Computers & Geosciences 148 (2021) 104621

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

A comparison of isometric and amalgamation logratio balances in compositional data analysis

Michael Greenacre [a,*], Eric Grunsky [b], John Bacon-Shone [c,1]

[a] Department of Economics and Business, Universitat Pompeu Fabra, and Barcelona Graduate School of Economics, Ramon Trias Fargas 25-27, Barcelona, 08005, Spain
[b] Dept of Earth & Environmental Sciences, 200 University Ave. W, Waterloo, Ontario, N2L 3G1, Canada
[c] Social Sciences Research Centre, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Keywords:
Amalgamation
Balance
Explained variance
Geometric mean
Logratio transformation
Subcompositional coherence

ABSTRACT

The isometric logratio transformation, in the form of what has been called a "balance", has been promoted as a way to contrast two groups of parts in a compositional data set by forming ratios of their respective geometric means. This transformation has attractive theoretical properties and hence provides a useful reference, but geometric means are highly affected by parts with small relative values. When a comparison between two groups of parts is required in practical applications, such as the investigation and construction of models, while making use of substantive domain knowledge, it is demonstrated that the logratio of two amalgamations serves as an alternative, interpretable form of balance. A geochemical data set is considered, which has been analyzed previously by transforming to a set of isometric logratio balances. An alternative approach, using a reduced set of pairwise logratios of parts, optionally involving prescribed amalgamations, is very close to optimal in accounting for the variance in this compositional data set. These simpler transformations also have an exact back-transformation to the original parts. This approach highlights for this dataset which compositional parts are driving the data structure, using variables that are easy to interpret and that map well to research-driven objectives.

1. Introduction

Compositional data analysis, known in short as CoDA, is the analysis of nonnegative data that carry only relative information; that is, the values for each sample are either observed as summing to 1 or are transformed to sum to 1 because their total is either fixed (e.g., 100%, 24 h) or not relevant to the research problem. A *composition* is a set of sample values with this unit-sum constraint. Compositional data are observed in many fields: geochemistry (the present context, e.g. mineral compositions), ecology (e.g. species relative abundances), biochemistry (e.g. fatty acid proportions), sociology (e.g. time budgets), geography (e.g. proportions of land use), political science (e.g. voting proportions), marketing (e.g. brand shares), genomics and microbiome research (e.g. proportions of bacterial species). The components of a composition are called its *parts,* and because of the unit-sum constraint, compositions exist in a mathematical simplex: three-part compositions in a triangle, four-part compositions in a tetrahedron, and so on for higher-dimensional simplexes.

In the approach to compositional data analysis by Aitchison (1986), based on data involving J compositional parts, various transformations have been proposed in the form of logarithms of ratios, or *logratios*, which take the sample points out of the simplex into real vector space. The raison d'être of logratios is that they have the property of *subcompositional coherence* (Aitchison, 1986), since they remain constant if parts are added (extending the composition) or removed (forming subcompositions). The original compositional data are not subcompositionally coherent, since their values change after *closing* (i.e. normalizing) the extended or reduced sets of parts to have unit sum.

The simplest examples of logratios are the log-transformed ratios of two parts of a composition, or pairwise logratios – the term "pairwise logratio" is denoted throughout this article by LR. For a J-part composition, with values denoted by $x_1, x_2, \ldots, x_J$, there are ½ $J(J-1)$ unique LRs:

$$\mathrm{LR}(j,j') = \log\left(\frac{x_j}{x_{j'}}\right) \quad j,j' = 1,\ldots,J,\ j<j' \tag{1}$$

* Corresponding author.
E-mail addresses: michael.greenacre@upf.edu (M. Greenacre), egrunsky@gmail.com (E. Grunsky), johnbs@hku.hk (J. Bacon-Shone).

[1] Authorship statement: MG[1] developed the statistical part of the paper and performed the data analyses, in collaboration with EG[2]. EG[2] provided geochemical justification for the statistical ideas and gave geochemical interpretation of the results. JB-S[3] helped with framing of arguments and suggested "what JA would do".

https://doi.org/10.1016/j.cageo.2020.104621
Received 23 June 2019; Received in revised form 1 September 2020; Accepted 24 September 2020
Available online 3 October 2020

Any subset of $J-1$ linearly independent LRs that includes all the compositional parts can generate all the other LRs. The simplest such subset is that of the additive logratios (ALRs), where a specific reference part is contrasted with all the other parts (here the last part is chosen as reference):

$$\mathrm{ALR}(j:J)=\log\left(\frac{x_j}{x_J}\right) \quad j=1,\ldots,J-1 \tag{2}$$

Depending on the chosen reference part, there are exactly J possible ALR subsets, each with $J-1$ elements.

The centered logratio (CLR) transformation is the logratio between each part and the geometric mean of all the parts in the composition. It consists of J elements, treating the parts symmetrically, and is defined as:

$$\mathrm{CLR}(j)=\log\frac{x_j}{\left(\prod_{j'=1}^{J}x_{j'}\right)^{1/J}} \qquad j=1,\ldots,J. \tag{3}$$

The CLRs are not linearly independent and cannot be described as subcompositionally coherent, since they involve all the parts. In fact, they are not intended to be used as alternative variables to represent the parts in each numerator: rather, it is the pairwise differences between CLRs that generate all the LRs. The CLRs serve a very useful computational purpose, for example providing a shortcut to analyzing the complete set of LRs in a biplot (Aitchison and Greenacre, 2002; Appendix A).

