SORT 47 (2) July-December 2023, 207-228 DOI: 10.57645/20.8080.02.6

40 years after Aitchison's article "The statistical analysis of compositional data". Where we are and where we are heading

Germà Coenders[1], Juan José Egozcue[2], Kamila Fačevicová[3], Carolina Navarro-López[4], Javier Palarea-Albaladejo[5], Vera Pawlowsky-Glahn[6], Raimon Tolosana-Delgado[7]

Abstract

The year 2022 marked 40 years since Aitchison published the article "The statistical analysis of compositional data". It is considered to be the foundation of contemporary compositional data analysis. It is time to review what has been accomplished in the field and what needs to be addressed. Astonishingly enough, many aspects seen as challenging in 1982 continue to lead to fruitful scholarly work. We commence with a bibliometric study and continue with some hot topics such as multi-way compositions, compositional regression models, dealing with zero values, non-logratio transformations, new application fields, and a number of current loose ends. Finally, a tentative future research agenda is outlined.

MSC: *62-02, 62-03, 62H99.*

Keywords: *Compositional data (CoDa), logratios, Aitchison geometry, multi-way compositions, zero replacement, compositional regression.*

[1] Dept. of Economics, University of Girona, Faculty of Economics and Business, Campus Montilivi, 17003 Girona, Spain. E-mail: germa.coenders@udg.edu

[2] Dept. of Civil and Environmental Engineering, UPC-Barcelona Tech, C2, Jordi Girona 1–3, 08034 Barcelona, Spain. E- mail: juan.jose.egozcue@upc.edu

[3] Dept. of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, 17. listopadu 12, 77146 Olomouc, Czech Republic. E-mail: kamila.facevicova@upol.cz

[4] Dept. of Applied Economics, University of The Balearic Islands, Faculty of Economics and Business, Cra. Valldemossa km 7.5, 07122 Palma, Spain. E-mail: carolina.navarro@uib.es

[5] Dept. of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain. E-mail: javier.palarea@udg.edu

[6] Dept. of Computer Science, Applied Mathematics and Statistics, University of Girona, 17003 Girona, Spain. E-mail: vera.pawlowsky@udg.edu

[7] Helmholtz-Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, Chemnitzer Str. 40, 09599 Freiberg, Germany. E-mail:r.tolosana@hzdr.de

Received: December 2022
Accepted: April 2023

1. Introduction

In 2022 we celebrated 40 years since the publication of the foundational article of contemporary compositional data (CoDa) analysis (Aitchison, 1982). The article and the methodology therein introduced has had a profound impact in statistics and in many scientific disciplines. It has been cited nearly 1,800 times according to the Web of Science (WoS) database and continues to be cited more often than ever. The 9^{th} edition of the International Workshop on Compositional Data Analysis took place in Toulouse, France, precisely in 2022. The present article results from the round table *40 years of Aitchison's seminal paper* held during such workshop. The list of authors is compiled in alphabetical order and all contributed equally as members of the discussion panel. Each author was responsible for the section corresponding to their intervention in the round table.

Most of the key concepts in the main body of the article by Aitchison (1982) continue to be used nowadays, especially the simplex sample space and the logratio transformations. A number of important topics (called *loose ends* using Aitchison's own words) were identified at the end of the article, and yet other issues were identified by the discussants. Some can be considered to be solved nowadays, such as principal component analysis of CoDa, modelling of independence, asymmetry of the additive logratio (alr) transformation, or the role of the geometry of the simplex. Some others continue to be *loose ends* in the third decade of the XXIst century, such as the treatment of essential zeros and measurement errors, or the joint analysis of compositions and other variables.

Nearly 20 years ago, Aitchison himself published a perspective article entitled "Compositional data analysis: where are we and where should we be heading?" (Aitchison and Egozcue, 2005). The title of the present piece of writing is an obvious tribute to that. As noted above, it builds on the discussions generated during the cited workshop and very much follows point-by-point the structure of Aitchisons's 1982 original paper, commenting on the ideas and points raised therein and subsequent developments, under an unavoidably biased and incomplete view, as perceived by a non-random sample of active scholars in the field, some of which had the immense fortune of working with John Aitchison. This 40th anniversary has attracted some interest with, as far as we are aware, a special issue by the Journal of Geochemical Exploration being in the works, and another article commemorating the anniversary, although not strictly guided by Aitchison's 1982 paper but offering a view of the discipline as a whole after this time (Greenacre et al., 2023). We hope to contribute to the literature by reviewing Aitchison's original insights under a contemporary view, and by providing up-to-date references on each important topic that was raised at the time, while suggesting a future research agenda.

The structure of the article is as follows. Section 2 presents a bibliometric study of the publications citing Aitchison (1982). In Section 3, the key features in Aitchison (1982) are identified, including the simplex sample space and the formal operations and transformations therein. In Section 4, multi-way compositions are presented, which include crossing two vector compositions as their most representative case. In Section 5, developments revolving around the use of a composition as a vector response or as a set

of explanatory variables are presented. In Section 6, the issue on how to deal with zeros is discussed, stressing the fact raised by Aitchison that the reason for the occurrence of zeros has to be taken into account. In Section 7, the *loose ends* are dealt with, both as presented by Aitchison and extracted from the subsequent discussion of the 1982 paper. The article concludes by suggesting a CoDa research agenda for the upcoming years. Some scientific fields where compositional data are increasingly used or may start to be used are identified, as well as some pending or needed methodological improvements, which are expected to expand the CoDa analysis toolbox or make it better suited to these fields.

2. Bibliometric study of Aitchison's 1982 article

Since the publication of the seminal work by John Aitchison in the *Journal of the Royal Statistical Society. Series B (Methodological)* in 1982, research in CoDa has found its way into a myriad of scientific disciplines. Through a bibliometric study based on the WoS database, this section updates (as of May 2022) the 2019 citation data previously

