Compositional analysis of dietary patterns

Accepted author version to appear in Statistical Methods for Medical Research

Author list: Solans M[1,2,3], Coenders G[2,1], Marcos-Gragera R[3,2], Castelló A[4,1,5], Gràcia-Lavedan E[6,7,1], Benavente Y[8], Moreno V[9,10,1,11], Pérez-Gómez B[12,1], Amiano P[13,1], Fernández-Villa T[14], Guevara M[15,1], Gómez-Acebo I[16,1], Fernández-Tardón G[17,1], Vanaclocha-Espi M[18], Chirlaque MD[19,1], Capelo R[20], Barrios R[21], Aragonés N[1,22], Molinuevo A[1], Vitelli-Storelli F[14], Castilla J[15,1], Dierssen-Sotos T[16,1], Castaño-Vinyals G[6,7,23,1], Kogevinas M[6,7,23,1], Pollán M[1,4], Saez M[2,1]

1. Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain Spain.
2. Research Group on Statistics, Econometrics and Health (GRECS), Universitat de Girona, Girona, Spain
3. Epidemiology Unit and Girona Cancer Registry, Oncology Coordination Plan, Department of Health, Autonomous Government of Catalonia, Catalan Institute of Oncology, Girona, Spain
4. Cancer Epidemiology Unit, National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain
5. Faculty of Medicine, University of Alcalá, Alcalá de Henares, Madrid, Spain
6. ISGlobal, Barcelona, Spain
7. Universitat Pompeu Fabra (UPF), Barcelona, Spain
8. Unit of molecular and genetic epidemiology in infections and cancer, Catalan Institute of Oncology (ICO-IDIBELL), Barcelona, Spain
9. Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO). Hospitalet de Llobregat, Barcelona, Spain.
10. Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). Hospitalet de Llobregat, Barcelona, Spain.
11. Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain
12. Cardiovascular and Metabolic Diseases Epidemiology Unit, National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain
13. Public Health Division of Gipuzkoa, BioDonostia Research Institute, Health Department, Basque Country, San Sebastian, Spain
14. Instituto de Biomedicina, Universidad de León, León, Spain.
15. Instituto de Salud Pública de Navarra, IdiSNA, Pamplona, Spain
16. Universidad de Cantabria - IDIVAL, Santander, Spain
17. IUOPA, Universidad de Oviedo, Asturias, Spain
18. Cancer and Public Health Area, FISABIO – Public Health, Valencia, Spain
19. Department of Epidemiology, Regional Health Authority, IMIB-Arrixaca, Murcia University, Murcia, Spain
20. Centro de Investigación en Recursos Naturales, Salud y medio Ambiente (RENSMA), Universidad de Huelva, Huelva, Spain
21. Departamento de Medicina Preventiva y Salud Pública, Facultad de Medicina, Universidad de Granada, Granada, Spain
22. Epidemiology Section, Public Health Division, Department of Health of Madrid, Madrid, Spain

23. IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

Corresponding author:

Prof. Marc Saez, PhD, CStat, CSci
Research Group on Statistics, Econometrics and Health (GRECS)
and CIBER of Epidemiology and Public Health (CIBERESP)
University of Girona
Carrer de la Universitat de Girona 10, Campus de Montilivi
17003 Girona, Spain
Tel 34-972-418338, Fax 34-972-418032
http://www.udg.edu/grecs.htm e-mail: marc.saez@udg.edu

Running title: Compositional analysis of dietary patterns

Abstract

Instead of looking at individual nutrients or foods, dietary pattern analysis has emerged as a promising approach to examining the relationship between diet and health outcomes. Despite dietary patterns being compositional (i.e. usually a higher intake of some foods implies that less of other foods are being consumed), Compositional Data Analysis (CoDA) has not yet been applied in this setting. We describe three CoDA approaches (compositional principal component analysis, balances and principal balances) that enable the extraction of dietary patterns by using control subjects from the Spanish multicase-control (MCC-Spain) study. In particular, principal balances overcome the limitations of purely data-driven or investigator-driven methods and present dietary patterns as trade-offs between eating more of some foods and less of others.

Key words: Compositional data analysis (CoDA), dietary patterns, epidemiology, principal balances, MCC-Spain.

1. Introduction

In nutritional epidemiological studies there is keen interest in identifying specific dietary components that may be related to particular health outcomes. Traditionally, research has focused on single dietary factors (i.e. nutrients, foods or food groups), even though individuals do not consume them in isolation. Thus, in the recent years most studies have shifted to dietary pattern analysis, which better captures overall dietary exposure and allows the cumulative and interactive effects between dietary factors to be evaluated(1,2). The foremost methods for extracting dietary patterns from a given population are *a priori* and *a posteriori* approaches(3). The former are investigator-driven or index-based analyses which use a numerical scoring system defined on the basis of previous scientific evidence. Thus, indexes may differ in design, structure, and interpretation of dietary guidance (e.g, multiple indexes describe adherence to a Mediterranean diet, using different food groups, weightings and cut-offs for recommended intakes(4)), but once there is agreement on which index to use, it eases comparability across populations. The later are data-driven methods that use principal component (PC) (or factor) analysis, cluster analysis, and related techniques, to derive dietary patterns. These patterns are more representative of the eating habits of the study population and, although their applicability to a different setting has been a major concern(1), recent evidence has proven that, under certain conditions, they may be used in different populations(5,6).

Usually, a higher intake of some foods implies that less of other foods are being consumed. In dietary interventions that advocate an increase or decrease of particular foods or nutrients, unless total caloric intake is modified, changes in one dietary component are accompanied by compensatory changes in others. The many food pyramids which have been built represent nothing more than ideal relative amounts of food groups within a total intake. Compositional data analysis (CoDA) is a standard family of statistical methods for analyzing the relative importance of magnitudes, and holds great potential in the context of dietary patterns. Within this family of CoDA methods, in more precise terms we refer to the so-called CoDA log-ratio approach. Although CoDA is a well-established statistical methodology in many scientific fields (e.g. geology, hydrology, or ecology)(7), it has only recently been used in health research. Health-related time-use research constitutes the most frequent application(8–19). However, relative information also lies at the core of the research interest in nutrition(20–23), cause-specific mortality(24), genomics(25,26) and microbiome(27–29). To our knowledge, no study has yet reported its application in the context of dietary patterns.