The isometric logratio (ILR) transformation, defined by Egozcue et al. (2003), has been promoted by several authors as the mathematically coherent way to express a compositional data set with respect to a new set of $J-1$ orthonormal basis vectors. A particular set of ILRs, called "balances", has been used as alternative variables in statistical analysis – see, for example, Egozcue and Pawlowsky-Glahn (2006), Mateu-Figueras et al. (2011), van den Boogaart and Tolosana-Delgado (2013), Hron et al. (2017) and Martín-Fernández et al. (2018). A single ILR balance contrasts two groups of parts, denoted by J_1 and J_2, as the logratio of their respective geometric means, with a scaling factor (Egozcue et al., 2003):

$$\begin{aligned}\mathrm{ILR}(J_1,J_2) &= \sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}}\log\frac{\left(\prod_{j\in J_1}x_j\right)^{1/|J_1|}}{\left(\prod_{j\in J_2}x_j\right)^{1/|J_2|}}\\ &= \sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}}\left(\frac{1}{|J_1|}\sum_{j\in J_1}\log(x_j)-\frac{1}{|J_2|}\sum_{j\in J_2}\log(x_j)\right)\end{aligned} \tag{4}$$

where $|J_1|$ and $|J_2|$ denote the number of numerator parts and denominator parts respectively. A particular set of $J-1$ linearly independent ILR balances can be obtained by hierarchical clustering, or recursive partitioning, of the parts (Egozcue and Pawlowsky-Glahn, 2005).

A special case of ILR balances is a set of pivot logratios (PLRs), which are a succession of $J-1$ balances where the numerator in the ratio is a single part and the denominator all those parts "to the right" in the ordered list of parts:

$$\begin{aligned}\mathrm{PLR}(j) &= \sqrt{\frac{|J_2|}{1+|J_2|}}\log\frac{x_j}{\left(\prod_{j'\in J_2}x_{j'}\right)^{1/|J_2|}}\\ &= \sqrt{\frac{|J_2|}{1+|J_2|}}\left(\log(x_j)-\frac{1}{|J_2|}\sum_{j'\in J_2}\log(x_{j'})\right)\end{aligned} \tag{5}$$

where $j=1,\ldots,J-1$ and J_2 is the set of parts $J_2=\{j+1, j+2, \ldots, J\}$ (Hron et al., 2017; Filzmoser et al., 2018). A PLR, with its single part in the numerator, has the advantage of being able to be expressed as an average of LRs where the numerator is the same for each LR. For example, the first PLR balance is, apart from the scalar multiplier, equal to $[\log(x_1/x_2)+\log(x_1/x_3)+\cdots+\log(x_1/x_J)]/(J-1)$. When it comes to choosing a set of balances, the PLR versions are the "first option" of Filzmoser et al. (2018, page 35).

Logratios of amalgamations of parts have not been widely used, although – paradoxically – parts used in compositional data analysis are themselves often pre-defined as amalgamations. Denoted here by SLR (standing for "summated logratio"), an amalgamation logratio is an alternative form of balance, defined more simply than (4) and (5) above as:

$$\mathrm{SLR}(J_1,J_2)=\log\frac{\sum_{j\in J_1}x_j}{\sum_{j\in J_2}x_j} \tag{6}$$

We call this an SLR balance for short, and it is a logratio, without any scaling factor, just like any other pairwise logratio of two parts. Amalgamations are often performed in chemistry based on the understanding of the stoichiometric balances, that is, variables that are part of a mineral composition and are additive in the simplex. This is evident when evaluating the molar relationships of elements within individual minerals. The number of moles that constitute the contribution of an element to a mineral is not based on compositions, but on the controlled placement of elements within the lattice structure of a mineral. Egozcue et al. (2013) indicate that within mineral structures, elements can substitute for each other and this implies that amalgamation is a reasonable action. Grunsky et al. (2008), Grunsky and Bacon-Shone (2011), Caritat and Grunsky (2013), and Grunsky and Kjarsgaard (2016) highlight the evidence of stoichiometric relationships in compositional data based on multi-element associations in principal component biplots that are based on the CLR transform. For example, the structure of plagioclase is complicated by the fact that the ratio of Si cations changes as Na and Ca substitute for each other in the crystal structure and can also contain up to 10% K, through a lattice transition and substitution with Na and Ca. The element associations in the biplots correspond with known element associations for specific minerals in geochemical datasets that are expressed in compositional form.

Like ILR balances, SLR balances are subcompositionally coherent with respect to adding parts to the composition, or removing parts that are not in the numerator or denominator groups. To show this, simply divide the sums in the numerator and the denominator by any part that occurs in the SLR and deduce the coherence from the fact that all ratios are subcompositionally coherent. Just like ILR balances, a set of $J-1$ linearly independent SLR balances involving all the parts can be inverted back to the original J part values, by solving a set of linear equations (Greenacre, 2020).

In this study the following questions are considered:

1. What is the meaning and the interpretation of a single ILR balance and a single SLR balance as univariate statistics? What do their values measure in practice?
2. What are the advantages and disadvantages of ILR balances?
3. What are the advantages and disadvantages of using SLR balances?

Section 2 describes a real geochemical data set used to illustrate the concepts and results, as well as the methodologies and software used and Section 3 gives the results for this data set. Sections 4 and 5 follow with a discussion and overall conclusion about the above-mentioned questions. Supplementary material is supplied, including additional tables, figures and code in R (R Core Team, 2019).