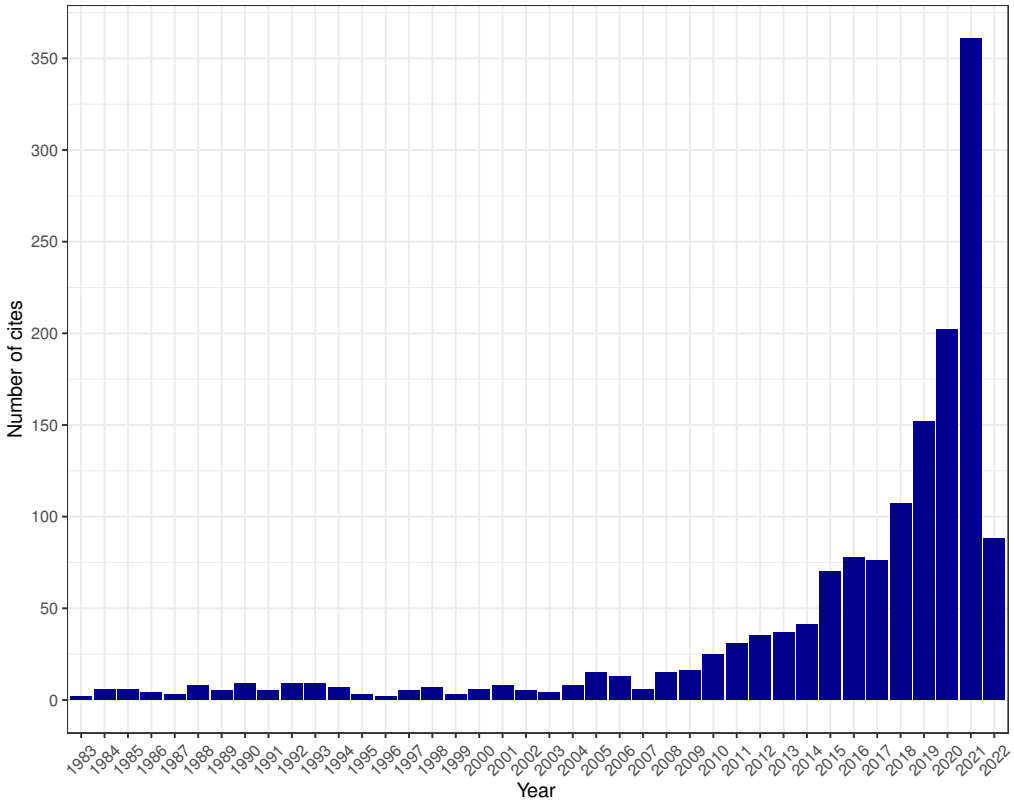

Figure 1. *Annual number of citations received by Aitchison's 1982 paper until May 2022, compiled from the WoS database.*

published (Navarro-López et al., 2021). It refers to main authors, institutions and research areas. It is important to note that the analysis presented here refers only to Aitchison's 1982 paper and, hence, it may not be representative of the entire CoDa field.

In the current bibliometric analysis, firstly, the temporal evolution of the number of papers citing Aitchison (1982) was analysed. The results show that the article has received citations uninterruptedly since its publication, although these were fairly low during the first years (1983-2007), not exceeding 10 citations in any year within the period (Figure 1). In contrast, the number of annual citations has grown exponentially since 2008, with more than 350 citations received in 2021 only.

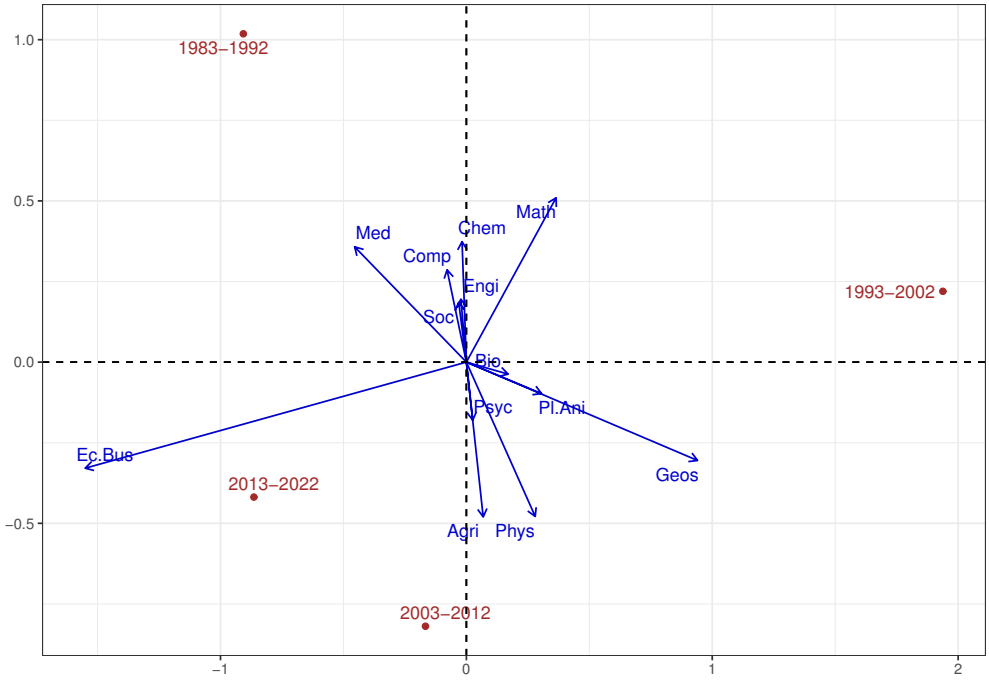

Figure 2. *Compositional biplot of cites by scientific categories contemplated in the WoS database by decade (from 1983 to 2022; Agri: agricultural sciences, Bio: biology and biochemistry, Chem: chemistry, Med: clinical medicine, Comp: computer science, Ec.Bus: economics and business, Engi: engineering, Geos: geosciences, Math: mathematics, Phys: physics, Pl.Ani: plant and animal science, Psyc: psychology, Soc: social science).*

The compositional biplot in Figure 2 displays the evolution of the relative importance of the fields of knowledge of the citing articles, by 10-year periods. The fields of clinical medicine, computer science, and chemistry stood out in the first ten years. The relative importance of the fields of mathematics and the geosciences was greater in the period 1993-2002. The subsequent 10-year period was dominated in relative terms by cites in the agricultural sciences and physics, whereas the last period studied was marked by the relative growth in citations in fields of economics and business. In absolute terms, and

over the entire 40-year period, the fields with the most citations are, in decreasing order, clinical medicine, biology-biochemistry, mathematics, engineering and chemistry. Even if the geosciences have been traditionally considered the field where CoDa methods have enjoyed most popularity, it can be observed that relative importance has shifted to other fields over the 40-year period.