The aim of this study was to apply and compare three CoDA approaches (compositional PC analysis, balances and principal balances) that enable dietary patterns to be extracted and later used as health outcome predictors. The methods chosen ranged from more data-driven approaches to more investigator-driven ones. We illustrate the methods with data from the

Spanish multicase-control (MMC-Spain) study. For this purpose, we selected a subset of food groups, which are typically used to describe adherence to a Mediterranean diet (MD).

2. Methods

Compositional Data Analysis basics

The use of CoDA started with Aitchison's seminal work(30,31) on chemical and geological compositions in which data are expressed as parts of a whole, commonly with a fixed sum(32). The term compositional analysis(33) was later coined to stress the fact that what is ultimately compositional is not the data, which may not have a fixed sum(8–19) and may not even constitute parts of the same whole or of any whole at all, but the analysis and research objectives which are expressed in terms of relative importance of magnitudes(34). In dietary research, this flexibility makes it possible to combine nutrients and food groups in the same analysis. Trichopoulou's MD index(35), which considers both food groups and fatty acids, is a good example. In the last three decades CoDA has provided a ready-to-use toolbox including software such as the R libraries SpiecEasi, compositions, zCompositions, propr and robCompositions (29,36–39), the stand-alone programs SparCC and CoDaPack (26,40), and accessible handbooks(7,36,41).

Let the composition **x** be a positive vector in a *D*-dimensional real space:

$$\mathbf{x} = (x_1, x_2, ..., x_D) \in R_+^D, \text{ with } x_j > 0 \quad \text{for all } j = 1, 2, ..., D, \tag{1}$$

where *D* is the number of parts, in our case, food groups or nutrients. In order to focus on the relative importance of the parts, the *closure* of **x** to a constant unit sum is common practice.

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

In our case, each subject would have a composition of each of the *D* food groups as proportions of total energy intake, total grams or portions per day or week, or whichever measurement units the data are expressed in. However, closure is by no means required. Regardless of whether closure is performed or not, the relative information carried out by the *D* parts should remain the same, ensuring the so-called *compositional equivalence* property(33).

z resides in an R_+^{D-1} subspace which is constrained by positivity and a fixed sum, called the *simplex*, with different operations, angles and distances from the real space. For this reason, most statistical workhorses such as correlation, variance and Euclidean distance are to a lesser or greater extent meaningless when applied to **z**. This has implications when studying dietary patterns with any correlation-based method, such as PC analysis(42). Finally, when it comes to statistical modelling, distributional assumptions of classical models are violated on **z**(7,43), since

constraints in **z** make it impossible to use unbounded probability distributions such as the normal distribution.

Transformations, association and variance

The most common CoDA approach is to express an original compositional vector of *D* parts into logarithms of ratios among parts or of ratios among geometric means of parts(31,44). There are six main arguments for log-ratios. First, log-ratios are unbounded and, once they have been computed, the normal distribution and other unbounded distributions can be used. Second, standard statistical analyses based on Euclidean geometry in the real space are appropriate. Third, log-ratios are compositionally equivalent, as they yield the same result regardless of whether they are computed from **x** or **z**. Fourth, log-ratios form the basis for defining association, distance and variance in a geometrically meaningful way. Fifth, log-ratios treat the numerator and denominator symmetrically. Sixth, and most important for the purposes of this article, logarithms, ratios and geometric means constitute a natural way of distilling the information about the relative importance of food groups and nutrients within the dietary patterns.

Log-ratios may be computed between each part and the geometric mean of all, in the so-called *centred-log ratios*:

$$\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 \ldots z_D}}\right) = \ln\left(\frac{x_j}{\sqrt[D]{x_1 x_2 \ldots x_D}}\right) \quad \text{with } j = 1, 2, \ldots, D. \qquad \textbf{(3)}$$

A higher centred log-ratio for a given subject on food group j means a higher relative importance of that food group within total intake. The sum of all centred log-ratios for a given subject is zero. Unlike the simple log transform which is commonly used in dietary research, the centred log-ratio of a given food group can only increase if at least some other decreases.

Total variance in a compositional data set is expressed by the sum of variances of all centred log-ratios:

$$\sum_{j=1}^{D} Var\left(\ln\left(\frac{z_j}{\sqrt[D]{z_1 z_2 \ldots z_D}}\right)\right) \qquad \textbf{(4)}$$

Proportionality between pairs of food groups is a valid alternative to correlation(45). The log-ratios between all $D(D\text{-}1)$ possible pairs of parts and their variances are computed for this purpose.

$$\text{Var}(\ln(z_j/z_k)) = \text{Var}(\ln(z_k/z_j)) \text{ with } j,k = 1,\ldots,D;\ j \neq k \qquad \textbf{(5)}$$

These variances can be arranged in a symmetric matrix with parts (i.e. food groups) defining both D rows and D columns, with the same layout as a correlation matrix. It is the so-called variation matrix(7,31). It can be shown that the sum of elements in the variation matrix is $2D$ times the total variance **(4)**. More advanced proportionality measures are available(38).

Var(ln(z_j/z_k)) is zero when z_j and z_k behave perfectly proportionally (e.g. individuals eating twice of one food group also eat twice of the other), which corresponds to perfect positive association. The further Var(ln(z_j/z_k)) is from zero, the lower the association. There is no clearly defined threshold representing lack of association, so that values in the matrix must be assessed comparatively.

A relevant issue in CoDA is the so-called *subcompositional coherence* principle(7). In dietary pattern terms, this concerns the decision on which food groups and nutrients to include in the analysis. Of course, including or excluding a food group does influence the results. However, the results obtained with a set of food groups or with a smaller subset of the former must be mutually coherent. The log-ratio methods in CoDA ensure that:

- Distances between subjects using the full set of food groups are equal or larger than when using a subset.
- Log-ratios and log-ratio variances involving pairs of food groups which are both in the full set and in the subset, are invariant.
- Geometrically speaking, subcompositions constitute an orthogonal projection of the whole composition.

Subcompositional coherence makes it possible to exclude from the analysis food groups which are not relevant to adherence to a particular dietary guidance.

Zero replacement

As it is well known, computing log-ratios implies that **x** and **z** may contain no zero values in any food group intake. Treatment of zeros in CoDA depends on the assumed reason for their occurrence, which is deemed more important than the sheer existence of zeros in itself(46).

