WHEN SIZE DOES NOT MATTER: COMPOSITIONAL DATA ANALYSIS IN MARKETING RESEARCH

Ferrer-Rosell, Berta is a Serra Húnter Fellow employed at the Department of Business Management, University of Lleida, Spain. Her research interests include e-marketing, e- tourism, compositional data analysis and tourist behaviour.

Martín-Fuentes, Eva is employed at the Department of Business Management, University of Lleida, Spain. Her research interests include tourism, marketing, social media, and electronic word of mouth.

Vives-Mestres, Marina is employed at the Department of computer Science, Applied Mathematics and Statistics, University of Girona, Spain. Her research interests include applied statistics in fields such as industrial processes, healthcare, materials and economics.

Coenders, Germà is employed at the Department of Economics, University of Girona, Spain. His research interests include survey research methods, structural equation models and compositional data analysis in management and sociology.

SUMMARY

Many research questions in marketing concern distribution of a whole or relative importance of magnitudes. Compositional Data Analysis focuses on testing relative hypotheses and solves the problems when analysing relative importance data with classical statistical methods. The chapter introduces and illustrates the method with data from an e-WOM platform.

This is a draft chapter. The final version is available in *Handbook of Research Methods for Marketing Management* edited by Robin Nunkoo, Viraiyan Teeroovengadum and Christian M. Ringle, published in 2021, Edward Elgar Publishing Ltd.
https://doi.org/10.4337/9781788976954.00009

INTRODUCTION

Compositional Data Analysis (CoDa) is the standard statistical methodology when data contain relative information or parts of a whole, typically (but not necessarily) with a fixed sum. The seminal work by Aitchison (1986) laid the foundations of CoDa based on the fields of chemistry and geology. Researchers in these fields are interested in the proportion of each component, since absolute amount (size of the sample) is irrelevant (Buccianti et al., 2006).

Nowadays, almost all hard sciences employ CoDa, and it has started to be used and expanded in the fields of marketing, communication and consumer behaviour, which often face similar research questions (Blasco-Duatis et al., 2019; Coenders and Ferrer-Rosell, 2020; Ferrer-Rosell et al., 2015, 2019, 2020; Ferrer-Rosell, Coenders and Martínez-Garcia, 2016; Ferrer-Rosell, Coenders, Mateu-Figueras, et al., 2016; Ferrer-Rosell and Coenders, 2018; Joueid and Coenders, 2018; Marine-Roig and Ferrer-Rosell, 2018; Morais et al., 2018; Vives-Mestres et al., 2016). All those applications show that what is ultimately compositional is the research question (not the data) focusing on relative rather than absolute values.

Typical compositional research questions in the marketing field are related either to the distribution of a whole (e.g., share or allocation) or to the relative importance (e.g., dominance, profile, prevalence). Morais et al. (2018) study the determinants of market share in the automobile industry. Joueid and Coenders (2018) analyse the role of marketing innovation in the relative importance of products with various innovation grades in the firm's portfolio. Marine-Roig and Ferrer-Rosell (2018) quantify the gap between the relative presence of contents in the two sides of tourist destination image (projected vs perceived image). Ferrer-Rosell, Coenders and Martínez-Garcia (2016) and Ferrer-Rosell and Coenders, (2018) segment tourists according to how they allocate their trip budget. Blasco-Duatis et al. (2019) study the relative importance of

contents in e-communication by means of twitter. In what follows we show the rationale for using CoDa in the context of one particular research question regarding on-line customer reviews.

When considering the reviews posted on a customer opinion platform the dominant type of review matters more than the total number of reviews. The dominance is usually computed as the count of each type of review out of the total number of reviews. The analysis of fixed-sum (or constant-sum) data, such as percentages adding to 100, poses considerable statistical challenges, which often are not properly addressed. Most classical statistical methods applied to percentages do not take into account the proportionality or the restricted nature of the data, and may cause serious problems, for example, when interpreting results: one percentage can only increase if one or more percentages decrease.

The chapter is organized as follows. In the next section we briefly introduce compositional data analysis. Then two sections follow on research validity and caveats of the method and the history of its use. Next, a personal experience using CoDa on data from hotel reviews in TripAdvisor and Booking.com is used to illustrate the methodology. The last sections provide our discussion and further reading. The syntax of the analyses is included in the Appendix.

ABOUT COMPOSITIONAL DATA ANALYSIS

What makes compositions special

Compared to absolute data, compositional data, such as count of each review type out of the total reviews, lie in a constrained space. Let $\mathbf{x}$ be a vector with D positive components $\mathbf{x} = (x_1,x_2,\ldots,x_D)$, with $x_j > 0$ for all $j = 1,2,\ldots,D$.