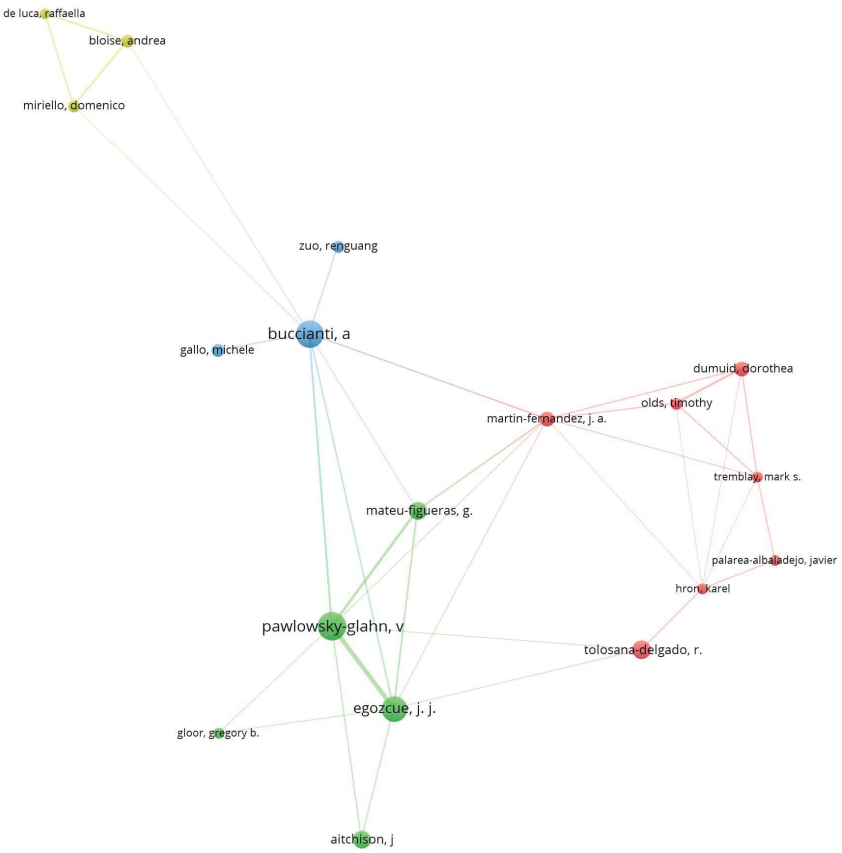

Figure 3. *Co-citation analysis of authors who have cited Aitchison's paper based on the WoS database. Node size indicates the number of citations received by an author; line thickness indicates multiple connections; line length is not relevant; the colours highlight groups of authors that are linked by a greater number of co-citations. Citation threshold set to 5 and showing the 100 most frequent co-citation connections.*

As to authors who have cited Aitchison's 1982 work the most, Vera Pawlowsky-Glahn (University of Girona, Spain) is at the top of the ranking, followed by Antonella Buccianti (University of Florence, Italy) and Juan José Egozcue (Technical University of Catalonia, Spain). Aitchison (1982) has been cited a total of 108 times by these three authors only. The co-citation analysis revealed four main groups (Figure 3). The right-most of these groups (in red) in Figure 3, consisting of seven authors, is centred on the authors Dorothea Dumuid and Tim Olds, both from the University of South Australia.

The group at the bottom left (in green) contains 5 authors and is mainly centred around Vera Pawlowsky-Glahn (University of Girona) and Juan José Egozcue (Technical University of Catalonia). A third group appears up the network (in blue), consisting of three authors and centred around Antonella Buccianti (University of Florence). Finally, the topmost group (in yellow), also consisting of three authors, is centred around Domenico Miriello and Andrea Bloise (University of Calabria). Note that the affiliations are those given at the time of citation.

3. Summary of essential features of Aitchison's 1982 paper

As stated in the introduction to Aitchison (1982), *the aim of the paper was to introduce a number of concepts of independence in the simplex, to relate these to some existing concepts, and to develop within the framework of rich new parametric classes of distributions appropriate statistical methods of analysis.* In some sense, it achieved its goal, as Aitchison (1982) is one of the most influential and cited papers in the field of CoDa analysis and its varied areas of application. Here, we intend to give an overview of the main concepts put forward in it, highlighting those that appeared particularly important to Aitchison himself, and those which became important in subsequent decades.

3.1. The simplex as sample space

One of the key concepts in Aitchison's approach is the insistence on the key role of the sample space of some random event. As stated right away in the first sentence of the abstract, *the simplex plays an important role as sample space in many practical situations where CoDa, in the form of proportions of some whole, require interpretation.* But already in the introduction, Aitchison states that *the simplex ... has proved to be an awkward space to handle statistically*. In Section 2.3 he states: *The idea of inducing a tractable class of distributions over some awkward sample space from a proven and well-established class over some simpler space is at least a century old. McAlister (1879), faced with the "awkward" sample space* $\mathbb{P}^1$ *—the positive semi-axis of the real line— saw that if he considered* $\mathbf{y} \in \mathbb{R}^1$ *to be* $\mathcal{N}(\mu, \sigma^2)$ *then the transformation* $\mathbf{x} = \exp(\mathbf{y})$ *would induce a useful "expnormal" distribution* $\Lambda(\mu, \sigma^2)$ *on* $\mathbb{P}^1$*: he, of course, expressed the idea in terms of the inverse, logarithmic, transformation and we are stuck with the name lognormal*. The approach put forward by McAlister was inspirational for Aitchison and led to the logratio methodology developed thereafter.

In the paper, cases with the simplex as sample space are mentioned which Aitchison does not consider to be compositional. More precisely, he writes: *There are also problems with simplex sample spaces where the data are not compositions; for example, probabilistic data in* $\mathbb{S}^d$ *occur in the analysis of subjective performance of inferential tasks (Aitchison, 1981). More complex product sample spaces, such as* $\mathbb{S}^d \times \mathbb{R}^C$ *or* $\mathbb{S}^d \times \mathbb{P}^C$*, also arise and succumb to the transformed normal technique.* This perspective has been nowadays reassessed, with discrete probability distributions being characterised

as ordinary discrete (finite-dimensional) compositions and continuous distributions being dealt with as infinite-dimensional compositions defined in a Bayes space (van den Boogaart, Egozcue and Pawlowsky-Glahn, 2014).

3.2. Parametric classes of distributions in the simplex and transformations

To introduce distributions on $\mathbb{S}^D$, where $\mathbb{S}^D$ stands for the D-part simplex, denoted in Aitchison (1982) by $\mathbb{S}^d$, with $d = D-1$ standing for the dimensionality of the set, Aitchison presented first the notation used for spaces and vectors, the definitions of basis of a composition, subcomposition, amalgamation and partition of order k, which stands for an amalgamation together with its associated subcompositions. Of these concepts, most appear in the compositional literature in their classical, conceptual form. The concept of basis of a composition has been later identified with the concept of composition itself through the definition of composition as an equivalence class (Aitchison, 1992). The concept of subcomposition prevails due to the principle of subcompositional coherence, an essential aspect of CoDa analysis, which has lately been reappraised (Greenacre et al., 2023).

Being the Dirichlet class of distributions in the simplex the best known at that time, it is logical that it is mentioned first. About it, Aitchison (1982) stated that *a major obstacle to its use in the statistical analysis of CoDa is that it seldom, if ever, provides an adequate description of actual patterns of variability of compositions.* The first reason to support this is attributed to the fact that isoprobability contours are convex and fail to describe concave patterns. The second reason is the strong independence structure derived from the relationship between the Dirichlet and the gamma classes.

