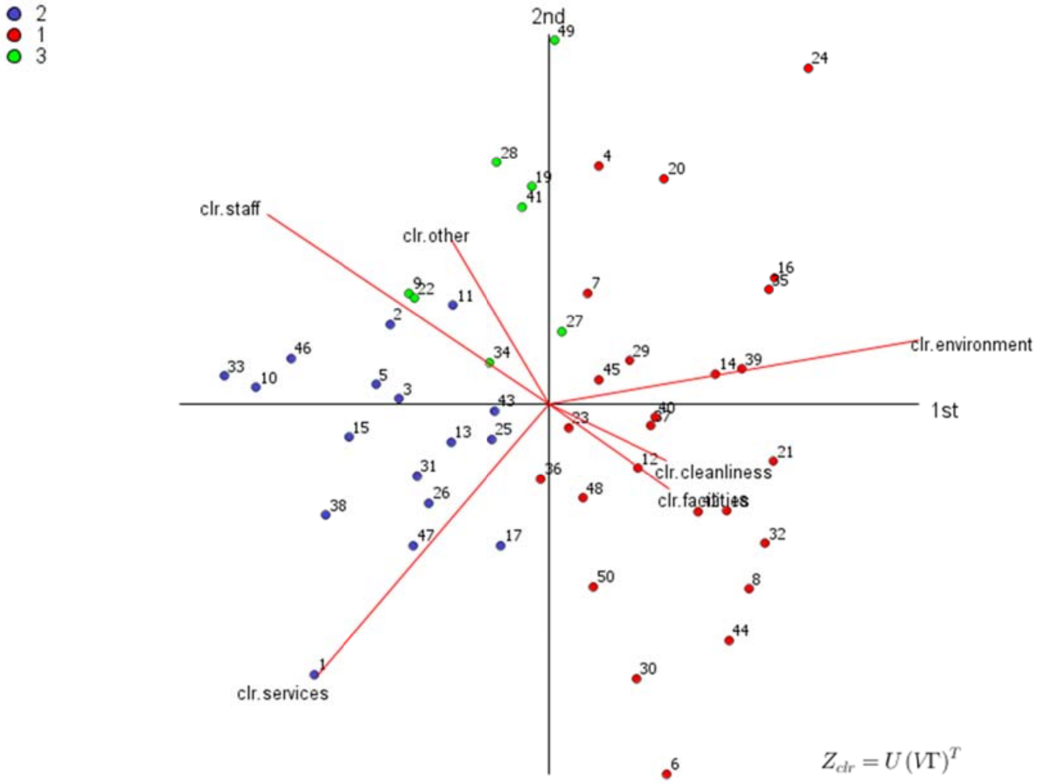
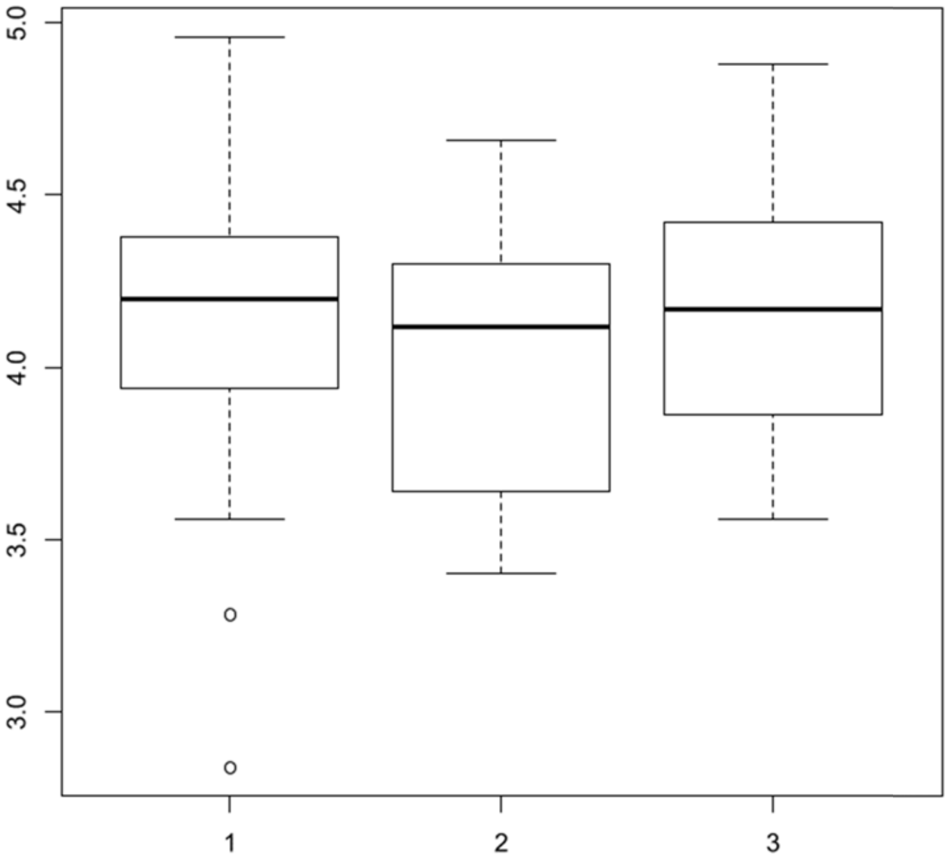