On the one hand, there are *absolute, essential or structural zeros*, which represent values that can only be zero given certain characteristics of the individuals (e.g. meat or fish intake in vegetarians). The presence of structural zeros may lead to different covariance structure of the variables of interest, and usually indicates that the choice of parts to be analysed is not meaningful to a certain subpopulation. Thus, data with absolute zeros should be considered as distinct subpopulations and should either be excluded (e.g. by analysing only non-vegetarians) or analysed separately (e.g. by using other dietary scores that better apply to vegetarians).

On the other hand, the so-called *rounded zeros, trace zeros, or zeros below detection limit* constitute parts which are believed to be present, but are not observed due to randomness or limitations of measurement (e.g. a retrospective food frequency questionnaire expressed in weekly portions may fail to record food groups which are consumed less frequently). They are, thus, analogous to missing data with the added information that they are below the detection limit (e.g. the gram equivalent of one portion per week). They can thus be imputed by means of the

EM algorithm if modified in such a way that no imputed value is allowed to be above the detection limit(47).

Extracting compositional information for its use as a predictor in statistical models

The *D* centred log-ratios **(3)** play an important role in distance-based statistical methods such as PC and cluster analysis, but they are not practical as predictors in statistical models for a number of reasons. The fact that they are perfectly collinear is their most often considered disadvantage, although computational solutions do exist(48). A more serious drawback is that each centred log-ratio, and hence each regression coefficient, is related to one particular component, which is not particularly useful when the interest of the researcher lies in dietary patterns as a whole.

Other forms of log-ratios, also called coordinates, are required, which can be interpreted in terms of dietary patterns and on which both geometrical operations and statistical models can be applied in a standard manner in a whole real space matching the (*D*-1)-dimensionality of the simplex(7). This approach is referred to as *working on coordinates* in the CoDA literature(49).

Egozcue *et al.*(44) establish several desirable properties that a set of log-ratios, and hence coordinates, must have in order to be used as variables in further statistical analyses. The most general expression of a log-ratio includes *r* parts in the numerator and *s* parts in the denominator, with possibly different exponents in the numerator ψ_{nj} and in the denominator ψ_{dj}:

$$\ln \frac{\left(x_{n1}^{\psi_{n1}} \ldots x_{nr}^{\psi_{nr}}\right)}{\left(x_{d1}^{\psi_{d1}} \ldots x_{ds}^{\psi_{ds}}\right)} \quad \textbf{(6)}$$

The coordinates fulfilling all desirable properties are the so-called *isometric log-ratios*, or *isometric log-ratio coordinates*, and have the following requirements:

- They must define an orthogonal (*D*-1)-dimensional basis in the simplex.
- The sum of exponents in the numerator of the log-ratio must equal the sum of exponents in the denominator.
- The sum of all squared exponents must be one.

It can be proven that *D*-1 isometric log-ratios capture all information in the compositional data set (44). Total variance of the *D*-1 isometric log-ratios equals the total variance of the *D* centred log-ratios **(4)**. Either the full set of *D*-1 isometric log-ratios or a subset may be used both as dependent and as explanatory variables in any standard statistical model. Using a subset of

these *D*-1 log-ratios is tantamount to an orthogonal projection into a lower-variance subspace. Three different approaches for computing either *D*-1 or a smaller number of isometric log-ratios are presented below.

Compositional PC coordinates

Aitchison(42) extended the well-known data-driven *PC analysis* procedure to the compositional case. The extension boils down to submitting the *D* centred log-ratios **(3)** to an otherwise standard PC analysis of the covariance matrix. *D*-1 PC scores with decreasing variance are extracted, from here on called *PC coordinates*. The PC coordinates are actually log-ratios **(6)** in which positive component loadings are the ψ_{nj} unequal exponents of parts in the numerator and negative loadings are the ψ_{dj} unequal exponents of parts in the denominator. The *D*-1 PC coordinates can be proven to fulfil all conditions for being isometric log-ratio coordinates. Either all *D*-1 PC coordinates or, more commonly, the first few of them explaining most of the variance, can thus be used as variables in further statistical analyses.

Balance coordinates

Isometric log-ratios can also be investigator-driven, on the basis of the investigator's research questions. As a general guideline to find *D*-1 investigator-driven isometric log-ratio coordinates Egozcue and Pawlowsky-Glahn(50) propose *balance coordinates*. Balance coordinates can be easily formed from a *sequential binary partition* (SBP) of parts. To create the first balance

coordinate, the complete composition $\mathbf{x} = (x_1, x_2, ..., x_D)$ 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 ($x_{n1,...,}$ x_{nr}) in the first group are placed in the numerator, and the s_k parts ($x_{d1,...,}x_{ds}$) 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(44):

$$y_k = \sqrt{\frac{r_k s_k}{r_k + s_k}} \ln \frac{\sqrt[r_k]{(x_{n1}...x_{nr})}}{\sqrt[s_k]{(x_{d1}...x_{ds})}} = \ln \frac{(x_{n1}...x_{nr})^{\sqrt{\frac{s_k}{r_k}}{r_k+s_k}}}{(x_{d1}...x_{ds})^{\sqrt{\frac{r_k}{s_k}}{r_k+s_k}}}, \text{ with } k = 1,...,D-1, \quad \textbf{(7)}$$

The corresponding expression **(6)** of the *k*th balance takes equal values $\psi_{nj} = \sqrt{\frac{s_k}{r_k(r_k + s_k)}}$ for all parts in the numerator, equal values, $\psi_{dj} = \sqrt{\frac{r_k}{s_k(r_k + s_k)}}$ for parts appearing in the denominator and $\psi = 0$ for parts appearing nowhere. Positive balance coordinates show a higher relative weight of parts in the numerator, and negative values show the opposite. Normally, all *D*-1 balance coordinates are kept for use as variables in further statistical analyses.

Unlike hierarchical cluster analysis, SBPs and hence balance coordinates are not driven by the data but can be tailored to the research questions of interest. For this purpose, SBPs may be constructed according to conceptual similarity of parts, to theoretically meaningful comparisons of numerator and denominator parts, or to trade-offs between numerator and denominator parts which extant knowledge expects to affect a health outcome. The total variance in the *D*-1 balance coordinates can thus be partitioned into the variance related to the research questions which have driven the construction of the SBP.

Balance coordinates have a visualization tool called the *CoDa-dendrogram*(51), also referred to as the balance dendrogram. It is a depiction of the SBP as a tree diagram. Each balance coordinate is represented on a horizontal axis between the two groups of parts which are divided at the corresponding SBP step. The vertical bar going up from each one of these axes represents the variance of that specific coordinate. The contact point is the coordinate mean, closer to the right set of parts if these parts are relatively more abundant, closer to the left set of parts if this set of parts is relatively more abundant, or just in the middle if the balance coordinate mean is zero. Box plots may be added to represent the balance coordinate medians and quartiles.