The strategy introduced by McAlister (1879) led John Aitchison to define several parametric classes of transformed-normal distributions on the simplex as an alternative to the Dirichlet, the main one being based on the alr representation of compositions. In addition to this one, he introduced the multiplicative logratio and the hybrid logratio, although the latter two have had only a short run in the literature.

Interestingly, when talking about logistic transformations in general, it is stated that *the* alr *is by no means the only transformation ... An obvious comment is that the exponential function used in the definition is not an essential feature; it could be replaced by any one-to-one transformation ..., though there are few transformations as tractable.*

About the *validity of transformed-normal models*, Aitchison writes: *Any statistical weapon designed to overcome such a resistant fortress as the simplex is unlikely to gain acceptance before undergoing proving tests as to its suitability to the terrain.* He then mentions several goodness-of-fit tests for multivariate normality and applies them to as many as 20 different data sets, reassuring the potential of transformed-normal models in practice.

While addressing the issue of *genesis models*, Aitchison (1982) introduces the concept of *perturbation* through central limit theory arguments. Over the years, *perturbation* has proven to be an essential operation in the analysis of CoDa, as it has the same prop-

erties for elements in the simplex as addition in the d-dimensional real space: it is an Abelian group operation.

3.3. Other key ideas and concepts

An extensive section is devoted to the analysis of independence, which we are not going to detail here as its impact on further developments has been really limited. Two types of structural analysis are addressed: the *extrinsic analysis*, interested in relationships between compositions and the basis they are derived from, and the *intrinsic analysis* interested in the composition *per se*. The first one lost interest when compositions were identified with equivalence classes, the second one did so because partitions of order k did not find their way into applications.

In Aitchison (1982) some concepts are mentioned marginally that have however flourished in later developments. These include the extent, if any, to which a composition depends on size or total or, in relation to linear models in geostatistics, the extent to which depth is explanatory of compositional pattern (see Sections 5 and 7). Also, the problem of zeros (see Section 6) and the presence of outliers and priority groups or categories are briefly discussed.

4. Multi-way compositions

One of the loose ends mentioned in Aitchison (1982) is the analysis of multi-way or multi-factorial CoDa. This data object results from a distribution of a mass into categories given by a combination of levels of two or more factors. Alternatively, it can be understood as a set of observations of a classical compositional vector over several time points or a multinomial realisation. Since Aitchison's paper, two main strategies for their analysis have appeared in the literature. Thus, Gallo (2015) relies on an adaptation of the so-called Tucker3 model from multi-way analysis, and Fačevicová, Filzmoser and Hron (2023) aims to generalise the ordinary logratio analysis of vector (one-way) compositions.

Focusing on the last-mentioned approach, multi-way compositions bring a new range of problems related to the relationships between the constituting factors. Egozcue, Díaz-Barrero and Pawlowsky-Glahn (2008) defined the orthogonal decomposition of compositional tables (two-way compositions) into independent and interaction parts. Fačevicová et al. (2023) extended the orthogonal decomposition to the multi-way case, where the interaction part can be further decomposed into objects preserving more specific sources of interaction. Moreover, they defined a system of orthonormal coordinates, which allows analysing both the whole multi-way relative structure and its decomposed parts using conventional statistical methods. An alternative perspective to the decomposition of two-way compositions is put forward by Egozcue and Maldonado (2021) that, motivated by the problem of exchange rate matrices, focuses on positive square matrices.

The possibility of isolating specific bits of information in a composition allows for understanding the overall multi-way structure more deeply, e.g by assessing the impor-

tance of each decomposed part in it. Moreover, when a set of multi-way compositions is available, it is possible to evaluate the effect of changes in a decomposed part on an external variable by regression analysis (Fačevicová, Kynčlová and Macku, 2021). The multi-way approach is also consistent with recent advances in the analysis of continuous compositions within the framework of Bayes spaces. This has made it possible to transfer the idea of orthogonal decomposition to the analysis of bivariate and even multivariate densities and their relationship with copulas Genest, Hron and Nešlehová (2023).

The methodology for the analysis of multi-way compositions currently presents several open challenges. Firstly, an objective comparative study of the different strategies proposed as mentioned above is needed. Secondly, the ordinary non-compositional approach based on generalised linear models (GLMs) is well established within the wider statistical community. It is then necessary to comprehensibly discuss how the multi-way compositional approach differs from that gold standard and, reciprocally, what elements of the GLM approach could be brought over into the compositional framework. Pioneering work along these lines has been done by Vencálek, Hron and Filzmoser (2020). Moreover, one popular direction in vector CoDa analysis focuses on the elemental information contained in pairwise logratios, which may provide a more natural interpretation than, in occasions, complex logratio representations (Hron et al., 2021). This idea can be extended to the case of multi-way compositions. Thus, in the case of compositional tables, the concept of pairwise logratios would be extended to simple four-part odds-ratios. Additionally, there are open questions about dealing with zeros or robustness and outlier detection in multi-way compositions, among others.

5. Compositional regression models

One of the key abilities provided by the transformation strategies mentioned in Aitchison (1982) is to allow for the construction of conditional models involving compositions. That is, to predict the behaviour of a composition in response to changes in one or more explanatory variables, or to express the dependence of an external variable on a composition. The move of expressing a response composition in terms of alr directly allowed to establish a linear model for each logratio as a function of the variable(s) of interest, opening the door to regression (Aitchison, 1982, p. 156), ANOVA and ANCOVA with compositional response. Swapping roles, it did not take much effort to define analogous linear and quasi-linear prediction methods for continuous (via multiple regression) and categorical (via linear and quadratic discriminant analysis, or via logistic regression; van den Boogaart and Tolosana-Delgado, 2013) responses where the composition plays the role of explanatory variable. Composition-on-composition regression could also be solved with the same idea: take logratios on both the response and the explanatory compositions and apply the desired predictive method with both sets of logratios.

Of the several logratio transformations most commonly used, the alr and ilr (isometric logratio) transformations are ideal for all these tasks, as they generally provide a full-rank (hence invertible) covariance matrix of the explanatory variable set. Even

the centred logratio transformation (clr) (Aitchison, 1983) is a valid choice, with some caveats: the clr is directly useful if the composition plays the role of response variable, but requires the use of some generalised inversion if the composition is placed in lieu of the explanatory variables. Indeed, models obtained with the data expressed in any one of these representations should, and for most models can, be unequivocally transformed into models for data in another logratio representation. Otherwise, one would not be obtaining an intrinsically compositional model, but a model for a specific logratio representation. The fact that such models are intrinsic and not logratio-dependent gives rise to the possibility to entirely define them using operations of the simplex as a Euclidean space, such as a perturbation-linear combination for regression with compositional response, or linear applications in the simplex for other more complex regression cases (van den Boogaart et al., 2021). What changes between the several transformations is rather the interpretation of the regression coefficients, and with them, the relevance of certain tests of linear independence (Coenders and Pawlowsky-Glahn, 2020). van den Boogaart et al. (2021) argue therefore that, barring particularities of a specific case study, only tests of subcompositional inference are interpretable in general terms, as these tests yield the same result no matter which logratio transformation is used.