2. Material and methods

2.1. The Aar Massif data set

This 10-part data set consists of geochemical compositions of the major oxides (SiO_2, TiO_2, Al_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5,

Fe_2O_3t) in 87 samples of glacial sediment in the Aar Massif, Switzerland (Tolosana-Delgado and von Eynatten, 2010), analyzed by van den Boogaart and Tolosana-Delgado (2013) and Martín-Fernández et al. (2018). The data are freely available in the compositions package (van den Boogaart et al., 2019) in R (R Core Team, 2019). These oxides have average percentages as low as 0.06% (MnO) and as high as 70.81% (SiO_2). The objective is to describe the patterns in the oxides in a meaningful and interpretable way, including the following three groupings of oxides based on geochemical considerations:

mafic: MgO, Fe_2O_3t, MnO

felsic: Na_2O, SiO_2, Al_2O_3, K_2O

carbonate-apatite: CaO, P_2O_5

(TiO_2 is not included in any group and thus forms a group of its own). Soils, sediments, igneous and metamorphic rocks are comprised, in whole or in part, of minerals. Minerals form under conditions governed by thermodynamics (temperature and pressure) and the bonds that the various elements form within a rigid framework, which govern mineral stoichiometry. Each mineral has a different stoichiometric form. Combining the chemistry of minerals in varying abundances will yield bulk geochemical signatures that represent a linear combination of the stoichiometric framework of the minerals. Our choice of combining the element oxides into three groups presents a provisional model of the geochemistry for the purposes of interpretation of various rock types (mafic, felsic, carbonate-apatite).

2.2. Methods

Apart from some standard statistical methods, the approach here focuses on the analysis of logratios of amalgamated parts (SLR balances), compared to the use of logratios of geometric means of parts (ILR balances). Comparisons are made in terms of (i) measurement, substantive meaning and interpretation, (ii) logratio variance explained, (iii) identification of ratios that account for the data structure, (iv) Procrustes correlation and (v) principal component analysis (PCA) of different forms of logratios.

2.2.1. Measurement, substantive meaning and interpretation

Here the scales of these two types of logratio balance are examined, namely what each balance is actually measuring and the substantive meaning of their values. The interpretation of the balances is examined using specific examples from the Aar Massif data set.

2.2.2. Explained logratio variance

Greenacre (2019, Section 2.2) gives a detailed description of the measure of total variance in a compositional data set, and also explains the difference between this definition and two other historical definitions, that of Aitchison (1986), which was modified by Pawlowsky-Glahn et al. (2007). To clear up misunderstandings highlighted by a reviewer, this definition is repeated here, with additional justification.

The *(weighted) logratio variance* of a compositional data set, with optional weights c_j ($j = 1, \ldots, J$) for the parts, is equal to the weighted average of the variances of the ½ J ($J - 1$) LRs, where $\mathrm{LR}(j, j') = \log(x_j/x_{j'})$ has weight $c_j\, c_{j'}$. There is no need to compute all the LRs since exactly the same quantity is obtained from the weighted average of the variances of CLRs (Greenacre, 2018, 2019), where CLR (j) has weight c_j:

$$\text{logratio variance} = \sum\sum_{j<j'} c_j c_{j'} \,\mathrm{var}[\mathrm{LR}(j,j')] = \sum_j c_j \mathrm{var}[\mathrm{CLR}(j)] \tag{7}$$

In the present application, weights for the parts (columns) will all be equal: $c_j = 1/J$ for all j, so that $c_j\, c_{j'} = 1/J^2$. Furthermore, the variance across samples is preferably computed dividing sums of squares by I and not by $I - 1$, which means that each sample (row) is equally weighted by a weight $r_i = 1/I$. Averaging is preferred here, as in Greenacre (2018, 2019), compared to the usual summed options of Aitchison (1986) and Pawlowsky-Glahn et al. (2007), although it makes no difference to percentages of explained variance that are computed subsequently, since the various definitions differ only by scalar multiples. The advantage of using the averaging is that it makes measures of logratio variance compatible across studies of different sizes, and also anticipates the introduction of differential weights for the parts (see the applications in Greenacre (2018, 2019) where weighting is used). Another advantage is that the result (7) above is equivalent to computing the average of all the squared elements of the double-centered matrix **Y** of log-transformed compositional data. When both samples and parts are equally weighted, this double-centered matrix is the following:

$$\mathbf{Y} = \left(\mathbf{I} - (1\,/\,I)\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\log(\mathbf{X})\left(\mathbf{I} - (1\,/\,J)\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) \tag{8}$$

where **1** denotes a vector of the appropriate number of ones in each case (i.e. I ones on the left for the column-centering and J ones on the right for the row-centering). Then the logratio variance in this case is equivalently:

$$\text{logratio variance} = \mathrm{trace}\left(\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\right)/(I\ J) \tag{9}$$

(these three equivalent ways of computing the logratio variance are performed in the accompanying R script).

Additionally, parts that are not deemed relevant to the research question, or do not contribute to the structure of the data, can be eliminated. This is, in effect, a subcomposition, but one that is directly related to the investigation of compositional structure or the research question. Given any explanatory variables, the amount and percentage of explained logratio variance can be computed in a laborious way by regressing each of the J CLRs on these variables, obtaining the parts of variance explained in each case (notice that all J CLRs have to be regressed, even though one is redundant, since they all count towards the total variance in (7)). A much shorter way is to realize that this set of regressions is encapsulated in the method of *redundancy analysis* (van den Wollenberg, 1977), or RDA. RDA can be used to obtain the percentage of explained variance in a simple matrix computation, where the J CLRs are projected in one operation onto the space of the explanatory variables.

The present application uses explanatory variables in the form of the LRs, ILR balances or SLR balances, so that the approach involves quantifying how much logratio variance is explained by a subset of logratios of one of the above three types, and then comparing the corresponding results – see Greenacre (2019).

2.2.3. Selecting LRs (pairwise logratios) to identify parts that explain data structure

To find a subset of LRs, Greenacre (2019), inspired by the work of Krzanowski (1987) on variable selection, proposed a stepwise process where LRs are selected that explain a maximum part of the logratio variance at each step. Identifying such a subset of LRs implies identifying a subcomposition of parts (i.e. those used in the LRs) that are the main drivers of the patterns in the data. Amalgamations that are pre-defined by the practitioner in the form of knowledge-driven groupings of the parts, can be included as candidates for creating LRs, since their amalgamated parts are simply considered to be additional parts. This means that SLR balances are similarly considered as LRs in the variable selection process.

