39. COMPOSITIONAL ANALYSIS OF TOURISM-RELATED DATA – Contributions by Berta Ferrer-Rosell

Approaching Compositional Data Analysis in Tourism

Compositional data analysis is the appropriate methodology to employ when dealing with data carrying relative information. Compositional data (CoDa) can be defined as arrays of positive numbers – called components, or parts – whose relative size is of interest to the researchers. In some cases, the components are parts of a whole and their sum is irrelevant or even constant. This is the case in geological and chemical analyses, which use proportions adding up to 1, or in time-use studies, where the sum is up to 24 hours. These two fields of application are the first and most traditional fields using CoDa (Aitchison, 1986).

It is taken for granted that the total time of a day is an uninteresting fact. A common time-use research interest is to know how the distribution of daily time of an individual – the time allocated to commuting, to work, to family and household tasks, to sleeping – may affect one's health, quality of life and life satisfaction. In other cases, components do not constitute any whole or do not have a constant sum, and the only key issue is that the researchers' interest and questions lie in the relative importance of components to one another (Egozcue & Pawlowsky-Glahn, 2019). This might be the case of the content uploaded on a tourism product, company or destination website or on a printed brochure. Larger websites or wider brochures may have more content of all types. There are, however, other cases in which both the relative importance and the total volume are of interest. This is the case with tourist expenditure. Destinations and tourism companies may like to know how travellers allocate their trip budget into different expenses – transportation, accommodation and food, activities, and others – as well as how much tourists spend during the trip as a whole.

As can be deduced, many research questions involving tourism-related data to analyse consumer (or company or destination) behaviour, are related to distribution of a whole (e.g., share or allocation) or to relative importance (e.g., prevalence, concentration, dominance). Possible questions tourism researchers may be interested in are, for instance, how does the relative popularity of search terms in Google relate to tourism market share? How do hospitality firms allocate their capacity to their product portfolio? How does time allocated to different types of activities at the destination relate to tourist satisfaction?

CoDa has started to be used in several fields of social science which often face similar research questions, such as education (Batista-Foguet, Ferrer-Rosell, Serlavós, Coenders & Boyatzis, 2015), finance (Carreras-Simó & Coenders, 2020; Linares-Mustarós, Coenders & Vives-Mestres, 2018), marketing (Morais, Thomas-Agnan & Simioni, 2017; Vives-Mestres, Martín-Fernández & Kenett, 2016), sociology (Hlebec, Kogovšek & Coenders, 2012), communication (Blasco-Duatis, Saez Zafra & Garcia Fernandez, 2018; Huertas, Ferrer-Rosell, Marine-Roig & Cristobal-Fransi, 2021), urban studies (Cruz-Sandoval, Ortego & Roca, 2020) and sustainability (Marcillo-Delgado, Ortego & Pérez-Foguet, 2019). In the field of tourism and hospitality, I have been the pioneer in the introduction of compositional data analysis. It has not been an easy task but being constant and persevering has really helped me to find a space in tourism literature.

Data carrying relative information (proportions) have characteristics which render most statistical workhorses (e.g. mean, correlation and distance) meaningless to a greater or lesser extent when applied to them. That is, Euclidean distance considers the pair of percentages 1% and 2% to be as mutually distant as 11% to 12%, while in the first pair the proportional difference is 100% and in the second it is less than 10%. Compositional data lies in a constrained space restricted to positiveness and sometimes to unit (or

100) sum, so that negative spurious correlations among the parts emerge (Pearson, 1897). The statistical and distributional assumptions of most classical statistical models are violated to a certain extent when applied to proportions (Pawlowsky-Glahn, Egozcue & Tolosana-Delgado, 2015). And statistical modelling with unbounded distributions (normal) is not feasible, as it results in confidence prediction intervals outside the [0,1] range. In other words, most standard and classical statistical methods applied to data carrying relative information do not consider the restricted nature of the data or the proportionality and may cause serious problems when the results are interpreted.