Fig. 4 Box plots of hotel score per cluster



Summing up, the application of the CoDa methodology in this example has made it possible to plot the relative importance of complaint topics for each specific hotel and to draw clusters with different complaint profiles, which are related to the hotel average score and could also be related to any other hotel characteristic.

3. Conclusion

As stated throughout the chapter, data carrying relative information have particular characteristics which may lead to interpretational difficulties, among other problems, when using standard statistical analyses. The proportional nature of data must then be taken into account from the onset. For instance, we cannot consider the distance between 1% and 2% to be the same as between 11% and 12%, which is what the Euclidean distance does. In the first pair the increase is 100%, while in the second pair, it is less than 10%. Most standard and classical statistical methods do not consider the restricted nature of the data expressed as proportions, and are subject to spurious correlations among the parts, and violation of the statistical and distributional assumptions, for instance, normality. The main advantage and also the appeal of the CoDa methodology is that it solves the aforementioned problems. It is also worth mentioning that once the data (components) have been transformed into logarithms of ratios, any present and future standard statistical technique may be used, since the relative importance of the parts is put on the table, and normality is recovered.

As stated in Ferrer-Rosell (2021), CoDa has already been used to analyze e-tourism data. The CoDa methodology in e-tourism is considered to be an ideal complement to content analysis techniques, and to research regarding dominance of contents in any kind of (online) source. Regarding the future of the CoDa methodology in e-tourism, apart from being a simple and straightforward tool to use when researchers focus on proportions, it also passes through considering the total (volume) of contents. The total has been considered to advantage in research about tourist expenditure, where it is interesting to analyze the distribution of the trip budget and the total trip budget in the same statistical model (Ferrer-Rosell et al., 2016b), but has not been used in e-tourism yet. In the e-tourism context, analyzing the composition of contents (which contents are more emphasized in online sources) in, for instance, social media, is as relevant as analyzing the total number of posts, or its ratio to the number of tourists at the destination, for instance. The total number of posts in social media according to the number of visitors determines how active the social media profiles are. Other possible developments are to take advantage of the usability of any statistical technique on the log-ratios, including the composition as dependent, explanatory or mediating variable in static or dynamic models, although more advanced log-ratio transformations than those presented in this chapter are sometimes needed (Filzmoser et al., 2018; Pawlowsky-Glahn et al., 2015).