Where in our illustration *D* represents the number of review types (also called components or parts) and x_i the number of reviews per each review type. Since what matters is the percentage of each part, it is common practice to close **x** to a constant sum. In the case of percentages (100) this yields the compositional vector **z** which lies in a bounded space between 0 and 100, known as the simplex:

$$\mathbf{z} = C(\mathbf{x}) = \left(\frac{x_1}{S},\ \frac{x_2}{S},\ \ldots,\ \frac{x_D}{S}\right) \cdot 100 = (z_1,\ z_2,\ \ldots,\ z_D)$$

$$\text{with } z_j > 0 \quad \text{for all } j = 1,2,\ldots,D; \quad \sum_{j=1}^{D} x_j = S; \quad \sum_{j=1}^{D} z_j = 100. \quad (1)$$

The so-called compositional equivalence property ensures that the relative information carried out by the *D* components remains the same regardless of whether the closure is performed (or not), or whether the constant sum is 1 (proportions) or 100. Since z_j are constrained by positiveness and 100 sum, using classical statistical methods will lead to meaningless results to a greater or lesser extent. In other words, one part can only increase if one or more of the others decrease(s), so negative spurious correlations among parts emerge (Pearson, 1897). On the other hand, the Euclidean distance considers the pair of percentages 1 and 2 to be as mutually distant as 11 and 12, while in the first pair the relative difference is far greater than in the second, so that, Euclidean distances among individual compositions are also meaningless. In addition, the common distributional assumptions of most classical models are also somehow violated on **z** (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). That is, modelling with unbounded distributions (e.g. normal) may lead to prediction interval limits outside the [0,100] range.

Exploring a composition

The closed geometric mean is the central tendency measure of a sample of *n* compositions, and it is expressed as $C(g_1,g_2,\ldots,g_D)$, where g_j is the sample geometric mean, of review type z_j for all *n* hotels.

The most common approach to analyse compositional data is to transform the original compositional vector of *D* parts into logarithms of ratios among parts (Aitchison, 1986; Egozcue et al., 2003). Log-ratios present a series of advantages; 1) they are unbounded, 2) tend to meet the distributional assumptions of classical statistical models, 3) carry all needed information about the relative importance of components and 4) are the basis for defining association and distance in a meaningful way. Log-ratios yield the same result when computed from **x** or **z** (Pawlowsky-Glahn et al., 2015). Log-ratios may, for instance, be computed among all possible pairs of parts:

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

The so-called centred log-ratio transformation is computed between each part and the geometric mean of all parts:

$$\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 \ldots z_D}}\right) \qquad (3)$$
$$\text{with } j = 1,2,\ldots,D.$$

Association in CoDa is computed as the proportionality between pairs of components (Lovell et al., 2015). The variance of a pairwise log-ratio is computed as:

$$Var\left(\ln\left(\frac{z_j}{z_k}\right)\right) \qquad (4)$$
$$\text{with } j < k;\ k = 2,3,\ldots,D;\ j = 1,2,\ldots,k-1.$$

Like correlations, variances (Equation 4) can be arranged in a symmetric matrix known as variation matrix were parts define *D* rows and *D* columns. The interpretation of

variances of log-ratios is as follows: hotel review types z_j and z_k behaving perfectly proportionally have a zero variance, and the other way around, the further the variance is from zero, the lower the association between parts.

The CoDa biplot

Compositional data require visualization tools to help researchers interpret large datasets with many components and individuals. Principal component analysis can be applied to compositional data when it is based on the covariance matrix and applied to the centred log-ratios (Equation 3). Results can then be plotted on a CoDa-biplot, whose accuracy is the percentage of explained variance by the two first dimensions.

The CoDa-biplot can be understood as the most accurate representation of the variation matrix in two dimensions. Rays emanating from a common origin represent parts and points represent individual compositions. The interpretation is as follows (see Aitchison and Greenacre, 2002; 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. Ratings that behave proportionally for all *n* (hotels in our example) appear close together.
2. The orthogonal projection of all individual compositions (hotels) in the direction defined by a ray shows an approximate ordering of the importance of that rating for all hotels.

Transformation in balance coordinates and sequential binary partition

Since the *D* centred log-ratios (Equation 3) are perfectly collinear and cannot be used as such in a statistical model, an alternative log-ratio transformation, named balance coordinates was developed in Egozcue and Pawlowsky-Glahn (2005). In fact, a *D*-part

composition subject to a fixed sum constrain is inherently *D*-1 dimensional. Standard statistical models can be applied on balance coordinates.

The appeal of balance coordinates is that they can be investigator-driven, that is, they can be built based on the investigator's research questions. Balance coordinates are formed from a sequential binary partition (SBP) of parts. To create the first balance coordinate, the complete composition is partitioned into two groups of parts: one for the numerator and the other for the denominator. In the following step, one of the two groups is further split into two new groups to create the second balance coordinate. In step k, when the y_k balance is created a group containing r_k+s_k parts is split into two: the r_k parts ($z_{n1},\ldots,z_{nr}$) in the first group are placed in the numerator, and the s_k parts ($z_{d1},\ldots,z_{ds}$) in the second group appear in the denominator. The balance coordinate obtained is a normalised log-ratio of the geometric means of each group of parts:

$$y_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{\sqrt[r_k]{\left(z_{n1} \cdots z_{nr}\right)}}{\sqrt[s_k]{\left(z_{d1} \cdots z_{ds}\right)}}, \text{ with } k = 1, \cdots, D-1. \quad (5)$$

Positive balance coordinates show a higher relative weight of parts in the numerator, while negative values show the opposite.