Principal balances

PC coordinates are a very efficient tool to compute isometric log-ratio coordinates. The fact that

the first few PC coordinates explain most of the variance makes them especially fit to summarize the composition into few variables for further statistical analyses. However, the PC coordinates obtained can be difficult to interpret as they generally involve all the parts of the composition with irregular ψ exponents(52). Being data driven, such coefficients would be recomputed each time the analysis was rerun on a different data set, thus making comparative research less practical.

On the other hand, balance coordinates compare readily identifiable groups of parts with equal exponents in the numerator and the denominator (actually, the geometric means of numerator and denominator parts) but they require the investigator to provide a SBP. It may prove difficult to provide a theory-driven SBP when *D* is large, or there may also be more than one SBP candidate. At best, selection of the SBP will always remain subjective to some extent. Besides, there is no guarantee that a small number of balance coordinates account for a large proportion of total variance.

The possibility of developing an intermediate approach sharing the best properties of PC coordinates and balance coordinates holds promise. The so-called *principal balances* first suggested by Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado(53), are data-driven balance coordinates, in which a large proportion of variance concentrates on a few coordinates, while comparing readily identifiable groups of parts with easy-to-interpret equal ψ exponents in the numerator and the denominator. These exponents can be easily kept for replication on other data

sets. The first principal balance is defined as the balance coordinate which maximizes explained variance. Subsequent principal balances, being orthogonal to the preceding ones, also maximize the explained remaining variance. Computing principal balances exactly fulfilling this definition requires an exhaustive search along all possible SBPs. A recommended heuristic method is to use the variation matrix among parts as if it was a squared Euclidean distance matrix, and cluster parts based on this matrix with Ward's clustering algorithm(52). The resulting classification tree diagram provides an SBP with balance coordinates which are close to being principal balances, from which, a CoDa-dendrogram can be represented. All *D*-1 principal balances, or, alternatively, those with the highest variance, can be used as variables in subsequent analyses. Alternative computationally intensive methods are described elsewhere(52).

3. Application example

Study population and data preprocessing

The example uses data from the MCC-Spain study, a multicentric case-control study launched to evaluate the influence of environmental exposures and their interaction with genetic factors in four common tumors in Spain. Additional information regarding the study design is provided elsewhere(54). In brief, between September 2008 and December 2013, subjects aged 20-85 with a histologically-confirmed newly-diagnosed cancer were recruited in 23 Spanish hospitals from 12

Spanish provinces. Simultaneously, population-based controls frequency-matched to cases, by age, sex and region were randomly selected from primary care centers within hospitals' catchment areas. For the current analysis, only control population was used. All participants signed an informed consent. Approval for the study was obtained from the ethical review boards of all recruiting centers.

Subjects were provided a semi-quantitative Food Frequency Questionnaire (FFQ), which was a modified version from a previously validated instrument in Spain to include regional products. It included 140 food items, and assessed usual dietary intake during the previous year. A subset of dietary components (in g/day intake) was selected for this illustrative example, based on a common pattern such as the MD as conceptualized by Trichopoulou *et al.* (35). Alcoholic beverages were excluded from our analyses, as they are known risk factors for several chronic diseases (e.g. cardiovascular conditions, cancer). Thus, for the current illustration, the following items were used:

x_1 Vegetables

x_2 Fruits and nuts (fruit)

x_3 Legumes

x_4 Fish and seafood (seafood)

x_5 Cereals

x_6 Meat

x_7 Dairy

x_8 Monounsaturated fats

x_9 Saturated fats

We assumed that respondents reporting no consumption of any type of meat, fish and seafood were vegetarians, and we treated them as absolute zeros by removing them. We then replaced trace zeros. Zero percentages (3.59 % overall; vegetables 0.44 %, fruit 0.99%, legumes 27.16%, seafood 0.96%, cereals 0.91%, meat 0.16%, dairy 1.73%, monounsaturated fats 0%, and saturated fats 0%) were acceptable for replacement with the modified EM algorithm(47). To check for the presence of multivariate outliers in the coordinate vector, squared Mahalanobis distances to the centre can be used(55). After removing cases above the 99.9 percentile of the χ^2 distribution with 8 degrees of freedom (167 cases), a sample size of 3,471 individuals was obtained.

Table 1 shows the variation matrix and centred log-ratio variances of the 9 dietary components. The variation matrix led to the observation that both fat types were those components which behaved most proportionally. This means that most individuals in our sample had either a high or low intake of both fat types and thus, the comparison of the fatty acid profile may contribute little to defining a useful dietary pattern. Similarly, cereals and meat behaved quite proportionally, in spite of the fact that the former represented MD and the latter a Western-like pattern. By contrast,

the highest variances corresponded to legumes, whose consumption tended to move away from that of any other component. This is also depicted by legumes having the highest centred log-ratio variance. The second and third highest variances corresponded to dairy products and fruit.

Table 1: Variation matrix and centred log-ratio variances of the 9 dietary components

Variation matrix **(5)**									
	Vegetables	Fruit	Legumes	Seafood	Cereals	Meat	Dairy	Monoun-saturated fats	Saturated fats
Vegetables	0.000	0.633	2.808	0.538	0.597	0.645	0.939	0.476	0.494
Fruit	0.633	0.000	3.201	0.720	0.714	0.897	0.995	0.669	0.677
Legumes	2.808	3.201	0.000	3.022	3.105	3.052	3.509	3.005	2.906
Seafood	0.538	0.720	3.022	0.000	0.521	0.508	0.942	0.434	0.423
Cereals	0.597	0.714	3.105	0.521	0.000	0.444	0.781	0.324	0.265
Meat	0.645	0.897	3.052	0.508	0.444	0.000	0.883	0.331	0.244
Dairy	0.939	0.995	3.509	0.942	0.781	0.883	0.000	0.725	0.586
Monounsaturated fats	0.476	0.669	3.005	0.434	0.324	0.331	0.725	0.000	0.099
Saturated fats	0.494	0.677	2.906	0.423	0.265	0.244	0.586	0.099	0.000
Centred log-ratio variances **(E4)**									
	0.062	0.096	0.487	0.062	0.053	0.059	0.117	0.036	0.027

PC coordinates

Table 2 shows the PC loadings and percentages of explained variance by PC coordinates. Together, three PCs (1, 2 and 3) accounted for 79.3% of the total variance of the *D* centered log-ratios **(4)** and could be used as a fair summary of diet composition. The first PC coordinate basically reflected the comparison between legumes and the rest of parts, with irregular

coefficients. The second basically compared dairy with seafood, vegetables and fruit. The third PC coordinate balanced meat with dairy and fruit.