The most common approach to dealing with compositional data consists of transforming the original compositional vector of *D* parts into logarithms of ratios between parts (Aitchision, 1986; Egozcue, Pawlowsky-Glahn, Mateu-Figueras & Barcelo-Vidal, 2003). Log-ratios have several advantages. They recover the unlimited space (from – infinite to + infinite), tend to fulfil the assumption of classical statistical models, constitute the natural way of distilling the relative information among parts, and form the basis for defining the association and the distance between parts in a meaningful way. In the CoDa literature, there are different ways of transforming the data. The simplest to compute and most popular transformation is the additive log-ratio transformation (alr), used by Aitchison (1982). This is simply the log-ratio of each component to the last. The centred log-ratio (clr) computes the log-ratio of each component over the geometric mean of all other components (Aitchison, 1983). Finally, the third log-ratio transformation is the most flexible because the denominator does not need to be the same in all ratios. This flexibility makes it easier to compute log-ratios which are interpretable with respect to research questions. This is the orthonormal log-ratio (olr) – also named isometric log-ratio (ilr). It draws from a sequential binary partition of parts to build coordinates comparing the relative importance of groups of parts to each other (Egozcue & Pawlowksy-Glahn, 2005). The appeal of CoDa is that once the data have been transformed, any kind of standard and well-understood statistical technique can be used with unbounded error-term distributions, while making sure results will be compositionally coherent. Log-ratios can be the explanatory or the dependent variables in any statistical model (see Coenders, Martín-Fernández & Ferre-Rosell, 2017 for examples and interpretation of results).

The most common compositional data techniques (centre, distance, association, biplots) and how to deal with zeros present in the data are explained in the several contributions commented on in the following sections, as well as in Coenders and Ferrer-Rosell (2020).

Analysis of tourist expenditure

Tourist expenditure analysis has been a recurrent topic in tourism literature, since it is a major concern for destinations, companies and marketers, and in general to the tourism industry. Tourist expenditure – rather than the number of tourists received – is becoming much more relevant for destinations and for the economic impact of tourism. Destinations are naturally interested in the local spending (expenses incurred at the destination) more than origin spending (expenses paid directly to tour operators at origin, for instance). Analysing how tourists spend their budget – that is, the analysis of expenditure composition – provides valuable information for destinations to make decisions regarding which market to target. If destination marketers seek to promote activities, they should focus their marketing efforts on those markets where tourists spend more of their trip budget.

Ferrer-Rosell, Coenders and Martínez-Garcia (2015) studied the drivers of the share of tourist expenditure allocated to the various categories of travel budget (transportation, accommodation and food, and activity expenses). It was the first publication to consider the trip budget as a composition, and thus, used CoDa. Isometric log-ratios of budget share were fitted to a MANOVA, with travellers' attributes as explanatory factors. The type of airline company used by individuals (low-cost or full-service) was used as a moderator variable to observe how travelling with one or the other type affected the distribution of trip budget expenses. As already mentioned, once expense variables (components or parts of the trip budget) have

been transformed into logarithms of ratios, any statistical technique with unbounded error-term distribution can be used, while ensuring results will be compositionally coherent and that the standard statistical assumptions hold. Actually, it is not advisable to carry out, for instance, a MANOVA model on raw expenditure share. There is a high risk of resulting in prediction intervals outside the [0,1] range (Ferrer-Rosell et al., 2015).

Another relevant contribution in this research line is that co-authored by Ferrer-Rosell, Coenders, Mateu-Figueras, and Pawlowsky-Glahn (2016) which represented the application of the *compositional data analysis with total* method development by Pawlowsky-Glahn, Egozcue and Lovell (2015). This study showed that the analysis of the absolute trip spending by parts (such as transportation), and of trip budget share (the percentage of transportation within the total trip budget) served different research objectives. The first type of analysis refers to how much tourists spend, while the second refers to how they spend the allocation (as was the case in Ferrer-Rosell et al., 2015). The study provided a new methodological tool to analyse the determinants of the tourist expenditure combining the analysis of budget share and absolute expenditure in the same model. It is worth noting that initially CoDa methodology was criticised for ignoring the total (volume) when it was available and of interest. The study drawn from Pawlowsky-Glahn, Egozcue & Lovell (2015) shows an alternative and flexible way to include the absolute expenditure in the analysis. It further indicates that it can be tailored to the research questions at hand, focusing on absolute expenditure on transportation, or expenditure made at the destination, for instance.