The stepwise procedure starts by first finding the LR that explains the maximum variance, then the one that adds the most explained variance to the first, and so on, described more fully by Greenacre (2018, 2019). The percentages of variance explained show how well these chosen subsets of LRs can serve as alternative variables to represent the compositional data set.

2.2.4. Procrustes correlation

The *logratio distance* between samples is defined explicitly by Greenacre (2018, 2019) and, like the logratio variance, differs from the Aitchison distance by a scalar multiple, being the result of an averaging rather than a summing. The inter-sample logratio distances can be displayed exactly in a ($J-1$)-dimensional Euclidean space, where their interpoint distances match the Euclidean distances based on all the ½ J ($J-1$) LRs (Aitchison and Greenacre, 2002). In order to see how closely the multivariate structure of these logratio distances between the samples, based on all the LRs, can be approximated by a smaller set of logratios, possibly including amalgamations, the Procrustes correlation between the sample positions in the respective spaces is computed (Krzanowski, 1987; Gower and Dijksterhuis, 2004; Legendre and Legendre, 2012, page 704) – see Greenacre (2019, Appendix) for the mathematical definition.

2.2.5. Principal component analysis of logratios and logratio analysis

In order to visualize the structure of compositional data, logratio analysis (LRA) is used to reduce the dimensionality of the data (Greenacre and Lewi, 2009; Greenacre, 2009, 2010a, 2011, 2018, 2019). LRA is defined as the PCA of all the LRs but can be performed more efficiently as a PCA of the CLRs. The result is shown as a biplot, where the biplot scaling option called the "form biplot" (see Aitchison and Greenacre, 2002) is used. The positions of the samples are such that their interpoint distances approximate the logratio distances. The individual parts are shown (these are the CLRs), but the interpretation is focused on the ½ J ($J-1$) links connecting pairs of parts. These links represent the biplot axes of the respective LRs onto which sample points can be projected. Although LRA is computed by a PCA of the CLR matrix, it should be thought of as the PCA of the matrix of LRs. When a reduced subset of LRs is selected, which in our application can include SLR balances, its structure is visualized and interpreted using PCA. LRA gives the same results for the samples and the same dimensional percentages of variance as the PCA of a full set of $J-1$ linearly independent ILR balances.

3. Results

3.1. Scale, meaning and interpretation

Here it is attempted to understand an ILR balance and an SLR balance when used as single variables, if they appear as explanatory variables and in multivariate analyses and thus beg interpretation. These aspects are best discussed around an example. The three groups of parts, mafic, felsic and carbonate-apatite, involve 3, 4 and 2 parts respectively. The ratio between mafic and carbonate-apatite, the two groupings with the lesser numbers of parts, is chosen as an example. The ILR balance, computed for one of the Aar Massif samples (specifically, sample 66), is (apart from the constant $\sqrt{6/5}$) equal to the average of the six LRs defined by the three numerator parts combined with the two denominator parts:

$$\begin{aligned}\text{ILR}:&\frac{1}{6}\left[\log\left(\frac{\text{MgO}}{\text{CaO}}\right)+\log\left(\frac{\text{Fe}_2\text{O}_3\text{t}}{\text{CaO}}\right)+\log\left(\frac{\text{MnO}}{\text{CaO}}\right)+\log\left(\frac{\text{MgO}}{\text{P}_2\text{O}_5}\right)\right.\\&\left.+\log\left(\frac{\text{Fe}_2\text{O}_3\text{t}}{\text{P}_2\text{O}_5}\right)+\log\left(\frac{\text{MnO}}{\text{P}_2\text{O}_5}\right)\right]\\&=\frac{1}{6}[-0.934-0.124-3.962+2.079+2.890-0.948]\\&=-0.167\end{aligned}$$

On the other hand, the SLR balance is a pairwise logratio made up of the summated values of the two respective groups:

$$\text{SLR}: \log\left(\frac{\text{MgO}+\text{Fe}_2\text{O}_3\text{t}+\text{MnO}}{\text{CaO}+\text{P}_2\text{O}_5}\right)=0.211$$

The scale, meaning and interpretation of the SLR balance is clear. The value of 0.211 corresponds to a ratio larger than 1, since the value 0.211 is positive, in fact the ratio is $e^{0.211}=1.235$, so there are about 24% more mafic parts than carbonate-apatite parts. A zero value for an SLR balance means that the sum of mafic parts equals the sum of carbonate-apatite parts, i.e. when the two groups of parts are truly balanced. The direct meaning of the ILR balance as an average of LRs is also clear, but can be less obvious to interpret. From the negative value of the ILR one might infer, incorrectly, that there were less mafic than carbonate-apatite parts. Of course, many different values of the numerator and denominator parts could likewise give an SLR balance of zero, but this is arguably more straightforward to interpret because of the fact that the two groups are considered as single agglomerated entities, just like two parts. If there is any interesting variation between the parts inside a group, this would be captured by LRs of parts or other SLR balances within the group.

3.2. Ratio selection, including ratios of amalgamations

For the Aar Massif data, the three amalgamations of mafic, felsic and carbonate-apatite, defined in Sect. 2.1, were also used to form LRs (i.e. SLR balances in this case) with the oxides or with the other amalgamations in the search for the set of LRs that maximized the explained variance of the compositional data set. Table 1 shows the 9 selected ratios, their cumulative explained variance, and the Procrustes correlations of the sample configurations with the exact sample configuration. Fig. 1 shows a graph of the solution, which involves felsic and carbonate-apatite but not mafic. The explained variance was 99.997%, only 0.003% short of 100%, with a Procrustes correlation of 0.993 between the sample configuration using the selected set of nine LRs and the exact sample configuration using all 45 original LRs. An even smaller subset of ratios can be considered: for example, with only four LRs more than 95% of the logratio variance is explained, with a Procrustes correlation of 0.976.