6. Dealing with zeros

Zero values do not fit into the logratio framework as they are not compatible with the ratio (if placed in the denominator) nor logarithm operations. More fundamentally, in the context of relative data, we can understand that a zero is not relative to anything. Although it is not a problem exclusive to CoDa, it has some additional difficulties here because a single zero part renders the whole composition only defined up to a multiplicative constant.

This issue is of course treated in Aitchison (1982) and the accompanying discussion, where *ad hoc* solutions like meaningful amalgamation of some parts or replacement by sensible positive amounts are advocated. In any case, any decision should consider the amount and the nature of the zeros. Insightful discussions about different kinds of zeros can be found in e.g. Bacon-Shone (2003), Martín-Fernández, Palarea-Albaladejo and Olea (2011), and van den Boogaart and Tolosana-Delgado (2013, Ch. 7). Although the issues with zeros, particularly so-called essential zeros, have prompted alternative representations for CoDa through e.g. Box-Cox, power or hyperspherical transformations (Greenacre, 2010), these come in conflict in different ways with the basic principles of CoDa analysis and the simplicial geometry as discussed in this contribution. We focus here on approaches building on the logratio methodology following Aitchison (1982).

6.1. Replacement strategies

The most prevalent class of zeros in CoDa applications so far concurs with the concept of type I left-censored data, i.e. small values that have been rounded off or have fallen below a limit of detection, so that the measured values cannot be distinguished from a

blank signal or background noise at a specified level of confidence. Thus, the actual value is unknown, but it must be somewhere in the lower tail of the data distribution below either a known threshold or the minimum observed value. Hence, zero replacement or imputation has become a common strategy in this setting.

Any zero imputation method should preserve the natural features of data defined on a simplex as well as their relative variation structure. From a non-parametric viewpoint, crude substitution of zeros by positive values or early proposals such as the additive replacement method found in Aitchison (1986) do not preserve the ratios between non-zero parts and, hence, distort the co-dependence structure of the data. Instead, Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2003) proposed a simple replacement by a prefixed fraction of the threshold while applying a multiplicative adjustment to the non-zero parts so that their ratios are preserved under data closure. Alternatively, model-based methods are feasible through logratio coordinate representations. Thus, a consistent parametric procedure was introduced in Palarea-Albaladejo, Martín-Fernández and Gómez-García (2007), based on the expectation-maximisation (EM) algorithm. Since then, a number of specialised imputation methods have been developed following the principles of the logratio approach in different flavours (Palarea-Albaladejo and Martín-Fernández, 2015). More recently, and partly motivated by the growing popularity of CoDa analysis in molecular sciences, the treatment of zeros in compositional count data has caught some attention (Martín-Fernández et al., 2015). These correspond to discrete-valued compositions derived from counting processes where zeros are usually associated to limited sampling or under-reporting issues, e.g. zeros for rare taxa due to limited sequencing depth in microbiome studies. Moreover, high-dimensionality is a common feature of some modern data formats that requires adapted methods (Palarea-Albaladejo et al., 2022).

6.2. Essential zeros

Being also known as structural or absolute zeros, this class does not correspond to censored values but to genuine zeros. Although the decision on whether a zero is considered essential or not lays ultimately on the analyst, it is clear that imputation would be at least debatable here. As suggested in Aitchison (1982), a sensible amalgamation of parts can circumvent the problem in some cases. In others, the zero patterns can actually reflect on a structure of subpopulations in the data, then suggesting to carry out separate statistical analyses. Otherwise, as anticipated by Aitchison (1982), no general methodology for essential zeros has been feasible so far and the ideas outlined therein remain current. As noted, *some form of conditional modelling separating out the zero may be possible* so that the zeros are incorporated to the own modelling. Proposals along these lines include the binomial conditional logistic normal model in Aitchison and Kay (2003), the Poisson-lognormal model in Bacon-Shone (2008) for discrete compositions, the multiplicative logistic normal mixture model in Stewart and Field (2011) or the logistic normal mixture model in Bear and Billheimer (2016). However, their applicability is limited to particular settings and subject to stringent conditions.

7. Loose ends in the 1982 paper

7.1. Loose ends according to J. Aitchison in 1982

At the end of his 1982 paper, J. Aitchison enumerates a list of topics, called *loose ends*, that in his opinion need further attention. The fact that these topics are mentioned supports the idea that he conceived the paper as a starting point and reveals his deep understanding of CoDa. One of the topics is multi-way compositions, which has already been discussed in Section 4. In the following paragraphs, we comment on how the remaining ones have been developed up to now, with some being closed and some others being active research areas currently.

Other transformations from $\mathbb{S}^D \to \mathbb{R}^{D-1}$ constitute a first controversial point which is permanently under discussion from different points of view. The main issue seems to be why CoDa should be transformed. The origin of such divergences is the initial definition of CoDa as vectors of positive components (parts) adding to a constant, say 1 or 100. This suggests that CoDa could contain zero parts and still add to the chosen constant. The modification of the logratio transformations should allow to include zero values in CoDa. The use of Box-Cox transformations or the like discussion in Aitchison (1982); Barceló, Pawlowsky-Glahn and Grunsky (1996) or representations on the first orthant of the hypersphere (Scealy and Welsh, 2011), are motivated by the inclusion of zeros in the analysis and by the use of tools for non-Euclidean geometries. The α-transformation approach by Tsagris and Stewart (2020) or the embedding in Greenacre (2010) are generalisations of the power-transformations. Logratio transformations exclude zeros in the data, but they were accepted in most CoDa analyses, even before Aitchison's contribution (Lewi, 1976). Beyond the alr and clr, the ilr (Egozcue et al., 2003) was added to the catalog of logratio transformations. Recently, the amalgamated balance logratio transformation (Sauro Graziano, Gozzi and Buccianti, 2020; Greenacre et al., 2023) has also been proposed. However, after Pawlowsky-Glahn and Egozcue (2001), the idea that logratio transformations should be considered as a representation of CoDa in coordinates better than as a pure transformation was consolidated. In particular, these coordinates can be orthogonal and Cartesian (straight axes) and they can be identified with the components of an ilr transformation. As precedents, Lewi (1976) used a particular case of ilr coordinates, Lindley (1988) proposed non-normalised ilr coordinates, and the principal component analysis for CoDa (Aitchison, 1983) also provided ilr-coordinates.