Appendix. CoDaPack menus used for the example

CoDaPack is an intuitive menu-driven freeware for CoDa developed by the *Research Group in Statistics and Compositional Data Analysis at the University of Girona*. The philosophy of CoDaPack is to reduce the analysis steps the users must perform by themselves. The program computes by itself the needed log-ratios for each type of analysis. CoDaPack can be downloaded at:

http://ima.udg.edu/codapack/

The *File* menu handles opening and saving data files, including importing and exporting them in a variety of formats, at the moment of writing this chapter **.xls**, **.csv**, **.txt** and **.RData**. Ideally the file contains some columns indicating a closed composition (Eq. 2) together with non-compositional numeric and categorical variables as wished by the researcher.

Zeros are not coded as “0”, but coded as below a certain detection limit, which may be different for each zero cell. For instance, if a value is known to be below 0.005, the data file entry in **xls**, **.csv**, and **.txt** formats is “<0.005”.

The *Irregular data*➢*Logratio-EM zero replacement* menu draws from the original closed composition (Eq. 2). The data file columns containing the parts in the closed composition are the variables to be selected by the user for analysis in this procedure and in most CoDaPack procedures. Even if the procedure is intended to replace rounded zeros with the EM method, it can also be used to replace count zeros if each zero is considered to have a detection limit equal to $1/S_i$, which must be entered as such in the original data file. For instance, if the total count for an individual composition is 40, all zeros in the row of that individual are coded as “<0.025”. A useful complement is the *Irregular data*➢*ZPatterns plot* menu, which computes percentages of zeros per part, and plots combinations of parts for which zeros tend to co-occur. This can be useful in suggesting feasible amalgamations if the percentage of zeros is very large. The *Data*➢*Manipulate*➢*Calculate new variable* menu can be used to create new variables such as the sum of the parts to be amalgamated.

The *Irregular data*➢*Atypicality index* menu draws from the original closed composition (Eq. 2) and computes a binary variable which marks outliers, if any, based on the desired percentile of the χ^2_{D-1} distribution for standard Mahalanobis distances The user can select the desired percentile under *Level of confidence*, for instance the result of computing $0.95^{(1/n)}$. The percentile itself is also stored in the data file for each individual. If any atypical values are encountered, the researcher may wish to remove them from the analysis by means of the *Data*➢*Filters*➢*Categorical filter* menu.

The *Data*➢*Transformations*➢*CLR* menu computes a new set of variables as the centered log-ratios (Eq. 4) from the selected components expressed as a closed composition (Eq. 2). The data file containing these transformed variables can be exported for use with the researchers’ favorite statistical software, in order to carry out any analysis not yet supported by CoDaPack.

The *Statistics*➢*Compositional statistics summary* menu draws from the raw closed compositional data (Eq. 2) and computes two types of descriptive statistics: the first related to pairwise log-ratios (Eq. 3) (a matrix with variances (Eq. 6) as in the variation matrix above the diagonal, and the means of pairwise log-ratios below the diagonal), and the second related to centered log-ratios (the variance of each centered log-ratio, adding up to the total variance (Eq. 8)) and the center, closed to unit sum.

The *Statistics*➢*Classical statistics summary* menu should not be applied to the raw composition, but can be correctly applied to the centered log-ratios (Eq. 4) which have been previously computed and stored by means of the *Data➢Transformations➢CLR* menu. Means, percentiles and standard deviations are especially useful.

The *Graphs➢CLR biplot* menu draws from the raw closed compositional data (Eq. 2) and computes the covariance biplot (*Cov.* default) and other types of biplots. First it shows a bidimensional plot, but it can be rotated to show the third dimension. The principal component dimensions can be added to the data file when selecting the *Add coordinates* option. Points can be colored by any categorical variable (*Groups* option). The individual cases can also be identified by row number. Next, the dimensions can be related to external numeric variables by means of correlation coefficients (*Statistics➢Classical statistics summary* menu, *Correlation matrix* option) or to external categorical variables by computing the dimension means per category (*Statistics➢Classical statistics summary* menu, *Mean* option, by introducing the categorical variable into *Groups*).