Table 2: PC loadings (ψ exponents) and percentages of explained variances by each PC coordinate

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Vegetables	0.053	-0.320	-0.158	0.632	0.584	-0.032	0.123	-0.037
Fruit	0.109	-0.428	-0.654	-0.456	-0.113	0.212	0.025	-0.017
Legumes	-0.939	0.066	-0.002	-0.039	-0.025	-0.009	-0.005	0.013
Seafood	0.104	-0.262	0.138	0.424	-0.761	-0.165	-0.067	-0.017
Cereals	0.136	0.006	0.187	-0.357	0.125	-0.745	0.362	0.079
Meat	0.117	0.029	0.461	-0.111	-0.006	0.604	0.514	0.145
Dairy	0.170	0.791	-0.424	0.182	-0.091	0.021	0.069	0.088
Monounsaturated fats	0.130	-0.003	0.225	-0.125	0.174	0.046	-0.682	0.555
Saturated fats	0.119	0.121	0.227	-0.149	0.112	0.069	-0.338	-0.809
% Variance	55.185	12.216	11.858	6.429	5.566	4.533	3.379	0.835

Balance coordinates

Figure 1 represents the CoDa-dendrogram corresponding to an investigator-driven SBP, in this case an adaptation of Trichopoulou's score of adherence to a Mediterranean dietary pattern(35). At the top of the SBP as the first partition in the dendrogram, y_1 separates food and nutrient groups presumed to fit a MD (x_1 to x_5 and x_8) in the numerator and those not related to a MD(x_6, x_7 and x_9) in the denominator. Thus, the y_1 balance coordinate is a score of adherence to a Mediterranean dietary pattern. If the population is heterogeneous with regard to adherence, this

coordinate should contribute a substantial part of total variance. In addition, below that first balance coordinate (y_1), both conglomerates of food groups are further subdivided sequentially. These subdivisions would address further research questions chosen by the investigator based on knowledge about the particular health outcome of interest. In the example presented here they would concern the importance of the relative intake of meat and dairy (y_3) for a health outcome, the importance of the relative intake of legumes and cereals (y_8), etc.

As PC coordinates do, balance coordinates have an implicit loading matrix following the ψ exponents **(7)**. **Table 3** presents the ψ exponents and the percentage of explained variance by each balance coordinate, which are further shown in **Figure 1.** The first balance coordinate (y_1) representing adherence to a MD explained only a small portion of the variance. This shows that most heterogeneity among eating patterns lies elsewhere, mainly in the balance coordinate opposing the two main types of grains (y_8), and in the balance coordinate between grains and the combination of fruit and vegetables (y_6), as shown by their higher percentages of variance and by the longer vertical segments going up from the balance coordinates in the CoDa-dendrogram in **Figure 1**. Overall, cereals were consumed more than legumes, as shown by the vertical bar representing y_8 closer to the cereal side. Along similar lines, fruit and vegetables were more prevalent than grains, as shown by the vertical bar representing y_6 closer to the fruit and vegetable side.

Table 3: Investigator-driven ψ exponents and percentages of explained variance by each investigator-driven balance

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Vegetables	0.236	0.000	0.000	-0.183	-0.224	0.500	0.707	0.000
Fruit	0.236	0.000	0.000	-0.183	-0.224	0.500	-0.707	0.000
Legumes	0.236	0.000	0.000	-0.183	-0.224	-0.500	0.000	0.707
Seafood	0.236	0.000	0.000	0.913	0.000	0.000	0.000	0.000
Cereals	0.236	0.000	0.000	-0.183	-0.224	-0.500	0.000	-0.707
Meat	-0.471	-0.408	0.707	0.000	0.000	0.000	0.000	0.000
Dairy	-0.471	-0.408	-0.707	0.000	0.000	0.000	0.000	0.000
Monounsaturated fats	0.236	0.000	0.000	-0.183	0.894	0.000	0.000	0.000
Saturated fats	-0.471	0.816	0.000	0.000	0.000	0.000	0.000	0.000
% Variance	11.731	2.837	9.666	7.764	7.482	19.607	6.925	33.988

[Insert Figure 1]

Figure 1: CoDa-dendrogram corresponding to the investigator-driven balance coordinates. Boxplots omitted for simplicity

Principal balances

The ψ exponent matrix and the percentages of explained variance are shown in **Table 4**. The first and second principal balances resembled the first and second PC coordinates. In both cases they were dominated by the ratio of legumes and dairy over most or all of the remaining food groups. The third principal balance is particularly interesting as it compared vegetables, fruit and seafood with cereals, meat and fat. The comparison between monounsaturated and saturated fat

had virtually no variance. Subjects were either eating more of both or less of both in nearly proportional terms (proportionality between these two parts corresponded to the lowest entry in the variation matrix). Together, the three first principal balances accounted for 76.9% of the total variance of the *D* centered log-ratios and could be used as a fair summary of diet composition. The lengths of the vertical bars above each principal balance in the dendrogram in **Figure 2** also show their variance.

Table 4: Principal balance ψ exponents and percentages of explained variance by each principal balance

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
Vegetables	-0.118	-0.134	0.436	-0.408	0.707	0.000	0.000	0.000
Fruit	-0.118	-0.134	0.436	0.816	0.000	0.000	0.000	0.000
Legumes	0.943	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Seafood	-0.118	-0.134	0.436	-0.408	-0.707	0.000	0.000	0.000
Cereals	-0.118	-0.134	-0.327	0.000	0.000	0.866	0.000	0.000
Meat	-0.118	-0.134	-0.327	0.000	0.000	-0.289	0.816	0.000
Dairy	-0.118	0.935	0.000	0.000	0.000	0.000	0.000	0.000
Monounsaturated fats	-0.118	-0.134	-0.327	0.000	0.000	-0.289	-0.408	0.707
Saturated fats	-0.118	-0.134	-0.327	0.000	0.000	-0.289	-0.408	-0.707
%Variance	54.838	11.849	10.175	7.908	5.887	4.421	3.835	1.087

[Insert Figure 2]

Figure 2: CoDa-dendrogram corresponding to the data-driven principal balances. Boxplots omitted for simplicity

Comparison between the compositional dietary patterns

The first investigator-driven balance coordinate had a high multiple correlation with the set of three first principal balances at 0.933. This shows that the first three principal balances contained virtually all the information in the investigator-driven pattern, but not the other way around because correlations between the investigator-driven pattern and each of the first three principal balances were relatively low at 0.601, –0.578, and 0.577, respectively. The fact that correlations were, at best, moderate is in accordance with the fact that the first investigator-driven balance coordinate only accounted for 11.7% of the variance in dietary composition.