SBPs and hence balance coordinates are flexible and can be tailored to the particular research question of interest. To this end, SBPs may be constructed according to conceptual similarity of parts, or in order to obtain balance coordinates that express theoretically meaningful comparisons of numerator and denominator parts. The resulting log-ratios can be included in a statistical model as variables.

RESEARCH VALIDITY AND CAVEATS

The most relevant limitation of CoDa is that, strictly speaking, it cannot deal with zero values because log-ratios cannot be computed. In case **x** or **z** contain zeros, they must

be replaced beforehand (Martín-Fernández et al., 2011) and treated depending on the assumed reason of the zeros occurrence. The situation is analogous to a missing data imputation problem under the restriction that, given the original zero values, imputed values cannot be large. The most common reasons for having zero values are; rounding or values below detection limits (continuous-valued compositions), and count compositions (integer compositions). The composition of occurrence of each review type per hotel clearly belongs to the second case.

Zero imputation started in a rather ad-hoc manner (Martín-Fernández et al., 2011), but nowadays rigorous statistically grounded procedures have been developed: the modified EM algorithm (Palarea-Albaladejo and Martín-Fernández, 2008) for the continuous case and the Geometric Bayesian Multiplicative (GBM) approach (Martín-Fernández, Hron, et al., 2015) for the count case. The references in this paragraph 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 around 15% or 20%.

HISTORY OF ITS USE

In his seminal work about CoDa, Aitchison (1982) included an example with geological data and another with economic data. Since then, the number of articles published using CoDa in both hard and social sciences has notably increased. In 1990 we found 14 articles in hard sciences and 1 article in social science citing the first handbook on compositional analysis by John Aitchison in 1986 (search made on Web of Science on 01/06/2018). In 2010 this number increased to 87 and 13 respectively, and from then on in the field of social sciences the number of articles published citing Aitchison (1986) never again decreased below five per year. The applications in fields related to management and business administration are relatively recent, and currently include

relative prices, finance, management education, organizational culture, and accounting, besides, of course, marketing.

Accessible handbooks have also contributed to extending the use of CoDa (Van den Boogaart and Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018; Pawlowsky-Glahn et al., 2015), as has dedicated user-friendly software (r library compositions by Van den Boogaart and Tolosana-Delgado, 2013; r library zCompositions by Palarea-Albaladejo and Martín-Fernández, 2015 and stand-alone program CoDaPack by Thió-Henestrosa and Martín-Fernández, 2005).

PERSONAL EXPERIENCE OF USING CODA IN MARKETING RESEARCH

Data

On April 2016, data was downloaded from TripAdvisor and Booking.com using an automatically controlled webscraper tool, developed in Python, using a library device called Python's Scrapy that simulated user navigation.

Out of the top 100 Euromonitor cities of 2016 (Geerts, 2016) , cities with at least 150 hotels have been selected for the analysis, finally resulting in the following city selection: Bangkok, Dubai, Hong Kong, Istanbul, London, New York, Paris, Rome, Singapore and Taipei. After city selection, hotels with at least 100 reviews have been retained for analysis. The final dataset contains 2,605 hotels: Bangkok 253, Dubai 134, Hong Kong 125, Istanbul 127, London 418, New York 227, Paris 627, Rome 440, Singapore 174 and Taipei 80.

The variables collected from TripAdvisor were: hotel name, address, city, country, global score, hotel category, number of reviews and number of reviews by types or categories (excellent, very good, average, poor, and terrible), and segment posting the review (families, business, solo, friends and couples).

TripAdvisor rating categories “excellent”, “very good”, “average”, “poor” and “terrible” conform a composition, in which each category is a component with its specific count of reviews, that is, out of the total reviews, how many are categorized under each component. In this case, the total number of reviews does not matter, what actually matters is the percentage of reviews included in each hotel rating category. The other TripAdvisor composition to analyse is the travelling group (segment), that is, total reviews distributed to the following components: “families”, “business”, “solo”, “friends” and “couples”. Table 1 shows the centres of both compositions.

Insert Table 1 near here

From Booking.com the variables downloaded were hotel name, address, city, country and also included hotel characteristics: number of rooms (mean 76 rooms), hotel category from 1-2 to 5 stars (1-2: 298, 3: 930, 4: 962 and 5: 415 hotels), hotel price category ordered from 1 to 5, 1 meaning cheap hotel and 5 meaning expensive hotel (1: 204 hotels, 2: 393, 3: 515, 4: 653 and 5: 840 hotels), and whether the hotel is member of a chain or not (yes: 748, not: 1857).