The *compositional data analysis with total* approach was also used in Ferrer-Rosell and Coenders (2017) to observe whether airline types (low-cost and full-service) had been converging regarding travellers' expenditure allocation and total trip expenditure. Repeated cross-sections were used and the aim was not to confound effects involving expenditure distribution with those involving expenditure volume. Users of both airline types converged in their allocation of the trip budget but diverged with regard to the total trip expenditure.

In Ferrer-Rosell, Coenders and Martínez-Garcia (2016) latent class modelling was used jointly with CoDa to segment tourists according to trip budget share, that is, according to proportions of total expenditure allocated to different expense concepts (transportation, accommodation and activities). Another study segmenting tourists based on how they distribute their trip budget is that of Ferrer-Rosell and Coenders (2018). Segmentations based on absolute expenditure and those based on share expenditure respond to different research interests, and in managerial terms, also serve different, but complementary purposes. In this sense, both studies contributed in methodological, theoretical and managerial terms. With results at hand, destinations might be able to know the segments to tackle, namely the segments that would bring more benefits to the destination.

Analysis of e-tourism content

Content analysis has been a frequently used method in tourism literature for quite a long time now (Camprubí & Coromina, 2016). In the same vein, how and what tourism stakeholders (e.g., destinations, companies) communicate online or offline is also of interest for researchers as well as for tourism sector agents to know the impact of marketing campaigns, for instance. In tourism communication, in the context of studies of destination image, branding or marketing, it might be of interest to depict the manner in which destination marketers emphasize certain content over others on a brochure, on the website or on social media. Regarding content dominance, when considering reviews posted on a customer opinion platform, or when considering posts published on a social media platform, the dominant type of review or post matters more than the total number of reviews or posts. The dominance is usually computed as the count of each type of review or post out of the total number of reviews or posts. Thus, data can be the count of a phenomenon, whose sum for an individual *i* is *Si*. For instance, a tourist company's total count of *Si* social-media posts can be classified into *D* content categories (parts of a composition).

One of the most relevant contributions in this research line, and the first one, is the publication by Marine-Roig and Ferrer-Rosell (2018), in which the (in)congruity (or gap) between projected and perceived tourist destination images was measured using compositional distance between proportions. An outstanding Mediterranean destination, Catalonia, was analysed based on three different information sources: induced (Catalan Tourist Board dossier), autonomous (Lonely Planet travel guide), and organic (UGC: user-generated content). UGC consisted of a random sample of 80,000 online travel reviews written in English by tourists who visited Catalonia during 2015. The common approach to differentiate the three information sources would be to compute differences between proportions directly, that is, by subtracting percentages of appearance of content in each source. However, this does not make sense because when considering proportionality of data, Euclidean distance does not take into account proportionality. Direct subtractions between proportions are not precise and are confusing. Using compositional distance (defined by Aitchison, 1983) based on clr transformation is the appropriate way to operate when dealing with proportions.

Compositional distance was also used in Lalicic, Marine-Roig, Ferrer-Rosell and Martin-Fuentes (2021) to observe differences between how guests of Airbnb perceive four main urban destinations in Spain. In relative terms, Airbnb reviews from Barcelona had more content about sports, while reviews from Madrid presented more content about the urban environment. In Seville, reviews contained relatively more content about food and wine, and in Valencia there was more about leisure and recreation. Considering all eight destination-image categories, Barcelona and Madrid are perceived as similar destinations, while Barcelona and Seville are perceived very differently.