The *Statistics➢Multivariate analysis➢Cluster➢k-means* menu draws from the raw closed compositional data (Eq. 2). The user selects the desired *Number of clusters*. The program shows the best solution out of 25 random sets of initial cluster centers. The results include the Calinski Index, the average silhouette width and a new variable containing cluster membership. The program can also select the number of clusters maximizing the Calinski index or the average silhouette width. The user may compute the cluster centers under the *Statistics➢Compositional statistics summary* selecting the closed compositional data (Eq. 2), the *Center* and *Groups* options. The user can also select the *Graphs➢Geometric mean barplot* menu by selecting the closed compositional data (Eq. 2) and introducing the cluster membership variable into *Groups*. The *Statistics➢Classical statistics summary* menu can be used to describe non-compositional variables separately by the group membership variable. The same can be accomplished with box plots by means of the *Graphs➢boxplot* menu. Finally, the biplot can be redrawn with the cases colored by cluster.

Acknowledgements

The authors acknowledge the support of the Catalan Government for the accreditation as Consolidated Research Groups TURESCO (2017 SGR 49) and COSDA (2017 SGR 656). This work was supported by the Spanish Ministry of Economy, Industry and Competitiveness (Grant id.: TURCOLAB ECO2017-88984-R), by the Spanish Ministry of Science, Innovation and Universities and FEDER (Grant id.: CODAMET RTI2018-095518-B-C21). Finally, authors also acknowledge the comments by Santi Thió-Henestrosa.

Cross references

Content analysis of travel reviews.

References

Aitchison J (1982) The statistical analysis of compositional data. J Roy Stat Soc B Met 44(2):139-177.

Aitchison J (1983) Principal component analysis of compositional data. Biometrika, 70(1):57-65.

Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman and Hall, London.

Aitchison J (2001) Simplicial inference. In: Marlos AGV, Richards DSP (eds) Algebraic methods in statistics and probability: AMS special session on algebraic methods in statistics. Contemporary mathematics series. American Mathematical Society, Providence, p 1-22.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distances. Math Geol 32(3):271-275.

Aitchison J, Greenacre M (2002) Biplots of compositional data. J Roy Stat Soc C App 51(4):375-392.

Bacon-Shone J (2003) Modelling structural zeros in compositional data. In: Thió-Henestrosa S, Martín-Fernández JA (eds) Proceedings of CoDaWork'03, the 1st compositional data analysis workshop.

Barceló-Vidal C, Martín-Fernández JA (2016) The mathematics of compositional analysis. Austrian J Stat 45(4):57-71.

Batista-Foguet JM, Ferrer-Rosell B, Serlavós R, Coenders G, Boyatzis RE (2015) An alternative approach to analyze ipsative data. Revisiting experiential learning theory. Front Psychol 6:1742.

Blasco-Duatis M, Coenders G (2020) Sentiment analysis of the agenda of the Spanish political parties on Twitter during the 2018 motion of no confidence. A compositional data approach. Revista Mediterránea de Comunicación 11(2):185-198

Blasco-Duatis M, Coenders G, Sáez M, Fernández-García N, Cunha I (2019) Mapping the agenda-setting theory, priming and the spiral of silence in twitter accounts of political parties. Int J Web Based Communities 15(1):4-24.

Blasco-Duatis M, Sáez-Zafra M, Fernández-García N (2018). Compositional representation (CoDa) of the agenda-setting of the political opinion makers in the main Spanish media groups in the 2015 General Election. Commun Soc 31(2):1-24.

Van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing compositional data with R. Springer, Berlin.

Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (2006) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London.

Carreras-Simó M, Coenders G (2020) Principal component analysis of financial statements. A compositional approach. Rev Métodos Cuant Econ Empresa, 29:18-37.

Coenders G, Ferrer-Rosell B (2020) Compositional data analysis in tourism. Review and future directions. Tour Anal, 25(1):153-168.

Coenders G, Martín-Fernández JA, Ferrer-Rosell B (2017) When relative and absolute information matter. Compositional predictor with a total in generalized linear models. Stat Model 17(6):494-512.