Fig. 2a shows the two-dimensional structure of the compositional data set, using LRA (i.e. the PCA of all 45 LRs, equivalent to the PCA of the CLRs). Fig. 2b shows the two-dimensional structure obtained from the PCA of the reduced set of nine LRs. The similarity in the configurations of the samples is clear, due to the Procrustes correlation that is almost 1. Notice that the Procrustes correlation of 0.993 is computed between the sample configurations in their respective full 9-dimensional spaces. If the Procrustes correlation is computed between the two-dimensional configurations in Fig. 2a and b, the correlation increases to 0.997 (the computations are given in the supplementary material).

3.3. Knowledge-driven intervention in the stepwise process

The completely automatic stepwise process, giving the results in Table 1, Figs. 1 and 2b, chooses the LR that gives the highest additional explained logratio variance at each step. In fact, there are several LRs competing for entry with very little difference in their explained variances. This opens the opportunity for the geoscientist to intervene in the process and choose an LR that is almost as good as the optimal one, but which is more meaningful in terms of describing the chemical processes.

As an example, the amalgamation mafic did not enter the stepwise process (Table 1 and Fig. 1), but its components MgO, Fe_2O_3t and MnO are clearly aligned in Fig. 2a and opposing the felsic parts Na_2O, SiO_2, Al_2O_3, K_2O. From the opposing positions of MgO and Na_2O in Fig. 2a, defining a long link between them and thus a high contribution to variance (Greenacre, 2013), it is no surprise that MgO/Na_2O is the ratio of choice in the first step of the algorithm. This optimal LR of a mafic part with respect to a felsic part has a maximum explained variance of

Table 1
The ratios that maximize additional variance explained at each step, their cumulative explained variance and Procrustes correlation with the exact logratio geometry.

	RATIO	Cum % of var.expl.	Procrustes correlation
1.	MgO/Na_2O	69.1	0.831
2.	K_2O/P_2O_5	89.3	0.944
3.	SiO_2/K_2O	93.4	0.962
4.	TiO_2/Na_2O	96.6	0.976
5.	SiO_2/Na_2O	98.7	0.984
6.	Felsic/Carb-Apat	99.3	0.986
7.	MnO/Carb-Apat	99.8	0.989
8.	Al_2O_3/MgO	99.9	0.991
9.	TiO_2/Fe_2O_3t	100.0	0.993

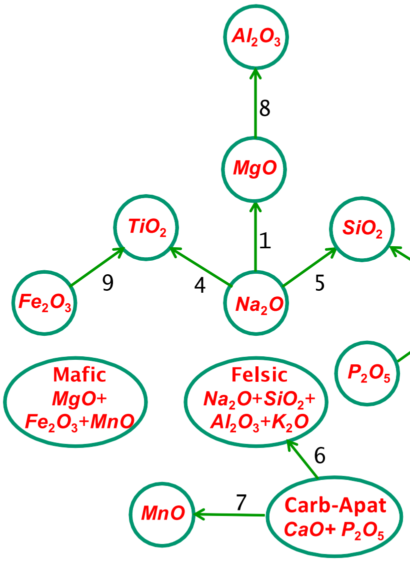

Fig. 1. Graph of the ratios in Table 1. The arrows point to the numerator of each ratio. The numbers refer to the ordering of the steps in Table 1. The Mafic amalgamation does not enter into any ratio in this solution.

69.1%, but in fact there were many such ratios contrasting mafic and felsic parts competing to enter, including the respective amalgamations, as shown by the top 10 ratios for entering at the first step (Table 2).

The ratio mafic/felsic contrast is of interest because, based on the geochemistry of igneous and metamorphic rocks, it is one of a few ratios that quantify the possible mineralogical combinations that might exist. Rather than the optimal pairwise ratio MgO/Na_2O entering, it is preferred that the logratio of mafic/felsic enters, which explains only 0.3% less than the optimal logratio (Table 2). After selecting this ratio as the first one, and then letting the stepwise process take its automatic course afterwards, a slightly different selection of logratios is obtained, explaining 99.994% of the logratio variance, and with a Procrustes correlation of 0.990, both values fractionally less than in Sect. 3.2. The PCA of this alternative set of logratios is shown in Fig. 3, where the configuration of samples is again practically identical to those in Fig. 2.

3.4. *Comparison of best single logratios of different types*

It is instructive to compare the best single logratios from different solutions, where "best" is measured in terms of highest percentage of explained logratio variance. The highest, by construction, is that obtained by the first principal component of the CLRs, which is an ILR involving non-integer powers of the parts (i.e. not a balance). In descending order, the best logratios are as follows, along with the number of parts involved in each case:

- first principal component: 71.2% (all 10 parts)
- first principal balance identified in Martín-Fernández et al. (2018): 70.2% (9 parts − CaO excluded)
- first LR of MgO/Na_2O in Table 1: 69.1% (2 parts)
- CLR of Na_2O: 68.6% (all 10 parts)
- PLR of Na_2O versus the other oxides: 68.6% (all 10 parts)

Notice that the CLR and the PLR of Na_2O versus the rest have identical explanatory power because they differ only by a scaling factor.