The topic of principal component analysis using alr coordinates was closed almost immediately in Aitchison (1983), where the alr was substituted advantageously by the clr representation. The theory of principal component analysis (PCA) with CoDa was mainly complete in 1983, but it was extended by Aitchison and Greenacre (2002) who focused on the interpretation of biplots.

Normal models with different transformations constitute the next loose end. After the 1982 paper, Aitchison continued studying families of distributions in order to conciliate the independence structure of the Dirichlet distribution and the logistic-normal family. The change of transformation from alr to clr was studied in van den Boogaart

and Tolosana-Delgado (2013) producing a singular normal distribution. The introduction of ilr coordinates allowed to define the normal distribution on coordinates in real space (Mateu-Figueras, Pawlowsky-Glahn and Egozcue, 2013). Nowadays, the study of distributions for CoDa is very much reduced to the distributions in the real space of ilr-coordinates. However, some aspects of joint normality on different types of coordinates remain open and call for further research. See Greenacre et al. (2023) for some proposed alternatives.

The topic of joint analysis of compositions and other variables recurrently appears from the very beginning of CoDa analysis (Lewi, 1976) and is also treated in the particular case of the sum of parts in Aitchison (1982). The article by van den Boogaart and Tolosana-Delgado (2008) discusses exploratory tools, dependence models, and display tools for co-dependence between compositions and variables on other scales (e.g. categorical, dichotomous, ordinal, and others). Including external variables in compositional PCA and associated biplots has been also studied (Daunis-i-Estadella, Thió-Henestrosa and Mateu-Figueras, 2011). However, a comprehensive characterisation of the structure of the joint sample space remains open.

The last identified loose end were the distributional problems to model different independence concepts. The independence concepts introduced in Aitchison (1982) were related to the dependence structure inherent to the Dirichlet distribution and can be considered settled. The initial studies on the logistic-normal distribution or variants thereof did not provide further insight into independence (internal or external). The definition of normality on the simplex using ilr-coordinates (Mateu-Figueras et al., 2013) clarified most of the different concepts of independence that can be discussed in terms of balance coordinates. For instance, the fact that a subcomposition can be viewed as an orthogonal projection within the CoDa geometry (Egozcue and Pawlowsky-Glahn, 2005), or that joining two subcompositions depends on a single balance considered as an ilr coordinate. This allows reducing independence issues to independence between real variables.

7.2. Some loose ends arising from the discussion in 1982

There were 20 discussants of the 1982 Aitchison's paper. Most of them set a number of interesting comments which were replied to by Aitchison at the end of the discussion. The asymmetry of alr coordinates was mentioned repeatedly in the discussion, namely by J.A. Anderson, A.P. Dawid, C.A.B. Smith, and R. Thomson. In the rejoinder by Aitchison, the clr transformation was mentioned as a solution to the asymmetry problem with the alr transformation, thus motivating the subsequent paper, Aitchison (1983), where clr, distance between compositions and compositional PCA were introduced, mostly closing the debate on this issue. Even so, some points which arose in the discussion can still be considered as *loose ends*, including the zero issue in Section 6. Let us comment on some others.

The discussant N.I. Fisher raised the role of the geometry of the sample space in writing *... one is ultimately better off working within the confines of the original geometry (of the circle, sphere, cylinder, ...), and with techniques particular thereto (vector*

methods, etc.) ... This point was retrieved more than a decade later – see Egozcue and Pawlowsky-Glahn (2019) for further details. Aitchison's idea that the simplex is an appropriate sample space for CoDa is nowadays complemented with an Euclidean space structure known as Aitchison geometry after the term was introduced in Pawlowsky-Glahn and Egozcue (2001).

J.A Anderson, N.I Fisher, D.A Preece and others mentioned the importance of error control in (log) ratios. In our opinion, a lot more can be done about this point. Mert, Filzmoser and Hron (2016) study how errors in raw measurements influence the computation of logratios. Alternative approaches consider cellwise uncertainties (Pospiech, Tolosana-Delgado and van den Boogaart, 2020), control for cellwise outliers (Štefelová et al., 2021), or multiple measurements of the composition to either estimate measurement error variance (Nguyen et al., 2020), obtain unbiased estimates in the presence of measurement error (Kogovšek, Coenders and Hlebec, 2013), or downplay the parts with the highest error variances (Hron et al., 2022).

8. Outlook

Over the last 40 years, the log-ratio approach to CoDa has shown to be a powerful and much needed tool in many fields of science. It provides a coherent, consistent and interpretable way to tackle the statistical analysis of CoDa. To conclude this review, we draft a tentative CoDa research agenda for the upcoming years, regarding both application fields and methodology.

8.1. New application fields

Historically, it is commonly quoted that awareness about issues related to the statistical analysis of CoDa was firstly raised in the geosciences. It was this area of application that mostly fuelled the early developments of the CoDa methods, as e.g. reflected by the illustrative examples in Aitchison (1982). Although it has continued to be amongst the leading application areas over time (see e.g. the recent perspective offered in Buccianti and Gozzi, 2021), certainly the range of scientific disciplines where CoDa analysis is being used has been significantly widened since then, particularly in the last 10-15 years. Without intending to be exhaustive, and being aware that we will be inevitably omitting a good number of other interesting novel applications; we note in the following some recent areas where CoDa methods have been successfully introduced and where, from our perspective, there exists good scope for future developments.

As to the natural and health sciences, molecular biology and the *omics* sciences in general represent a good example of fields where awareness of the compositional character of many types of data generated therein have notably grown in recent years (Gloor et al., 2017; Greenacre et al., 2023; Quinn et al., 2018). Something similar has happened in physical activity and time-use epidemiology. In less than a decade, the use of CoDa analysis in this area has rocketed once the dominant paradigm has been shifted from studying the effects of activity types individually to conceptualising the

daily 24-hour activity pattern as a whole mix of inter-dependent behaviours, with the CoDa approach being recognised as a unifying statistical analysis framework (Chastin et al., 2015; Dumuid et al., 2020). Another recent health-related example is the study of diets, representing trade-offs between food groups or macronutrient groups in total food intake (Trinh et al., 2019). Moreover, the environmental sciences are also becoming a fruitful area of CoDa applications, with traditional interest in soil pollutant composition expanding to pollutants in water (Bondu et al., 2020), air (Sánchez-Balseca and Pérez-Foguet, 2020), and smoke emissions (Weise et al., 2020).