Daunis i Estadella J, Thió i Fernández de Henestrosa S, Mateu i Figueras G (2011) Two more things about compositional biplots: quality of projection and inclusion of supplementary elements. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) Proceedings of the 4th international workshop on compositional data analysis.

Egozcue JJ, Pawlowsky-Glahn V (2019) Compositional data: the sample space and its structure. TEST 28(3):599-638.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35(3):279-300.

Van Eijnatten FM, van der Ark LA, Holloway SS (2015) Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis. Qual Quant 49(2):559-579.

Ferrer-Rosell B (2021) Compositional analysis of tourism-related data In: Correia A, Dolnicar S (eds) Women's voices in tourism research. Contribution to knowledge and letters to future generations 2021. The University of Queensland, Brisbane, p 182-188.

Ferrer-Rosell B, Coenders G (2017) Airline type and tourist expenditure: Are full service and low cost carriers converging or diverging? J Air Transp Manag 63:119-125.

Ferrer-Rosell B, Coenders G (2018) Destinations and crisis. Profiling tourists' budget share from 2006 to 2012. J Destin Mark Manag 7:26-35.

Ferrer-Rosell B, Coenders G, Martínez-Garcia E (2015) Determinants in tourist expenditure composition - the role of airline types. Tour Econ 21(1):9-32.

Ferrer-Rosell B, Coenders G, Martínez-Garcia E (2016a) Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. Tour Anal 21(6):589-602.

Ferrer-Rosell B, Coenders G, Mateu-Figueras G, Pawlowsky-Glahn V (2016b) Understanding low cost airline users' expenditure patterns and volume. Tour Econ 22(2):269-291.

Ferrer-Rosell B, Marine-Roig E (2020) Projected versus perceived destination image. Tour Anal, 25(2-3):227-237.

Ferrer-Rosell B, Martin-Fuentes E, Marine-Roig E (2020) Diverse and emotional: Facebook content strategies by Spanish hotels. Inform Technol Tour, 22(1):53-74.

Ferrer-Rosell B, Martin-Fuentes E, Marine-Roig E (2019) Do hotels talk on Facebook about themselves or about their destinations? In: Pesonen J, Neidhardt J (eds) Information and communication technologies in tourism 2019. Springer, Cham, p 344-356.

Ferrer-Rosell B, Martin-Fuentes E, Vives-Mestres M, Coenders G (2021) When size does not matter: compositional data analysis in marketing research. In: Nunkoo R, Teeroovengadum V, Ringle C (eds) Handbook of research methods for marketing management. Edward Elgar, Cheltenham: p 73-90.

Filzmoser P, Garrett RG, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. Comput Geosci 31(5):579-587.

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Math Geosci 40(3):233-248.

Filzmoser P, Hron K, Templ M (2018) Applied compositional data analysis with worked examples in R. Springer, New York.

Fry T (2011) Applications in economics. In: Pawlowsky-Glahn V. Buccianti A (eds) Compositional data analysis. Theory and applications. Wiley, New York, p 318–326.

Gabriel KR (1971) The biplot-graphic display of matrices with application to principal component analysis. Biometrika 58(3):453-467.

Godichon-Baggioni A, Maugis-Rabusseau C, Rau A (2019) Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. J Appl Stat 46(1):47-65.

Greenacre M (2017) 'Size'and 'shape' in the measurement of multivariate proximity. Methods Ecol Evol 8(11):1415-1424.

Greenacre M (2018) Compositional data analysis in practice. Chapman and Hall/CRC press, New York.

Hruzová K, Rypka M, Hron K (2017) Compositional analysis of trade flows structure. Austrian J Stat 46(2):49-63.

Hu N, Zhang T, Gao B, Bose I (2019) What do hotel customers complain about? Text analysis using structural topic model. Tour Manag 72:417-426.Joueid A, Coenders G (2018) Marketing innovation and new product portfolios. A compositional approach. J Open Innov Technol Mark Complex 4:19.

Kogovšek T, Coenders G, Hlebec V (2013) Predictors and outcomes of social network compositions. A compositional structural equation modeling approach. Soc Netw 35(1):1-10.