The single LR of MgO/Na_2O, involving only two parts, compares very favorably with the others in terms of percentage of explained logratio variance, which involve the complete set of 10 parts (except the first principal balance, which includes 9 parts). This LR, found with minimal computational effort (see Greenacre, 2019 for algorithmic timing issues), explains only 1.1 percentage points less than the first principal

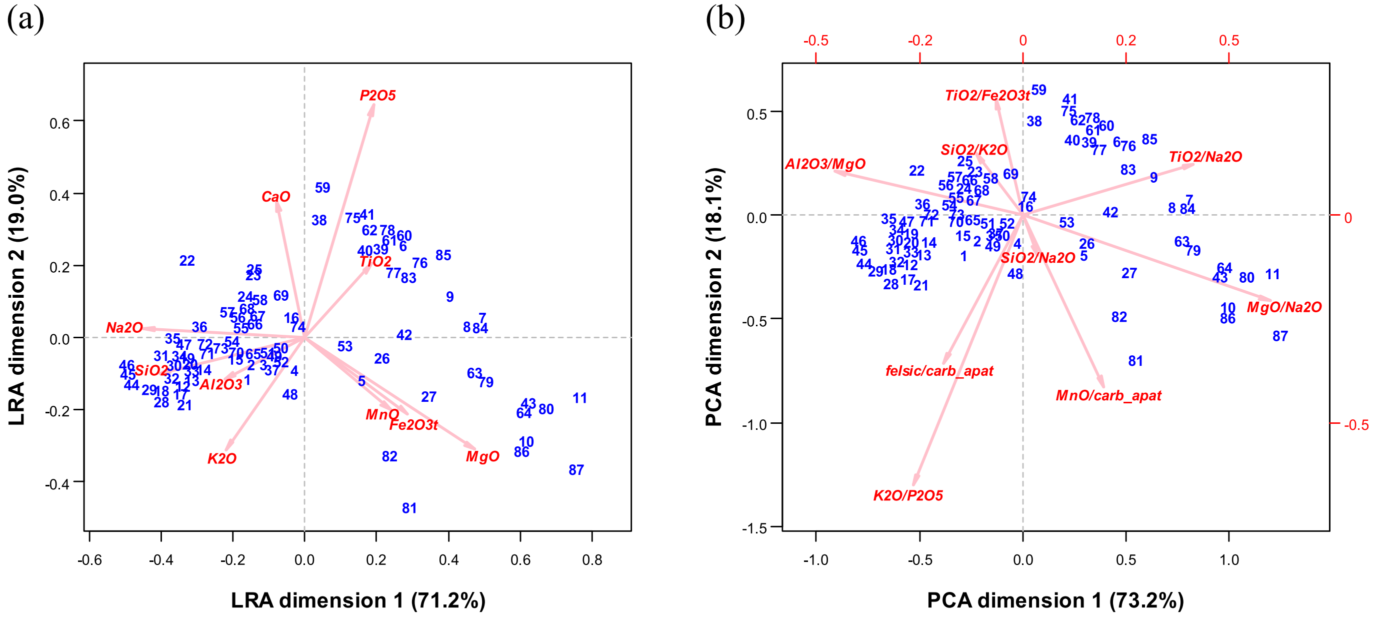

Fig. 2. (a) Logratio analysis (LRA) of the Aar massif data set; (b) PCA of the nine selected logratios. The contribution biplot scaling is used.

Table 2
The top 10 ratios competing to enter in the first step of the logratio selection process, showing their explained variances in descending order and Procrustes correlations.

	RATIO	Cum % of var.expl.	Procrustes correlation
1.	MgO/Na2O	69.1	0.831
2.	mafic/Na2O	69.0	0.831
3.	MnO/felsic	68.9	0.830
4.	mafic/felsic	68.8	0.829
5.	mafic/Al2O3	68.6	0.829
6.	Fe2O3/felsic	68.6	0.828
7.	Fe2O3/Na2O	68.6	0.828
8.	Fe2O3/Al2O3	68.1	0.825
9.	MgO/felsic	67.8	0.824
10.	MgO/Al2O3	67.7	0.823

balance, which involves an exhaustive and costly search algorithm. This good behavior of simple LRs has been found in different applications, for example Greenacre (2018, 2019) and Graeve and Greenacre (2020).

Notice that we are not interested in investigating the very large number of potential SLR balances, as Martín-Fernández et al. (2018) do in their search for principal balances. Our approach is rather to use knowledge-driven amalgamations chosen by experts, who choose groupings of parts that make substantive sense, in this case geochemical sense.

4. Discussion

Various articles and books on compositional data analysis, such as Mateu-Figueras et al. (2011), Pawlowsky-Glahn et al. (2015), Fačevicová et al. (2016), Kynčlová et al., (2017) and Filzmoser et al. (2018, page 35), maintain that using ILR balances, or at least transformations to orthonormal coordinates, is strongly recommended for further statistical analysis. By contrast, it is remarkable that Aitchison himself, when referring to the isometric logratio transformation, referred to its useful theoretical properties and "the elegance of the algebraic-geometric (Hilbert space) structure of the simplex", but cautioned that "it is easy to fall into the pure-mathematical trap that all compositional problems must depend on this structure, that all statistical problems should be addressed in terms of coordinates associated with orthonormal, isometric bases" (Aitchison, 2008, page 20).

There are two main claimed benefits of ILR balances: first, the definition of a new set of coordinates for the data, which preserve the "Aitchison geometry"; and second, their role in contrasting two groups of parts. These two aspects are considered in turn.

Concerning the geometry, a set of $J-1$ independent ILR balances needs to be defined in order to provide a new set of coordinates for a J-part compositional data set. These then provide an isometric transformation of the compositional data to a $(J-1)$-dimensional vector space defined by the ILR coordinates. The logratio distances between the samples are identical to the Euclidean distances between the ILR coordinates, which is a favourable property of its definition, since exactly $J-1$ ILR balances serve to reproduce the $(J-1)$-dimensional geometry perfectly.

However, it is not necessary to satisfy perfectly this exact reproduction of the sample space by the ILR coordinates. For example, when one performs a logratio analysis, as in Fig. 2a where just the first two dimensions are displayed, this is intentionally approximating the geometry by attempting to separate non-random from random variation in the compositional data set, effectively discarding the minor dimensions. These lesser components may represent either random effects or undersampled processes (Grunsky and Kjarsgaard, 2016). So it seems reasonable that some non-informative variability in the compositional data set be removed by appropriate and meaningful transformations and variable selection. As proposed by Krzanowski (1987), selecting key variables that explain almost 100% of the variance and that closely approximate the original sample configuration using much fewer variables presents an easier alternative for the practitioner.