Hotels were merged automatically from two sources (TripAdvisor and Booking.com), being the city the base to merge the hotel (note that there are hotels with the same name in different cities). The same hotel name in the same city was clearly the same hotel if the hotel name on one site is entirely contained in the name on the other site. If none of the previously mentioned points were fulfilled, then we applied the Ratcliff text similarity method (Ratcliff and Metzener, 1988): given two hotel names the algorithm calculates the number of common characters; if hotels had a lower than 85% similarity, then they were discarded.

On an opinion platform, the number of reviews (called volume) may be important because it allows businesses to be more visible and is a sign of popularity (Martin-Fuentes et al., 2020). However, the valence (e.g. positive, negative, or neutral review) bring users with even more information of the product and service and influence other consumers; positive reviews may encourage other consumers to book and negative reviews may discourage them (Dellarocas, 2003). Moreover, the category of the review allows researchers to calculate the real score of a hotel on TripAdvisor.

Two different analyses have been carried out to show how to apply the most common CoDa tools: linear models and visualization. The first is a multivariate analysis of covariance (MANCOVA), in order to answer the following research question: are hotel characteristics affecting hotel reviews composition ("terrible", "poor", "average", "very good" and "excellent")? The second analysis uses the CoDa-biplot, creating a visual representation of the two TripAdvisor compositions: type of reviews by hotel rating and type of reviews by travelling segment with the aim of observing associations between them.

Analyses have been carried out with R software, version 3.4.2.

Linear model

Balance coordinates have been computed after replacing the 0.69% of count zeros in the hotel rating composition. Figure 1 shows visual representation of the SBP as a tree diagram. Drawing from the "excellent" component, the composition is split into two groups, "excellent" versus the other four components. In the following step, the group with the other four components is again split into two groups resulting in the component "very good" versus the other three components. The last partition splits the component "poor" versus "terrible". The balance coordinates are:

$$y_1 = \sqrt{\frac{1\cdot 4}{1+4}} \ln \frac{x_5}{\sqrt[4]{x_4 x_3 x_2 x_1}}$$

$$y_2 = \sqrt{\frac{1\cdot 3}{1+3}} \ln \frac{x_4}{\sqrt[3]{x_3 x_2 x_1}}$$

$$y_3 = \sqrt{\frac{1\cdot 2}{1+2}} \ln \frac{x_3}{\sqrt{x_2 x_1}} \qquad (6)$$

$$y_4 = \sqrt{\frac{1\cdot 1}{1+1}} \ln \frac{x_2}{x_1}$$

Insert Figure 1 near here

Balance coordinates must be introduced in a multivariate model as dependent, and since we have continuous and categorical predictors, a MANCOVA is performed. Global model tests such as Pillai's trace are invariant to how the balance coordinates are constructed (Martín-Fernández, Daunis i Estadella, et al., 2015). However, individual tests referring to each balance coordinates are not invariant and are interpreted according to the interpretability of each balance.

Table 2 shows the estimates and individual tests for each predictor. As regards the global model tests, all five predictors' Pillai's Traces are significant (p-values< 0.001).

Insert Table 2 near here

When interpreting the model, the compositional nature of data has to be taken into account by considering which component increases when others decrease. The positive effect on y_1 show that hotels located in the cities with statistically significant estimates (Rome, London, Paris, Taipei, Dubai and New York) are associated with an increase in the frequency of the "excellent" rating at the expense of a decrease in the

geometric mean of all other ratings with respect to Bangkok hotels. On the other hand, the negative effects of Rome, London and New York on y_3 show that hotels located in these three cities are associated with a decrease in the frequency of the "average" rating at the expense of an increase in the geometric mean of "poor" and "terrible" ratings.

Being member of a chain also affects y_1, that is, the positive effect shows that hotels members of a chain tend to have a higher frequency of "excellent" ratings at expense of all other rating categories. Regarding price, all estimates are significant, meaning that price category is associated with ratings as confirmed by Martin-Fuentes (2016). In fact, the higher the price category the higher the estimate, and thus, as the price category increases it also increases the frequency of the "excellent" rating, at the expense of the geometric mean of the other ratings. The negative effect of number of rooms on y_1, shows that hotels with more rooms tend to have a higher geometric mean of all four ratings except for "excellent", at the expense of "excellent" ratings. Regarding y_4 and cities, only the effect of hotels located in Dubai is significant. This negative effect shows that being located in Dubai is associated with an increase in the frequency of "poor" rating at expense of a decrease in the frequency of "terrible", when comparing this city with Bangkok.

CoDa biplot

Biplots can be extended to the case in which there is more than one composition (Filzmoser et al., 2018; Kynčlová et al., 2016). To this end the centred log-ratios (Equation 3) of both compositions are combined into a single dataset, which is submitted to a principal component analysis based on the covariance matrix. Both hotel rating and travel group compositions are represented together in Figure 2. The percentage of explained variance by the two first dimensions is 72.9%.