In the same vein, the multiple correlation between first investigator-driven balance coordinate and the first three PC coordinates was 0.941. It can thus be argued that PC coordinates and principal balances perform the job of summarizing the information in diet composition almost equally well. Principal balances would be preferable on the grounds that they are easily interpretable and lend themselves more readily to replication and comparison. In fact, the correlations between the first three PC coordinates and their corresponding principal balances were –0.999, 0.868, and –0.706, respectively (it must be noted that the negative signs have no particular implication. All correlations may be turned into positive by reversing the numerator and denominator of one of the coordinates which are being correlated).

4. Comments

This article is the first to compare several CoDA methods to extract compositional information in dietary patterns, in a manner that is appropriate for their later use as predictors of health outcomes. Predictors can be, alternatively, the first balance coordinate, all balance coordinates, the first few PC coordinates, or the first few principal balances. The model and estimation method will be dictated only by the characteristics of the dependent variable and the research design. For instance, natural choices can be a linear model for a continuous health outcome(56), a probit or logit model for an ordered or unordered categorical health outcome(57), and a Cox regression for survival time. Predictors are introduced in a standard manner, the model of choice is estimated with standard software, and standard predictions, diagnostics, residuals and goodness of fit measures can be used.

Interpreting the effects of PC coordinates in a statistical model does not differ from common practice when using standard principal components, once coordinates themselves have been interpreted. The main difference is that PC coordinates in CoDA always imply trade-offs between eating more of some food group(s) and less of other(s), which does not need to be the case in standard principal component analysis.

The effects of balance coordinates or principal balances in a statistical model refer to the impact on the dependent variable when increasing all parts with positive exponents by a common factor and decreasing all parts with negative exponents by another common factor. For instance, a

positive effect of the first balance coordinate y_1 on a health outcome would be interpreted as follows. Increasing vegetable, fruit, legume, seafood, cereal and monounsaturated fat intake all by the same proportion while decreasing meat, dairy and saturated fat intake all by the same proportion is related to a better health outcome.

Thus, in the context of dietary research, CoDA puts emphasis on the fact that any dietary pattern constitutes a trade-off between eating more of some foods and less of others. The relative importance of food groups, nutrients, or a combination of both, lies at the core of the research interest. No pattern derived by CoDA will ever imply eating more of all food groups or less of all food groups. This nicely fits both the intuitive notion of dietary pattern and usual practice in dietary recommendation.

CoDA offers a diverse toolbox which enables researchers to benefit from the best features of data-driven and investigator-driven methods. The closest to being a data-driven approach are PC coordinates extracted from compositional PC analysis, which is carried out in the same way as standard PC, once data have been appropriately transformed. On the other hand, the closest to being an investigator-driven approach are balance coordinates, in other words, log-ratios of geometric means of dietary components which the investigator wishes to compare or relate. The first balance coordinate is an attractive substitute for classical indexes of adherence to the MD such as Trichopoulou's, which are discrete(4). The balance coordinate has i) continuous unbounded distribution that may better lend itself to classic statistical models with, for instance,

normally distributed variables, and ii) computation not reliant on sample-derived medians, tertiles and the like. However, adherence to the MD admittedly does not explain a large proportion of variance in dietary compositional patterns in our control sample. In the opposite extreme, including all *D*-1 investigator-driven balance coordinates as predictors of any health outcome would take into account all variance in dietary patterns at the expense of parsimony.

The recently developed principal balances share some features with data-driven methods and others with investigator-driven methods. Like the investigator-driven method, they can be understood as log-ratios of geometric means of dietary components, with the added attractive property that most of the variance concentrates on a few principal balances, which enables parsimonious models when used as explanatory variables. Like the data-driven method, they can be understood as an equivalent to PC coordinates constrained to have equal loadings, which are easier to interpret and to replicate in comparative research. Rather than suggesting that one approach is superior under all circumstances, each method is designed to answer questions in a different way. When predicting a health outcome, data-driven analysis focuses on the variation in intakes whereas investigator-driven analysis focuses on predefined dietary guidelines. Each approach has unique strengths and limitations, and their relative merits can ultimately depend on how well they predict each particular health outcome. In some cases, the patterns with the highest explanatory power on a health outcome may not be those with the highest variance or those based on previous theoretical knowledge. In such cases, using all *D*-1 balance coordinates

may be more appropriate than using just the first few PC coordinates, principal balances or one or several investigator-driven indexes (11,12,21,24).

As regards limitations of the proposed approaches, it must also be taken into account that, in spite of some attempts(29,58), CoDA is not fully developed for sparse data tables, in other words, those with large proportions of zeros(46), which would be the case when subdividing food groups in great detail (e.g. separated weekly intake of beef, pork, rabbit, lamb, horse, poultry and other meats). The case of structural zeros is also currently underdeveloped in CoDA(59) and it can be problematic to treat nondrinkers, vegetarians, vegans or even subjects of certain religions. Two limitations concern the present study rather than the methods themselves. First, our analysis was based on classical CoDA, but robust alternatives are available(24,39). Second, for simplicity purposes, the illustration has been restricted to the subset of food groups and nutrients which are relevant to MD, but the whole set of groups available from the FFQ could have been used(60).

The goal of dietary pattern analysis is to examine the multiple dimensions of the diet simultaneously relative to a given outcome. In this respect, CoDA provides an interesting alternative perspective. The proposed approaches seem to hold promise for investigating the relationships between dietary patterns and diseases, given the compositional nature of research questions about diet.