Kwok L, Yu B (2013) Spreading social media messages on Facebook: an analysis of restaurant business-to-consumer communications. Cornell Hosp Q 54:84-94.

Linares-Mustarós S, Coenders G, Vives-Mestres M (2018) Financial performance and distress profiles. From classification according to financial ratios to compositional classification. Adv Account 40:1-10.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015) Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol 11(3):e1004075.

Marine-Roig E, Ferrer-Rosell B (2018) Measuring the gap between projected and perceived destination images of Catalonia using compositional analysis. Tour Manag 68:236-249.

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (1998) A critical approach to non-parametric classification of compositional data. In: Rizzi A, Vichi M, Bock HH (eds) Advances in data science and classification. Springer, Berlin, p 49-56

Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2015) Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Model 15(2):134-158.

Martín-Fernández JA, Palarea-Albaladejo J, Olea RA (2011) Dealing with zeros. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis. Theory and applications. Wiley, New York, p 47-62.

Martin-Fuentes E (2016) Are guests of the same opinion as the hotel star-rate classification system? J Hosp Tour Manag 29:126-134.

Morais J, Thomas-Agnan C, Simioni M (2018) Using compositional and Dirichlet models for market share regression. J Appl Stat 45(9):1670-1689.

Ortells R, Egozcue JJ, Ortego MI, Garola A (2016) Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indices. In: Martín-Fernández JA, Thió-Henestrosa S (eds) Compositional data analysis. Springer proceedings in mathematics & statistics, Vol. 187. Springer, Cham, p 145-166.

Palarea-Albaladejo J, Martín-Fernández JA (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Comput Geosci 34(8):902-917.

Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemomet Intell Lab 143:85-96.

Di Palma MA, Gallo M (2019) External information model in a compositional perspective: evaluation of Campania adolescents' preferences in the allocation of leisure-time. Soc Indic Res 146(1-2):117-133.

Pawlowsky-Glahn V, Buccianti A (2011) Compositional data analysis. Theory and applications. Wiley, New York.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modelling and analysis of compositional data. Wiley, Chichester.

Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlations which may arise when indices are used in the measurements of organs. Proc R Soc Lond 60:489-498.

Roberts ME, Stewart BM, Airoldi EM (2016) A model of text for experimentation in the social sciences. J Am Stat Assoc 111(515):988-1003.

Russell MA (2014) Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more. O'Reilly, Sebastopol, CA.

Sanz-Sanz E, Martinetti D, Napoleone C (2018) Operational modeling of peri-urban farmland for public action in Mediterranean context. Land Use Policy 75:757-771.

Song H, Seetaram N, Ye S (2019) The effect of tourism taxation on tourists' budget allocation. J Destin Mark Manag 11:32-39.

Thió-Henestrosa S, Martín-Fernández JA (2005) Dealing with compositional data: The freeware CoDaPack. Math Geol 37(7):773-793.

Vives-Mestres M, Martín-Fernández JA, Kenett R (2016) Compositional data methods in customer survey analysis. Qual Reliab Eng Int 32(6):2115-2125.

Voltes-Dorta A, Jiménez JL, Suárez-Alemán A (2014) An initial investigation into the impact of tourism on local budgets: A comparative analysis of Spanish municipalities. Tour Manag 45:124-133.

Yoo KH, Lee W (2017) Facebook marketing by hotel groups: impacts of post content and media type on fan engagement. In: Sigala M, Gretzel U (eds) Advances in social media for travel, tourism and hospitality: new perspectives, practice and cases 2017. Taylor and Francis, London, p 131-146.

Zhou X, Ferrer-Rosell B, Coenders G (2017) Use of social media as e-marketing tool. Comparison of Weibo posts of big and small hotels in China. In: Correia A, Kozak M, Gnoth J, Fyall A (eds) The art of living together. 7th advances tourism marketing conference. CEFAGE - Universidade do Algarve, Faro, p 127-131.

Index