Distances between the vertices of the rays of two parts of the same composition are interpreted as an approximation of the standard deviation of their corresponding pairwise log-ratio. Parts that behave proportionally for all individuals appear close together. This is the case for “poor” and “terrible” reviews; or reviews written by families and couples. The cosine of the angle between two rays corresponding to different compositions is approximately equal to the correlation between their two centred log-ratios. For instance, when the frequency of reviews written by solo travellers increases in relative terms the frequency of “average”, “poor” and “terrible” reviews also increases in relative terms.

Hotels on the right side of the biplot have a comparatively high proportion of “excellent” and “very good” reviews and a comparatively low proportion of “poor” and “terrible” reviews. Hotels on the top of the biplot have a comparatively high proportion of reviews written by business travellers, and a comparatively low proportion of reviews written by families, couples and friends.

As in standard biplots, hotel points could be labelled according to the values of an external variable, or their scores on the first two dimensions could be used as data in subsequent statistical analyses.

Insert Figure 2 near here

DISCUSSION

This chapter constitutes the first overview of CoDa in the marketing management field. We have strived to keep to the point and write an easy-to-follow invitation to use the method which can be complemented in the annotated further reading section.

A common statistical saying goes that one should not “fit square pegs into round holes”: compositional questions have to be answered with compositional methods. Aitchison pointed out in 1997 that “compositional data analysis is simple”, and we agree with that. CoDa is the simple way out for researchers who want to be on the safe side when interpreting their results by taking into account the relative information of interest. An added appeal of the CoDa methods we have presented in this chapter is that, once the data have been transformed, researchers can use standard and well understood statistical tools.

Given the fact that many research questions and measures in marketing are relative rather than absolute, it can be argued that CoDa methodology has not yet reached its full usage potential. To give just a few suggestions, CoDa can be used on a general basis to analyse survey questionnaires with the common response format “Please divide 100 points among the following product attributes according to the importance they have to you”, which is quite common in product development and market segmentation. Moreover, audience research is focused on the concept of share, which is compositional by nature. Likewise, consumer time use adds up to 24 hours a day and can only be studied in relative terms, to give just a few suggestions.

Analysing the relative importance of reviews by type is important in marketing because it has been demonstrated that the valence is correlated with the expectations created by the customers about a product or a service and its intention to purchase (Mauri and Minazzi, 2013) and influences other consumers’ decisions (Vermeulen and Seegers, 2009). Moreover, the relative importance of the reviews according to the traveller group that has posted them is a source of segmentation that allows for better positioning (Martin-Fuentes et al., 2018).

Ultimately CoDa can be traced back to the old distinction between “size” and “shape” in morphometrics and multivariate statistics. It goes without saying that CoDa focuses on shape, as if size would not matter. This is not always true, and more advanced CoDa methods, which are beyond the scope of this chapter, consider both shape and size.

REFERENCES

Aitchison, J. (1982), “The statistical analysis of compositional data”, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 139–177.

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Aitchison, J. and Greenacre, M. (2002), “Biplots of compositional data”, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 51 No. 4, pp. 375–392.

Blasco-Duatis, M., Coenders, G., Sáez, M., Fernández García, N. and Cunha, I. (2019), “Mapping the agenda-setting theory, priming and the spiral of silence in twitter accounts of political parties”, *International Journal of Web Based Communities*, Vol. 15 No. 1, pp. 4–24.

Coenders, G. and Ferrer-Rosell, B. (2020), “Compositional data analysis in tourism. Review and future directions”, *Tourism Analysis*, Vol. 25, No. 1, pp. 153–168.

Van den Boogaart, K.G. and Tolosana-Delgado, R. (2013), *Analyzing Compositional Data with R*, Vol. 122, Springer, Berlin.

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

Dellarocas, C. (2003), “The digitization of word of mouth: Promise and challenges of online feedback mechanisms”, *Management Science*, Vol. 49 No. 10, pp. 1407–1424.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2005), “Groups of parts and their balances in compositional data analysis”, *Mathematical Geology*, Vol. 37 No. 7, pp. 795–282.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003), "Isometric logratio transformations for compositional data analysis", *Mathematical Geology*, Vol. 35 No. 3, pp. 279–300.

Ferrer-Rosell, B. and Coenders, G. (2018), "Destinations and crisis. Profiling tourists' budget share from 2006 to 2012", *Journal of Destination Marketing and Management*, Vol. 7, pp. 26–35.

Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2015), "Determinants in tourist expenditure composition - The role of airline types", *Tourism Economics*, Vol. 21 No. 1, pp. 9–32.

Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2016), "Segmentation by tourist expenditure composition: An approach with compositional data analysis and latent classes", *Tourism Analysis*, Vol. 21 No. 6, pp. 589–602.

Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2016), "Understanding low-cost airline users' expenditure patterns and volume", *Tourism Economics*, Vol. 22 No. 2, pp. 269–291.

Ferrer-Rosell, B., Martin-Fuentes, E. and Marine-Roig, E. (2019), "Do Hotels Talk on Facebook About Themselves or About Their Destinations?", *Information and Communication Technologies in Tourism 2019*, pp. 344–356, Springer, Cham.