Declaration of Conflicting Interests

The authors declare that there is no conflict of interest

Funding

The study was partially funded by the “Accion Transversal del Cancer", approved on the Spanish Ministry Council on the 11th October 2007; by the Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP); by the Instituto de Salud Carlos III-FEDER (PI08/1770, PI08/0533, PI08/1359, PS09/00773-Cantabria, PS09/01286-León, PS09/01903-Valencia, PS09/02078-Huelva, PS09/01662-Granada, PI11/01403, PI11/01810, PI11/02213, PI12/00488, PI14/01219, PI14/0613, PI15/00069, PI15/00914, PI15/01032, PI17/01280, PI09/0914, IJCI-2014-20900); by the Spanish Ministry of Health (CB06/02/1002); by the Spanish Ministry of Economy and Competitiveness (MTM2015-65016-C2-1-R); by the Catalan Government- Agency for Management of University and Research Grants (AGAUR) (2014SGR551, 2017SGR656, 2017SGR733, 2017SGR723, 2017SGR1085); by the University of Girona (MPCUdG2016/069, GDRCompetUdG2017/19); by the Fundación Marqués de Valdecilla (API 10/09); by the Junta de Castilla y León (LE22A10-2); by the Consejería de Salud of the Junta de Andalucía (PI-0571-2009, PI-0306-2011, salud201200057018tra); by the Conselleria de Sanitat of the Generalitat Valenciana (AP_061/10); by the Regional Government of the Basque Country; by the Consejería de Sanidad de la Región de Murcia; by the European Commission (FOOD-CT-2006-036224-HIWATE); by the Spanish Association Against Cancer (AECC)

Scientific Foundation; by the Fundación Caja de Ahorros de Asturias; and by the University of Oviedo. ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

References

1. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. *Curr Opin Lipidol*. 2002; 13(1):3–9.
2. Jacques PF, Tucker KL. Are dietary patterns useful for understanding the role of diet in chronic disease? *Am J Clin Nutr* 2001; 73:1–2.
3. Reedy J, Wirfalt E, Flood A, Mitrou PN, Krebs-Smith SM, Kipnis V, et al. Comparing 3 dietary pattern methods--cluster analysis, factor analysis, and index analysis--With colorectal cancer risk: The NIH-AARP Diet and Health Study. *Am J Epidemiol* 2010; 171(4):479–487.
4. Mila-Villarroel R, Bach-Faig A, Puig J, Puchal A, Farran A, Serra-Majem L, et al. Comparison and evaluation of the reliability of indexes of adherence to the Mediterranean diet. *Public Health Nutr* 2011;14(12A):2338–2345.
5. Castelló A, Buijsse B, Martin M, Ruiz A, Casas AM, Baena-Canada JM, et al. Evaluating the applicability of data-driven dietary patterns to independent samples with a focus on measurement tools for pattern similarity. *J Acad Nutr Diet* 2016;116(12):1914–1924.
6. Castelló A, Lope V, Vioque J, Santamariña C, Pedraz-Pingarrón C, Abad S, et al. Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. *Br J Nutr* 2016; 116:734–742.
7. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. *Modeling and analysis of*

compositional data, Chichester: Wiley, 2015; p.1-247.

8. Pedišić Ž, Dumuid D, Olds TS. Integrating sleep, sedentary behaviour, and physical activity research in the emerging field of time-use epidemiology: definitions, concepts, statistical methods. *Kinesiol Int J Fundam Appl Kinesiol* 2017; 49:10–11.
9. Chastin SFM, Palarea-Albaladejo J, Dontje ML, Skelton DA. Combined efects 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* 2015; 13;10:e0139984.
10. Carson V, Tremblay MS, Chaput J-P, Chastin SFM. Associations between sleep duration, sedentary time, physical activity, and health indicators among Canadian children and youth using compositional analyses. *Appl Physiol Nutr Metab* 2016; 41: 294–302.
11. Dumuid D, Pedišić Ž, Stanford TE, Martín-Fernández J-A, Hron K, Maher CA, et al. The compositional isotemporal substitution model: A method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Stat Methods Med Res* 2017. Epub ahead of print 20 November 2017. DOI: 10.1177/0962280217737805
12. Dumuid D, Stanford TE, Martin-Fernández J-A, Pedišić Ž, Maher CA, Lewis LK, et al. Compositional data analysis for physical activity, sedentary time and sleep research. *Stat Methods Med Res* 2017. Epub ahead of print 30 May 2017. DOI: 10.1177/0962280217710835

13. Dumuid D, Olds T, Lewis LK, Martin-Fernandez JA, Katzmarzyk PT, Barreira T, et al. Health-related quality of life and lifestyle behavior clusters in school-aged children from 12 countries. *J Pediatr* 2017; 183:178–183.

14. Dumuid D, Maher C, Lewis LK, Stanford TE, Martin Fernandez JA, Ratcliffe J, et al. Human development index, children's health-related quality of life and movement behaviors: a compositional data analysis. *Qual Life Res* 2018; 27:1473–1482.

15. Dumuid D, Lewis LK, Olds TS, Maher C, Bondarenko C, Norton L. Relationships between older adults' use of time and cardio-respiratory fitness, obesity and cardio-metabolic risk: a compositional isotemporal substitution analysis. *Maturitas* 2018; 110: 104-110.

16. Hunt T, Williams MT, Olds TS, Dumuid D. Patterns of time use across the chronic obstructive pulmonary disease severity spectrum. *Int J Environ Res Public Health* 2018; 15:533.

17. Foley L, Dumuid D, Atkin AJ, Olds T, Ogilvie D. Patterns of health behaviour associated with active travel: a compositional data analysis. *Int J Behav Nutr Phys Act* 2018;15:26.

18. Talarico R, Janssen I. Compositional associations of time spent in sleep, sedentary behavior and physical activity with obesity measures in children. *Int J Obes* 2018. Epub ahead of print 5 March 2018. DOI: 10.1038/s41366-018-0053-x.

19. Gupta N, Mathiassen SE, Mateu-Figueras G, Heiden M, Hallman DM, Jørgensen MB, Holtermann A. A comparison of standard and compositional data analysis in studies addressing group differences in sedentary behavior and physical activity. *Int J Behav Nutr*

Phys Act 2018;15:53.