The use of the term "orthonormal coordinates" for the ILR balances in this context is widespread – see, for example, van den Boogaart and Tolosana-Delgado (2013, p. 45) – but it is confusing: the ILR balances are themselves not orthonormal, nor are they orthogonal, they are rather

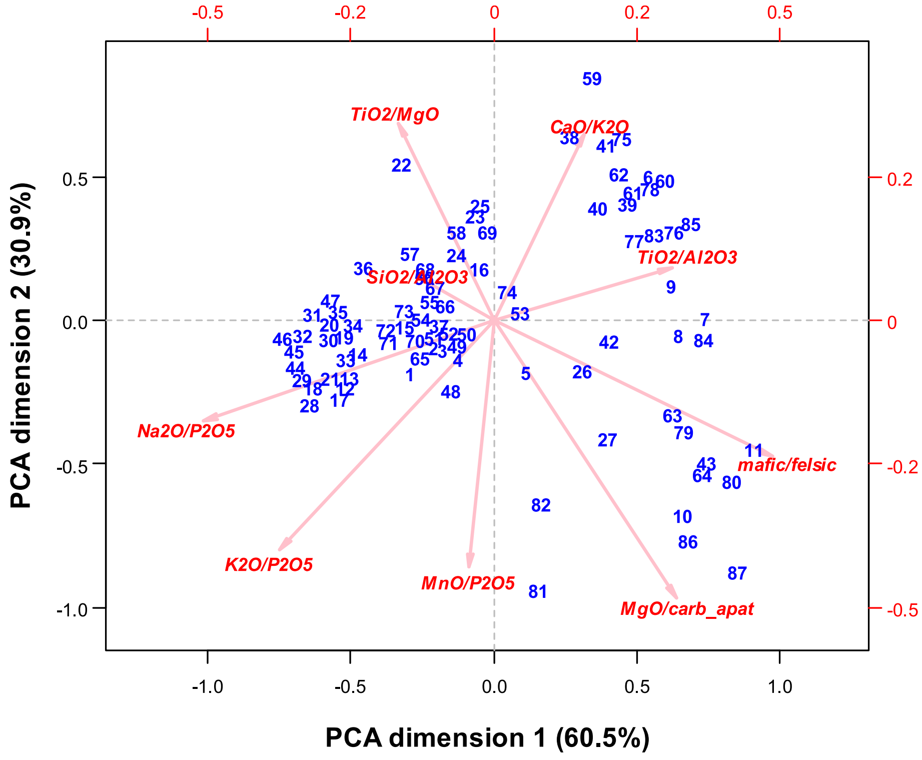

Fig. 3. PCA biplot of the 9 logratios selected after the Mafic/Felsic logratio is chosen at the first step.

coordinates with respect to an orthonormal basis. This is different from the principal coordinates of the samples (i.e. PCA "scores") with respect to the principal component basis vectors: the former are orthogonal, being scaled versions of the left singular vectors of a singular-value decomposition, and the latter are orthonormal, being the right singular vectors – see, for example, Greenacre (2010b). SLR balances are also not orthogonal and they are not even defined with respect to an orthonormal basis – they are data transformations based on expert knowledge. When it comes to the claimed benefit that ILR balances are "easily interpreted in terms of grouped parts of a composition" (Pawlowsky-Glahn et al., 2015, p. 38), this claim is arguable, as an ILR balance is actually a log-contrast formed from the average of multiple LRs.

For an example of the interpretational difficulties, we examine an election example in Pawlowsky-Glahn et al. (2015, p. 41), where the authors give contrasting election percentages for four leftwing parties in the numerator with those of two rightwing parties in the denominator:

$$\sqrt{\frac{4\ \cdot\ 2}{4+2}}\log\frac{(x_1x_2x_5x_6)^{1/4}}{(x_3x_4)^{1/2}}$$

stating that "if someone is interested in knowing which wing has obtained more votes and in evaluating their relative difference, the [above] balance between the left group versus the right group (…) provides this quantitative information" and "the sign of the balance points out which group obtained more votes, and the value gives the size of the difference in log relative scale". Even the simple example of the four leftwing parties obtaining 15% each of the votes, and the rightwing 20% each, contradicts the above assertion, which would conclude that the rightwing won the election whereas the leftwing obtained 60% of the votes.

When it comes to combining parts, the specialist often has knowledge about the possible models that the empirical relationships might reveal and amalgamations can make use of this knowledge. Moreover, amalgamations can be applied when there are problems with the number of degrees of freedom (e.g. too many parts and too few samples in a modeling exercise) and when a preliminary examination of the data suggests that some amalgamations are useful. They can also partially solve the problem of zeros in compositional data, when parts with zeros are meaningfully combined with other parts, such as with hierarchical problems.

Logratios of geometric means (i.e. ILR balances) and logratios of amalgamations (i.e. SLR balances) have been rarely compared in the literature. An exception is Mateu-Figueras and Daunis-i-Estadella (2008), who compare SLR and ILR balances for a five-part data set, performing an ANOVA in each case between the means of three pre-defined sample groups. The mean of the SLR balance is found to be non-significant between groups, whereas the mean of the ILR is significant. The significant group difference in the case of the ILR balance is conditional on the groupings of parts being justifiable according to the objectives of the study, and requires further investigation of the balance's component LRs – for example, the presence of a rare part can radically affect the values of the ILR balance, while hardly affecting the SLR balance.