Ferrer-Rosell, B., Martin-Fuentes, E. and Marine-Roig, E. (2020), "Diverse and emotional: Facebook content strategy by Spanish hotels", *Journal of Information Technology and Tourism*, Vol. 22 No. 1, pp. 53-74.

Filzmoser, P., Hron, K. and Templ, M. (2018), *Applied Compositional Data Analysis: With Worked Examples in R*, Springer, Cham.

Geerts, W. (2016), *Top 100 City Destinations Ranking*, available at: http://blog.euromonitor.com/2016/01/top-100-city-destinations-ranking-2016.html.

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

Joueid, A. and Coenders, G. (2018), "Marketing innovation and new product portfolios.

A compositional approach", *Journal of Open Innovation: Technology, Market, and Complexity*, Vol. 4 No. 2, p. 19.

Kynčlová, P., Filzmoser, P. and Hron, K. (2016), "Compositional biplots including external non-compositional variables", *Statistics*, Vol. 50 No. 5, pp. 1132–1148.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S. and Bähler, J. (2015), "Proportionality: a valid alternative to correlation for relative data", *PLoS Computational Biology*, Vol. 11 No. 3, p. e1004075.

Marine-Roig, E. and Ferrer-Rosell, B. (2018), "Measuring the gap between projected and perceived destination images of Catalonia using compositional analysis", *Tourism Management*, Vol. 68, pp. 236–249.

Martín-Fernández, J.-A., Palarea-Albaladejo, J. and Olea, R.A. (2011), "Dealing with zeros", in Pawlowsky-Glahn, V. and Buccianti, A. (Eds.), *Compositional Data Analysis. Theory and Applications*, pp. 47–62, Wiley, New York.

Martín-Fernández, J.A., Daunis i Estadella, J. and Mateu i Figueras, G. (2015), "On the interpretation of differences between groups for compositional data", *SORT: Statistics and Operations Research Transactions,* Vol. 39 No. 2, pp. 231-252.

Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P. and Palarea-Albaladejo, J. (2015), "Bayesian-multiplicative treatment of count zeros in compositional data sets", *Statistical Modelling*, Vol. 15, pp. 134–158.

Martin-Fuentes, E. (2016), "Are guests of the same opinion as the hotel star-rate classification system?", *Journal of Hospitality and Tourism Management*, Vol. 29, pp. 126–134.

Martin-Fuentes, E., Mateu, C. and Fernandez, C. (2018), "Does verifying users influence rankings? Analyzing Booking.com and Tripadvisor", *Tourism Analysis*, Vol. 23 No. 1, pp. 1–15.

Martin-Fuentes, E., Mateu, C. and Fernandez, C. (2020), "The more the merrier? Number of reviews versus score on TripAdvisor and Booking.com", *International Journal of Hospitality & Tourism Administration*, Vol. 21, No. 1. pp. 1-14.

Mauri, A.G. and Minazzi, R. (2013), “Web reviews influence on expectations and purchasing intentions of hotel potential customers”, *International Journal of Hospitality Management*, Vol. 34, pp. 99–107.

Morais, J., Thomas-Agnan, C. and Simioni, M. (2018), "Using compositional and Dirichlet models for market share regression", *Journal of Applied Statistics*, Vol. 45 No. 9, pp. 1670–1689.

Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2008), “A modified EM alr-algorithm for replacing rounded zeros in compositional data sets”, *Computers & Geosciences*, Vol. 34 No. 8, pp. 902–917.

Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2015), “zCompositions—R package for multivariate imputation of left-censored data under a compositional approach”, *Chemometrics and Intelligent Laboratory Systems*, Vol. 143, pp. 85–96.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015), *Modeling and Analysis of Compositional Data*, Wiley, Chichester.

Pearson, K. (1897), “Mathematical contributions to the theory of evolution - on a form of spurious correlation which may arise when indices are used in the measurement of organs”, *Proceedings of the Royal Society of London*, Vol. 60 No. 359–367, pp. 489–498.

Ratcliff, J.W. and Metzener, D.E. (1988), “Pattern-matching-the gestalt approach”, *Dr Dobbs Journal*, Vol. 13 No. 7, pp. 46.

Thió-Henestrosa, S. and Martín-Fernández, J.A. (2005), “Dealing with compositional data: the freeware CoDaPack”, *Mathematical Geology*, Vol. 37 No. 7, pp. 773–793.

Vermeulen, I.E. and Seegers, D. (2009), “Tried and tested: The impact of online hotel reviews on consumer consideration”, *Tourism Management*, Vol. 30 No. 1, pp. 123–127.

Vives-Mestres, M., Martín-Fernández, J.-A. and Kenett, R.S. (2016), “Compositional data methods in customer survey analysis”, *Quality and Reliability Engineering*

International, Vol. 32 No. 6, pp. 2115–2125.