Glancing over the published literature suggests that CoDa analysis is still generally underused in the social sciences as a whole. Some areas in which CoDa appear naturally include political science (Nguyen, 2019), demography (Lloyd, Pawlowsky-Glahn and Egozcue, 2012), content analysis and text mining (Marine-Roig and Ferrer-Rosell, 2018), and questionnaire surveys, which often use a question format asking respondents to distribute 100 points among a set of choices (van Eijnatten, van der Ark and Holloway, 2015). Moreover, CoDa applications in economics and business are also gaining momentum. Early examples focused on expenditure distribution, either by families in their homes, or by overseas tourists (Ferrer-Rosell, Coenders and Martínez-Garcia, 2015). Other natural fields of application which are even newer are market share analysis (Morais, Thomas-Agnan and Simioni, 2018) and portfolio analysis (Belles-Sampera, Guillen and Santolino, 2016). In some instances, data are directly given by ratios, as in relative prices and exchange rates (Maldonado, Egozcue and Pawlowsky-Glahn, 2021) or accounting ratios in financial statement analysis (Linares-Mustarós, Coenders and Vives-Mestres, 2018). Even if an uncritical use of standard statistical methods has repeatedly been proved misleading here, CoDa methods are used only rarely. A challenge for the CoDa community is therefore not just to look for such new application fields, but also to carefully discuss the CoDa approach and its principles, in comparison with traditional approaches, with subject-matter experts so that they are definitely persuaded.

8.2. New methodological developments

Methodological developments in CoDa analysis have been traditionally driven by applications in diverse fields, providing new scientific insights and perspective through statistical analysis based on log-ratio coordinate representations. There are still fields which compositional methods have hardly interacted with. For instance, modern developments in data science and machine learning is an obvious emerging area of interest to expand CoDa ideas and methods into (Tolosana-Delgado et al., 2019). Also novel application in physics may pose interesting challenges in relation to e.g. endmember unmixing, scaling invariances in physical laws, or natural scales of physical magnitudes.

Within the mathematical and statistical sciences, there is also plenty of scope for stimulating interconnections and cross-disciplinary developments. Examples include functional data (Hron et al., 2016; Genest et al., 2023) and weighted analysis (Hron et al., 2022) building on the theory of Bayes spaces (van den Boogaart et al., 2014), algebra (Egozcue et al., 2011), graph theory (Greenacre, 2021) or compositional differ-

ential equations (Egozcue and Jarauta-Bragulat, 2014). In relation to high-dimensional compositions, there are at least two current topics which generate intense discussions: the measurement of associations between parts (Egozcue, Pawlowsky-Glahn and Gloor, 2018; Erb and Notredame, 2015) and variable selection (Calle, Pujolassos and Susin, 2023; Greenacre, 2019; Greenacre et al., 2023; Shi et al., 2016).

Acknowledgements

This article was supported by the Spanish Ministry of Science and Innovation/AEI/10.13039/501100011033 and by ERDF – A way of making Europe [grant numbers PID2021-123833OB-I00; PID2021-125380OB-I00], the Department of Research and Universities of the Generalitat de Catalunya [grant number 2021SGR01197], and the Czech Science Foundation [grant number 22-15684L].

References

Aitchison, J. (1981). Some distribution theory related to the analysis of subjective performance in inferential tasks. In Taillie, C., Patil, G. P., and Baldessari, B., editors, *Statistical Distributions in Scientific Work*, volume 5, pages 363–386, Dordrecht (NL). Reidel Publishing Company.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 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 & Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press), London (UK). 416 p.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379.

Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850.

Aitchison, J. and Greenacre, M. (2002). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4):375–392.

Aitchison, J. and Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In Daunis-i-Estadella, J. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona. Available at https://dugi-doc.udg.edu/handle/10256/652.

Bacon-Shone, J. (2003). Modelling structural zeros in compositional data. In Daunis-i-Estadella, J. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03,*

The 1st Compositional Data Analysis Workshop, Girona (E). Universitat de Girona. Available at https://dugi-doc.udg.edu/handle/10256/661.

Bacon-Shone, J. (2008). Discrete and continuous compositions. In Daunis-i-Estadella, J. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona. Available at https://dugi-doc.udg.edu/handle/10256/712.

Barceló, C., Pawlowsky-Glahn, V., and Grunsky, E. (1996). Some aspects of transformations of compositional data and the identification of outliers. *Mathematical Geology*, 28(4):501–518.

Bear, J. and Billheimer, D. (2016). A logistic normal mixture model for compositional data allowing essential zeros. *Austrian Journal of Statistics*, 45(4):3–23.

Belles-Sampera, J., Guillen, M., and Santolino, M. (2016). Compositional methods applied to capital allocation problems. *Journal of Risk*, 19:15–30.

Bondu, R., Cloutier, V., Rosa, E., and Roy, M. (2020). An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada). *Applied Geochemistry*, 114:104500.

Buccianti, A. and Gozzi, C. (2021). The whole versus the parts: The challenge of compositional data analysis (coda) methods for geochemistry. In Filzmoser, P., Hron, K., Martín-Fernández, J. A., and Palarea-Albaladejo, J., editors, *Advances in Compositional Data Analysis*, pages 253–264, Cham (CH). Springer International Publishing.

Calle, M. L., Pujolassos, M., and Susin, A. (2023). coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies. *BMC bioinformatics*, 24(1):82.

Chastin, S. F. M., Palarea-Albaladejo, J., Dontje, M. L., and Skelton, D. A. (2015). Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PloS One*, 10(10):e0139984.

Coenders, G. and Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT-Statistics and Operations Research Transactions*, 44(1):201–220.

Daunis-i-Estadella, J., Thió-Henestrosa, S., and Mateu-Figueras, G. (2011). Including supplementary elements in a compositional biplot. *Computers & Geosciences*, 37(5):696–701.

Dumuid, D., Pedišić, Ž., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K., and Olds, T. (2020). Compositional data analysis in time-use epidemiology: what, why, how. *International Journal of Environmental Research and Public Health*, 17(7):2220.

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L., and Mateu-Figueras, G. (2011). Elements of simplicial linear algebra and geometry. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional*

Data Analysis: Theory and Applications, pages 141–157, Chichester (UK). John Wiley & Sons.

Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2008). Compositional analysis of bivariate discrete probabilities. In Daunis-i-Estadella, J. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona. Available at https://dugi-doc.udg.edu/handle/10256/717.

Egozcue, J. J. and Jarauta-Bragulat, E. (2014). Differential models for evolutionary compositions. *Mathematical Geosciences*, 46(4):381–410.

Egozcue, J. J. and Maldonado, W. L. (2021). An interpretable orthogonal decomposition of positive square matrices. In Filzmoser, P., Hron, K., Martín-Fernández, J. A., and Palarea-Albaladejo, J., editors, *Advances in Compositional Data Analysis*, pages 1–18, Cham (CH). Springer.

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

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

Egozcue, J. J., Pawlowsky-Glahn, V., and Gloor, G. B. (2018). Linear association in compositional data analysis. *Austrian Journal of Statistics*, 47(1):3–31.

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

Erb, I. and Notredame, C. (2015). How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135(1-2):21–36.