20. Decarli A, Ferraroni M. Compositional data analysis and diversity indices: different approaches to define the role of nutrients in the study of diet-cancer relationship. In: [Abstract] *Proceedings from the first International Conference on Dietary Assessment Methods*, S. Paul, Minnesota, 20–23 September 1992. *Am J Clin Nutr* 1994; 59: 306S.
21. Leite MLC. Applying compositional data methodology to nutritional epidemiology. *Stat Methods Med Res* 2016; 25:3057–3065.
22. Leite MLC, Prinelli F. A compositional data perspective on studying the associations between macronutrient balances and diseases. *Eur J Clin Nutr* 2017; 71:1365–1369.
23. Trinh HT, Morais J, Thomas-Agnan C, Simioni, M. Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Stat Methods Med Res* 2018. Epub ahead of print 23 April 2018. DOI: 10.1177/0962280218770223.
24. Mert MC, Filzmoser P, Endel G, Wilbacher I. Compositional data analysis in epidemiology. *Stat Meth Med Res* 2018; 27:1878-91.
25. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics 2018. Epub ahead of print 28 March 2018 DOI: 10.1093/bioinformatics/bty175.
26. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012; 8:e1002687.

27. Pinto JR, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. Balances: a new perspective for microbiome analysis. *bioRxiv* 2017. Epub ahead of print 15 November 2017. DOI: 10.1101/219386.

28. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann of Epidemiol* 2016; 26:330-335.

29. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015; 11: e1004226.

30. Aitchison JA. The statistical analysis of compositional data. *J R Stat Soc Ser B* 1982; 44:139–177.

31. Aitchison JA. *The statistical analysis of compositional data. Monographs on Statistics and Applied Probability*. London: Chapman and Hall, 1986, p.416.

32. Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V. *Compositional data analysis in the geosciences : from theory to practice*. London: Geological Society; 2006, p.212.

33. Barceló-Vidal C, Martín-Fernández JA. The mathematics of compositional analysis. *Austrian J Stat* 2016; 45:57–71.

34. Azevedo Rodrigues L, Daunis-i-Estadella J, Mateu-Figueras G, Thió-Henestrosa S. Flying in compositional morphospaces: Evolution of limb proportions in flying Vertebrates. In: Pawlowsky-Glahn A, Buccianti V (eds) *Compositional Data Analysis: Theory and Applications*. Chichester: John Wiley & Sons, Ltd; 2011. pp. 235–54.

35. Trichopoulou A, Costacou T, Bamia C, Trichopoulos D. Adherence to a Mediterranean diet and survival in a Greek population. *N Engl J Med* 2003; 348:2599–608.

36. van den Boogaart KG, Tolosana-Delgado R. *Analyzing Compositional Data with R*. Heidelberg: Springer, 2013, p. 1–258.

37. Palarea-Albaladejo J, Martín-Fernández JA. ZCompositions - R package for multivariate imputation of left-censored data under a compositional approach. *Chemom Intell Lab Syst* 2015; 143:85–96.

38. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: An R-package for identifying proportionally abundant features using compositional data analysis. *Scientific Reports* 2017; 7:16252.

39. Templ M, Hron K, Filzmoser P. robCompositions: an R-package for robust statistical analysis of compositional data. In: Pawlowsky-Glahn A, Buccianti V (eds) *Compositional Data Analysis: Theory and Applications.* Chichester: John Wiley & Sons, Ltd; 2011, pp. 341–55.

40. Thió-Henestrosa S, Martín-Fernández JA. Dealing with compositional data: The freeware CoDaPack. *Math Geol* 2005; 37:773–793.

41. Pawlowsky-Glahn A, Buccianti V. *Compositional Data Analysis: Theory and Applications*. Chichester: John Wiley & Sons, Ltd, 2011, p. 1–378.

42. Aitchison J. Principal component analysis of compositional data. *Biometrika* 1983; 70:57–65.

43. Aitchison J. Simplicial inference. In: Marlos AGV and Richards DSP (eds) *Algebraic Methods in Statistics and Probability: AMS Special Session on Algebraic Methods in Statistics. Contemporary mathematics Series*. Providence: American Mathematical Society; 2001. pp. 1–22.

44. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. *Math Geol* 2003; 35:279–300.

45. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A valid alternative to correlation for relative data. *PLoS Comput Biol* 2015; 11:e1004075.

46. Martín-Ferńandez JA, Palarea-Albaladejo J, Olea RA. Dealing with zeros. In: Pawlowsky-Glahn A, Buccianti V (eds) *Compositional Data Analysis: Theory and Applications.* Chichester: John Wiley & Sons, Ltd; 2011, pp. 43–58.

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

48. Bruno F, Greco F, Ventrucci M. Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop. *Environ Ecol Stat* 2015; 22:445-63.

49. Mateu-Figueras G, Pawlowsky-Glahn V, Díaz-Barrero JL. The principle of working on coordinates. In: Pawlowsky-Glahn A, Buccianti V (eds) *Compositional Data Analysis: Theory and Applications*. Chichester: John Wiley & Sons, Ltd; 2011, pp. 29–42.

50. Egozcue JJ, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. *Math Geol* 2005; 37:795–828.

51. Pawlowsky-Glahn V, Egozcue JJ. Exploring compositional data with the CoDa-Dendrogram. *Austrian J Stat* 2016; 40:103–113.

52. Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosona-Delgado R. Advances in Principal Balances for Compositional Data. *Math Geosci* 2018; 50:273–298.

53. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Principal Balances. In: Egozcue JJ, Tolosana-Delgado R Ortego MI (eds)*The 4th International Workshop on Compositional Data Analysis CoDaWork2011*, Sant Feliu de Guíxols, Spain, 9-23 May 2011; Girona: University of Girona. pp. 1–10.

54. Castano-Vinyals G, Aragones N, Perez-Gomez B, Martin V, Llorca J, Moreno V, *et al*. Population-based multicase-control study in common tumors in Spain (MCC-Spain): rationale and study design. *Gac Sanit* 2015; 29:308–315.

55. Filzmoser P, Hron K. Outlier detection for compositional data using robust methods. *Math Geosci* 2008; 40:233-48.

56. Tolosana-Delgado R, van den Boogaart KG. Linear models with compositions in R. In: Pawlowsky-Glahn A, Buccianti V (eds) *Compositional Data Analysis: Theory and Applications.* Chichester: John Wiley & Sons, Ltd; 2011, pp. 356–71.

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

58. Mert MC, Filzmoser P, Hron K. Sparse principal balances. *Stat Model An Int J* 2015;

15:159–174.

59. Roel V, Antonio K, Claeskens G. Unraveling the predictive power of telematics data in car insurance pricing. (November 7, 2017). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2872112

60. Castello A, Boldo E, Perez-Gomez B, Lope V, Altzibar JM, Martin V, et al. Adherence to the Western, Prudent and Mediterranean dietary patterns and breast cancer risk: MCC-Spain study. *Maturitas* 2017 ;103:8–15.