ANNOTATED FURTHER READING

Van den Boogaart and Tolosana-Delgado (2013) is the first guide to CoDa with R, using a library developed by the authors. The book contains many step-by-step examples from different fields and covers most standard statistical methods.

Pawlowsky-Glahn et al. (2015) is a comprehensive guide to the CoDa methodology. It does not follow any software in particular but rather focuses on principles, interpretation and threats to the validity of conclusions.

Filzmoser et al. (2018) is another guide to CoDa with R, using a library developed by the authors and including some novel features, among which we highlight missing data, high-dimensional compositions, robust methods and two-way compositions.

APPENDIX

```
library(compositions)
library(zCompositions)
#data frame "data" contains the rating category components
#in columns 2 to 6.
#travel group components are columns 7 to 11
#computing percentage of zero values
zPatterns(data[,2:6],label=0)
zPatterns(data[,7:11],label=0)
#imputation by GBM method
comp_rating<-acomp(cmultRepl(data[,2:6]))
comp_group<-acomp(cmultRepl(data[,7:11]))
#center
mean(comp_rating)
mean(comp_group)
#linear model
#Defining an SBP for balance coordinates
#(1: numerator, -1: denominator, 0 not present)
W<-matrix(c(
  1,-1,-1,-1,-1,
  0,1,-1,-1,-1,
  0,0,1,-1,-1,
  0,0,0,1,-1), nrow=5)
rownames(W)<-c("excellent","very good","average",
"poor", "terrible")
colnames(W)<-c("y1","y2","y3","y4")
W
#balance computation
comp_ratingbalances<-ilr(comp_rating,gsi.buildilrBase(W))
#estimation
model=lm(comp_ratingbalances~factor(city)+factor(chain)
+factor(price)+factor(stars)+rooms, data=data)
summary(model)
anova(model)
#biplot
#merging centred log ratios of both compositions
comps2<-cbind(clr(comp_rating),clr(comp_group))
#principal component analysis
pcx=princomp(comps2)
#biplot data point labels
data$point="."
biplot(pcx, xlabs=data$point)
```

TABLES

Table 1. Geometric mean (centres) of components per types of Reviews and segments

Composition hotel rating					
Excellent	Very good	Average	Poor	Terrible	Total
33.1	42.3	16.2	5.1	3.3	100
Composition group travelling					
Families	Solo	Friends	Business	Couples	
20.2	8.9	13.8	15.0	42.1	100

Table 2. Parameter estimates and individual tests for each predictor

Variable	Category	y_1	y_2	y_3	y_4
Intercept		-1.0653***	0.4857***	0.7554***	-0.0191
City	Singapore	-0.1198	-0.1361*	-0.0050	0.0132
	Bangkok	0	0	0	0
	Istanbul	0.0523	-0.0793	0.0171	-0.0964
	Hong Kong	0.0876	0.1633*	0.1378**	0.0105
	Rome	0.1820*	-0.1281*	-0.1293***	-0.0255
	London	0.3017***	-0.1523**	-0.1990***	-0.0407
	Paris	0.4620***	0.1412**	-0.0441	0.0307
	Taipei	0.5984***	0.6246***	0.4166***	0.0561
	Dubai	0.7680***	0.2530***	0.0248	-0.1154*
	New York	0.9939***	0.2184***	-0.1013*	-0.0528
Chain membership	No	0	0	0	0
	Yes	0.2833***	0.1442***	0.0617**	0.0612**
Price	1	0	0	0	0
	2	0.7906***	0.6777***	0.3497***	0.2281***
	3	1.2799***	1.0049***	0.4797***	0.2995***
	4	1.9116***	1.3888***	0.5974***	0.4122***
	5	2.5133***	1.6380***	0.6251***	0.3931***
Stars	1-2	0	0	0	0
	3	0.1089	0.0724	0.0139	0.0133
	4	0.0756	-0.1070	-0.1365***	-0.0079
	5	0.5660***	-0.2549***	-0.3015***	-0.0340
Nº of rooms		-0.0009***	-0.0004***	0.0000	0.0000
F-tests		<0.001	<0.001	<0.001	<0.001
R-Squared		0.4707	0.3164	0.1536	0.0544