Moving to tourism companies' communication, in Ferrer-Rosell, Marine-Roig and Martin-Fuentes (2020), the content posted on Facebook was analysed. The aim was to unveil the content strategy of the two types of content (hotel information vs. destination information) for hotels located in the two most visited cities in Spain, Barcelona and Madrid. In this case, 5,900 Facebook posts were categorized into hotel-related information and destination-related information. The composition of internal information included the content categories (components) of rooms, restaurants and other facilities. The composition of external information included heritage, urban, nature and sports, and gastronomy. The isometric log-ratios were fitted to a t-test (to compare hotels from Barcelona and Madrid) and results showed that hotels from Barcelona published more posts on Facebook about themselves (hotel services), while hotels from Madrid posted more content related to the destination.

Another line of research within the tourism-content analysis and tourism communication is presented in Huertas, Ferrer-Rosell, Marine-Roig and Cristobal-Fransi (2021). The aim of this research was to analyse the treatment in the Spanish press of controversial issues regarding Airbnb and its evolution. CoDa allowed us to get what were the most dominant topics in all media sources and to observe the evolution of topics along the timeline of 2016 through 2018. The results showed that topics treated mainly from a negative perspective evolved towards a more positive vision.

Concluding and personal remarks

My tourism compositional journey began in 2012, while I was writing my PhD thesis at the University of Girona, thanks to my supervisor Germà Coenders. The University of Girona is the home of a widely recognized research group in compositional data analysis, and since I was analysing tourist trip budgets and considering it compositional, my supervisor pushed me to take the "Week CoDa Course" held at the University of Girona in summer 2012.

Thus, the thesis entitled "Tourism Demand in Spain: Trip Duration and Budget Structure – A Comparison of Low Cost and Legacy Airline Users" represented the first time CoDa was applied to the tourism field to analyse a traditional and widely studied topic, tourism expenditure. Now, a bit less than 10 years later, I am a councillor on the CoDa-Association board.

I realised that the research questions outlined regarding the analysis of tourism expenditure were leading me to consider the compositional nature of the data I had. Thus, I started to see the world of tourism from another perspective. We can find compositional data everywhere: aggregated data of origin–destination flows are also of compositional nature, as well as the time tourists dedicate to different activities at the destination. Moreover, tourism firms allocate their capacity and resources to their products and services, and the financial ratios of tourism companies are also compositional, as their interest also lies in the relative size of accounting. How a restaurant menu is designed (having more fish plates or meat or vegetables), depending for instance on the location (e.g., seafront, mountain destination), is also compositional. All kinds of content (videos, pictures, text) posted on social media (e.g., Instagram) or the total minutes a film dedicates to showing the destination where it was filmed (e.g., Woody Allen's *Vicky Cristina Barcelona*) are susceptible to being classified into predefined content categories.

Jointly with Germà Coenders, I have recently published an article entitled "Compositional data analysis in tourism. Review and future directions" (Coenders & Ferrer-Rosell, 2020). This is a review of CoDa methodology used in fields which could be easily transferred to tourism (e.g., economy, marketing). It also presents future directions with several ideas tourism researchers might be interested in analysing. I was also invited to contribute to the *Encyclopedia of Tourism Management and Marketing* with a methodological chapter on "Compositional Data Analysis in Tourism".

Apart from the publications presented in this chapter, I have other publications (e.g., book chapters) using CoDa, and it is also worth mentioning I have been invited to talk about CoDa in tourism at several conferences (e.g., ENBIS 2015, COMPSTAT 2018), at the European IFITT Masterclass on e-Tourism 2019 and at the CoDa-Day 2021. I have also lectured several seminars presenting research featuring CoDa at national and international universities (e.g., Virginia Tech), I am an instructor at the CoDa course powered by the University of Girona, and I have created a 10-hour CoDa course in social sciences with a special focus on research applications and where the free software CoDaPack is used to practice with real data.