Fačevicová, K., Filzmoser, P., and Hron, K. (2023). Compositional cubes: a new concept for multi-factorial compositions. *Statistical Papers*, 64(3): 955–985.

Fačevicová, K., Kynčlová, P., and Macku, K. (2021). Geographically weighted regression analysis for two-factorial compositional data. In Filzmoser, P., Hron, K., Martín-Fernández, J. A., and Palarea-Albaladejo, J., editors, *Advances in Compositional Data Analysis*, pages 103–124, Cham (CH). Springer.

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

Gallo, M. (2015). Tucker3 model for compositional data. *Communications in Statistics - Theory and Methods*, 44(21):4441–4453.

Genest, C., Hron, K., and Nešlehová, J. G. (2023). Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copulas-based representation. *Journal of Multivariate Analysis*, 198, 105228.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 15:2224.

Greenacre, M. (2010). Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences*, 42:129–134.

Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, 51:649–682.

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and its Application*, 8:271–299.

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., and Quinn, T. (2023). Aitchison's compositional data analysis 40 years on: A reappraisal. *Statistical Science*, 38(3): 386–410

Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M., and Matys-Grygar, T. (2021). Analysing pairwise logratios revisited. *Mathematical Geosciences*, 53(7):1643–1666.

Hron, K., Menafoglio, A., Palarea-Albaladejo, J., Filzmoser, P., Talská, R., and Egozcue, J. J. (2022). Weighting of parts in compositional data analysis: Advances and applications. *Mathematical Geosciences*, 54(1):71–93.

Hron, K., Menafoglio, A., Templ, M., Hruzová, K., and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350.

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

Lewi, P. J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneimittelforschung*, 26(7):1295–1300.

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

Lindley, D. V. (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 3*, pages 307–326, New York (NY). Oxford University Press.

Lloyd, C. D., Pawlowsky-Glahn, V., and Egozcue, J. J. (2012). Compositional data analysis in population studies. *Annals of the Association of American Geographers*, 102:1–16.

Maldonado, W. L., Egozcue, J. J., and Pawlowsky-Glahn, V. (2021). No-arbitrage matrices of exchange rates: Some characterizations. *International Journal of Economic Theory*, 17:375–389.

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

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278.

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*, 15(2):134–158.

Martín-Fernández, J. A., Palarea-Albaladejo, J., and Olea, R. (2011). Dealing with zeros. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*, pages 43–58, Chichester (UK). John Wiley & Sons.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *SORT - Statistics and Operations Research Transactions*, 37(1):29–56.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London*, 29:367–376.

Mert, M. C., Filzmoser, P., and Hron, K. (2016). Error propagation in isometric log-ratio coordinates for compositional data: Theoretical and practical considerations. *Mathematical Geosciences*, 48:941–961.

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

Navarro-López, C., González-Morcillo, S., Mulet-Forteza, C., and Linares-Mustarós, S. (2021). A bibliometric analysis of the 35th anniversary of the paper "The statistical analysis of compositional data" by John Aitchison (1982). *Austrian Journal of Statistics*, 50(2):38–55.

Nguyen, H. D., Tran, K. P., Celano, G., Maravelakis, P. E., and Castagliola, P. (2020). On the effect of the measurement error on Shewhart t and EWMA t control charts. *The International Journal of Advanced Manufacturing Technology*, 107(9):4317–4332.

Nguyen, T. H. A. (2019). *Contribution to the statistical analysis of compositional data with an application to political economy*. PhD thesis, Université Toulouse 1, Toulouse (FR).

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*, 143:85–96.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Gómez-García, J. A. (2007). Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39(7):625–645.

Palarea-Albaladejo, J., Martín-Fernández, J. A., Ruiz-Gazen, A., and Thomas-Agnan, C. (2022). lrSVD: an efficient imputation algorithm for incomplete high-throughput compositional data. *Journal of Chemometrics*, 36(12):e3459.

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

Pospiech, S., Tolosana-Delgado, R., and van den Boogaart, K. G. (2020). Discriminant analysis for compositional data incorporating cell-wise uncertainties. *Mathematical Geosciences*, 53:1–20.

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878.

Sánchez-Balseca, J. and Pérez-Foguet, A. (2020). Spatio-temporal air pollution modelling using a compositional approach. *Heliyon*, 6(9):e04794.

Sauro Graziano, R., Gozzi, C., and Buccianti, A. (2020). Is Compositional Data Analysis (CoDA) a theory able to discover complex dynamics in aqueous geochemical systems? *Journal of Geochemical Exploration*, 211:106465.

Scealy, J. L. and Welsh, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(3):351–375.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.

Štefelová, N., Alfons, A., Palarea-Albaladejo, J., Filzmoser, P., and Hron, K. (2021). Robust regression with compositional covariates including cellwise outliers. *Advances in Data Analysis and Classification*, 15(4):869–909.

Stewart, C. and Field, C. (2011). Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(1):45–69.

Tolosana-Delgado, R., Talebi, H., Khodadadzadeh, M., and van den Boogaart, K. G. (2019). On machine learning algorithms and compositional data. In Egozcue, J. J., Graffelman, J., and Ortego, M. I., editors, *Proceedings of CoDaWork2019, the 8th International Workshop on Compositional Data Analysis*, Barcelona (E). Catalan Polytechnic University-BarcelonaTECH. Available at https://webs.camins.upc.edu/codawork2019/proceedings/book-proceedings-CoDaWork2019-correctedv.pdf.

Trinh, H. T., Morais, J., Thomas-Agnan, C., and Simioni, M. (2019). Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical Methods in Medical Research*, 28(8):2305–2325.

Tsagris, M. and Stewart, C. (2020). A folded model for compositional data analysis. *Australian & New Zealand Journal of Statistics*, 62(2):249–277.

van den Boogaart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian and New Zealand Journal of Statistics*, 56(2):171–194.

van den Boogaart, K. G., Filzmoser, P., Hron, K., Templ, M., and Tolosana-Delgado, R. (2021). Classical and robust regression analysis with compositional data. *Mathematical Geosciences*, 53(5):823–858.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2008). Mixing compositions and other scales. In Daunis-i-Estadella, J. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona. Available at https://dugi-doc.udg.edu/handle/10256/743.

van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer, Berlin (DE). 258 p.

van Eijnatten, F. M., van der Ark, L. A., and Holloway, S. S. (2015). Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis. *Quality & Quantity*, 49(2):559–579.

Vencálek, O., Hron, K., and Filzmoser, P. (2020). A comparison of generalised linear models and compositional models for ordered categorical data. *Statistical Modeling*, 20(3):249–273.

Weise, D. R., Palarea-Albaladejo, J., Johnson, T. J., and Jung, H. (2020). Analyzing wildland fire smoke emissions data using compositional data techniques. *Journal of Geophysical Research: Atmospheres*, 125(6):e2019JD032128.