P-value: <0.001***; 0.001**; 0.01*

FIGURES

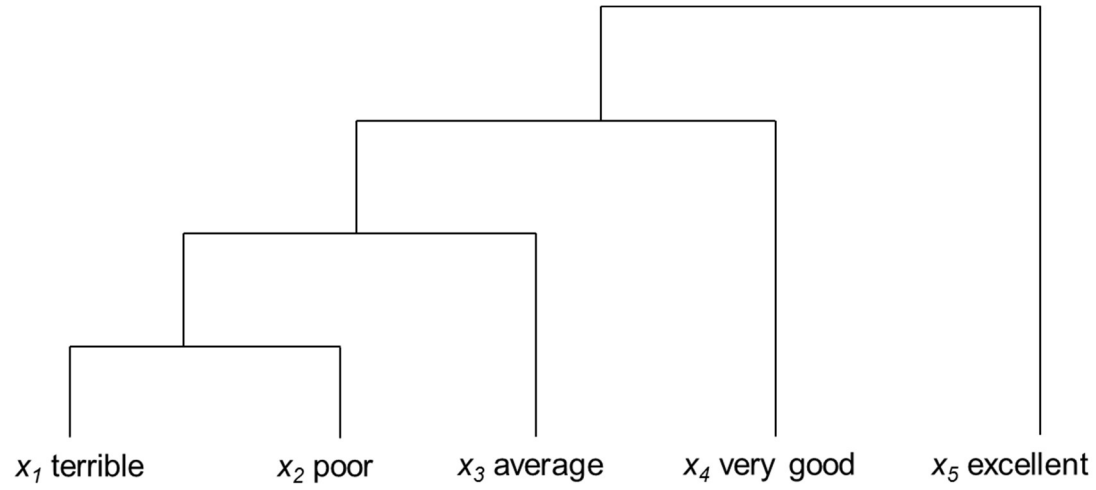

Figure 1.

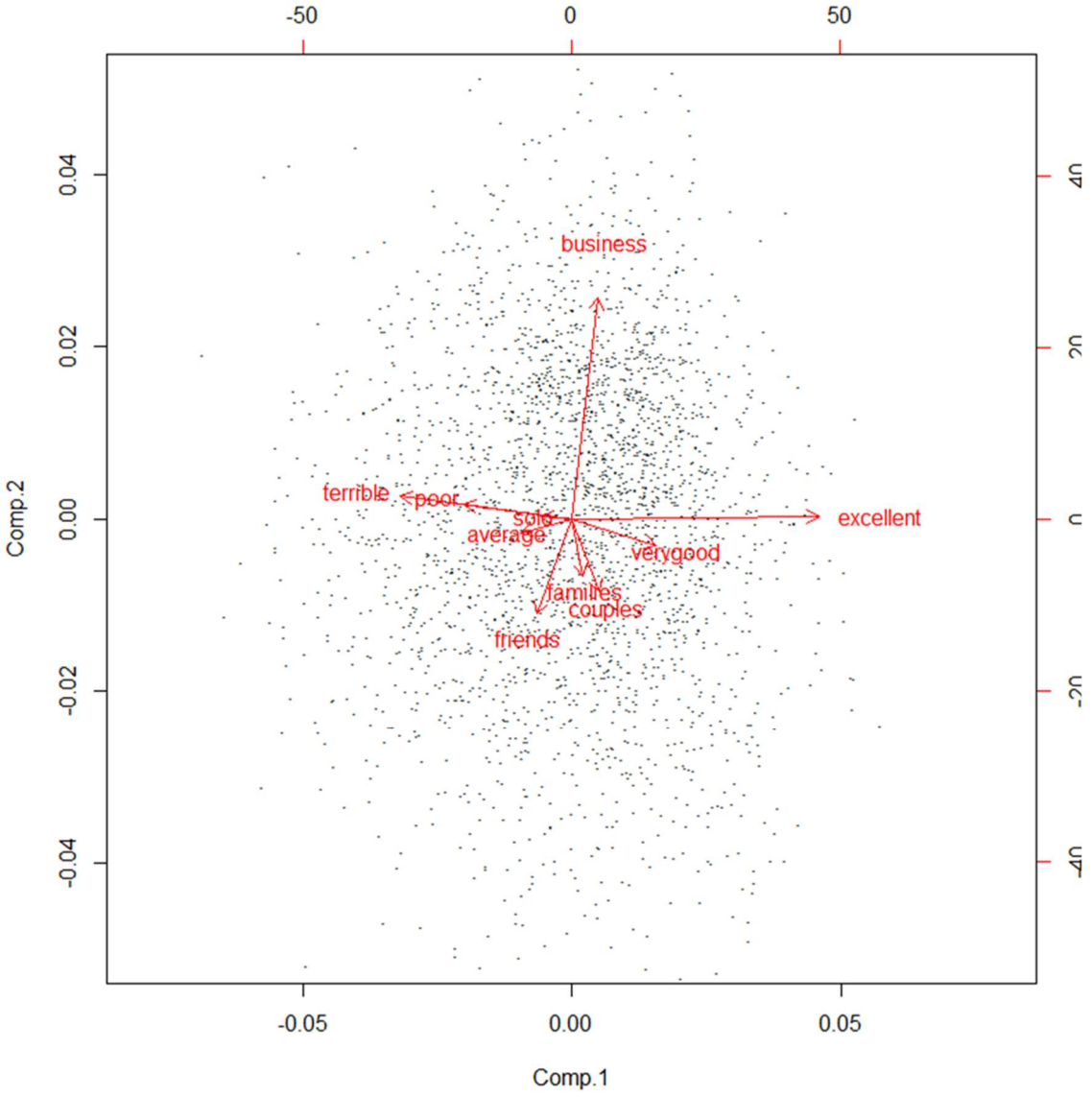

Figure 2.