Acknowledgements

I am deeply grateful to my past and present (and future) colleagues for travelling with me in this journey.

Written by Berta Ferrer-Rosell, University of Lleida, Spain
Read Berta's letter to future generations of tourism researchers

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44*(2), 139-160.

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*. London, UK, Chapman and Hall.

Batista-Foguet, J.M., Ferrer-Rosell, B., Serlavós, R., Coenders, G., & Boyatzis, R.E. (2015). An Alternative Approach to Analyze Ipsative Data. Revisiting Experiential Learning Theory. *Frontiers in Psychology, 6*, 1742.

Blasco-Duatis, M., Saez Zafra, M., & Garcia Fernandez, 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. *Communication & Society, 31(2)*, 1-24.

Camprubí, R., & Coromina, L. (2016). Content analysis in tourism research. *Tourism Management Perspectives, 18*, 134–140.

Carreras-Simó, M., & Coenders, G. (2020): Principal component analysis of financial statements. A compositional approach. *Revista de Métodos Cuantitativos para la Economía y la Empresa, 29*,18-37.

Coenders, G., & Ferrer-Rosell, B. (2020). Compositional data analysis in tourism. Review and future directions. *Tourism Analysis, 25(1)*, 153-168.

Coenders, G., Martín-Fernández, J.A., & Ferrer-Rosell, B. (2017). When relative and absolute information matter. Compositional predictor with a total in generalized linear models. *Statistical Modelling, 17*(6), 494-512.

Cruz-Sandoval, M., Ortego, M.I., & Roca, E. (2020). Tree ecosystem services, for everyone? A compositional analysis approach to assess the distribution of urban trees as an indicator of environmental justice. *Sustainability,12(3)*, 1215.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology, 37(7)*, 795–828.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *Test, 28*(3), 599-638.

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

Ferrer-Rosell, B., & Coenders, G. (2017). Airline type and tourist expenditure: Are full service and low-cost carriers converging or diverging? *Journal of Air Transport Managment, 63*, 119-125.

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

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

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

Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G., & Pawlowsky-Glahn, V. (2016). Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics, 22*, 269-291.

Ferrer-Rosell, B., Martin-Fuentes, E., & Marine-Roig, E. (2020) Diverse and emotional: Facebook content strategy by Spanish hotels. *Journal of Information Technology & Tourism*, *22(1)*, 53-74.

Hlebec, V., Kogovšek, T., & Coenders, G. (2012). The Measurement Quality of Social Support Survey Measurement Instruments. *Metodološki Zvezki, Advances in Methodology and Statistics, 9(1)*, 1-24.

Huertas, A., Ferrer-Rosell, B., Marine-Roig, E., & Cristobal-Fransi, E. (2021). Treatment of the Airbnb controversy by the press. *International Journal of Hospitality Management*, *95*, 102762.

Lalicic, L., Marine-Roig, E., Ferrer-Rosell, B. & Martin-Fuentes, E. (2021). Destination image analytics for tourism design. *Annals of Tourism Research, 86*, 103100.

Linares-Mustaros, S., Coenders, G., & Vives-Mestres, M. (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting, 40*, 1-10.

Marcillo-Delgado, J. C., Ortego, M.I., & Pérez-Foguet, A. (2019). A compositional approach for modelling

SDG7 indicators: Case study applied to electricity access. *Renewable and Sustainable Energy Reviews,107*, 388-398.

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

Morais, J., Thomas-Agnan, C., & Simioni, M. (2018). Using compositional and Dirichlet models for market share regression. *Journal of Applied Statistics, 45(9)*, 1670-1689.

Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015). Tools for compositional data with a total. *Statistical Modelling, 15(2)*, 175–190.

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

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*. Proceedings of the Royal Society, 60, 489–498.

Vives-Mestres, M., Martín-Fernández, J. A., & Kenett, R. S. (2016). Compositional data methods in customer survey analysis. *Quality and Reliability Engineering International, 32(6)*, 2115-2125.