An additional benefit of the transformation to a set of ILR balances is that it reduces the J-part data set of rank $J-1$ to one of $J-1$ variables that are linearly independent, and whose covariance matrix is easily inverted in matrix computations such as multiple regression and Mahalanobis distance (Egozcue and Pawlowsky-Glahn, 2005). This would be an advantage when using standard software in computer packages that assume the covariance matrix to be nonsingular, but in a computing environment such as R (R Core Team, 2019), the generalized inverse can easily be used directly on the singular covariance matrix of the J CLRs, for example, giving identical results. Moreover, any set of ALRs, or – more generally – any linearly independent set of $J-1$ LRs, has a nonsingular covariance matrix and induces the same Mahalanobis distances as those obtained using ILR balances and serves as an equivalent set of independent LRs in multiple regression, for example. The function **rda**() in the R package **vegan** (Oksanen et al., 2015), which performs redundancy analysis and requires matrix inversion when projecting the matrix of response variables onto the explanatory variables, anticipates singular cross-product matrices and deals with them as a matter of course.

A criticism repeatedly raised about using amalgamations is that they are not linear for CoDA (see, for example, Egozcue and Pawlowsky-Glahn, 2006, p. 155). In terms of geochemistry and mineralogy, amalgamations must be done in the simplex because the stoichiometric formulae are constructed based on crystal structure. However, to amalgamate using geometric means through ILR balances or some other multiplicative measure will not represent anything that is stoichiometrically meaningful. ILR balances might be interpretable within the context of chemical equilibrium equations, but in the case presented here, we are not combining chemical reactions. Rather, we are combining elements, in an additive way, that are common to specific processes, either through direct substitution within a crystal lattice structure or as part of a molecular structure that defines the crystal lattice. As demonstrated in this study, amalgamations can represent geochemical processes and their relevance can be assessed objectively by the logratio variance accounted for, and a mathematical argument unrelated to practical issues should not impede their use. John Aitchison himself said, referring to ILR balances and the orthonormal basis property, that "it is not that such structure is unimportant, but that we must not let pure mathematical ideas drive us into making statistical modeling more complicated than it is necessary" (Aitchison, 2008, p. 12). Aitchison himself proposed the use of amalgamations, which he defined in Aitchison (1986, p. 36–38), as a practical way of dealing with the problem of grouping of parts, especially when parts form hierarchies (Aitchison, 2008, Sect. 6.3). When considering variable transformations in the context of the logistic normal distribution, he also pointed out a difficulty with amalgamations that "there is no way of expressing the logarithm of a sum of components in terms of the logarithms of the components" (Aitchison, 1986, p.122).

SLR balances do obey the principle of subcompositional coherence but have none of the other elegant mathematical properties of ILR balances. A set of $J-1$ SLR balances does not explain exactly 100% of the logratio variance, as the same number of ILR balances can, but can come very close to 100% explained variance for all practical purposes, and are very simple to interpret. SLR balances can be included in the search process to find a small set of interpretable variables that effectively replace the complete set of LRs. The practitioner can intervene in the stepwise process, as demonstrated in a study of fatty acid compositions by Graeve and Greenacre (2020) and in Sect. 3.3. Notice that in fatty acid studies the ratio of polyunsaturated to saturated fatty acids (PUFA/SFA) is a common ratio to include in any analysis, and these two groupings of fatty acids are amalgamations and would never be defined by biochemists using geometric means.

5. Conclusion

Our overall conclusion is that, on the one hand, ILR balances have attractive theoretical properties, especially if one requires a transformation of the J-part compositional data set to a new set of exactly $J-1$ variables, which are equivalent to the original data in the sense of reproducing exactly the geometric structure of all the LRs. One problem with ILR balances is their complex interpretation, as we have shown, and possible misinterpretation of what constitutes a balance of two groups of parts. On the other hand, SLR balances have none of these theoretical properties, so they cannot reproduce the geometric structure exactly but can come very close to doing so. SLR balances are simply additional transformations of the compositional data that make substantive practical sense, and their contribution to any analysis of compositional data can be easily measured in terms of explained variance, a standard measure in modeling. SLR balances should not be rejected on theoretical grounds but accepted for their practical benefits, as no transformation that makes scientific sense, both causally and empirically, can be rejected as undesirable in research. With most geochemical datasets, the interpretations are subjective. However, if there is a geochemical model with which the geochemical data can be tested, then the interpretation can be objective, and SLR balances are an important tool in this respect.

The responses to the specific questions posed in Section 1 are as follows.

1. *Meaning and interpretation of ILR and SLR balances*: An ILR balance is proportional to the average of several LRs, as many as the product of numbers of parts in the numerator and denominator. An ILR balance should not be misinterpreted as the logratio of two amalgamations, or even two averages, of the original part values. An SLR balance can be directly interpreted in terms of the two amalgamations.
2. *Advantages and disadvantages of ILRs*: A full set of independent ILR balances with respect to an orthonormal basis forms an alternative set of variables that perfectly represents the geometry of all the LRs in the compositional data set, which is a notable mathematical property as a theoretical reference. The full set of ILR balances has a non-singular covariance structure that makes it useful for standard software that requires the inversion of a covariance or cross-product matrix. But notice that this is only a computational problem for software where the use of a generalized inverse has not been anticipated. Conversely, it is arguable whether single ILR balances have inherent meaning as summary univariate statistics, or as responses or explanatory variables in modeling. ILR balances are problematic with datasets containing many zeros, since zeros are replaced by small numbers and geometric means are strongly affected by rare parts.
3. *Advantages and disadvantages of SLR balances*: Amalgamating parts is a straightforward and meaningful way of combining parts in all applications of compositional data analysis, including geochemical applications. Logratios involving amalgamations, i.e. SLR balances, are transformations just like simple pairwise logratios (LRs) and can contribute, along with LRs of single parts, to constructing a set of transformations that explains the quasi-totality of the logratio variance. Thanks to the amalgamation process, SLRs reduce the number of zeros and thus alleviate the problem of data zeros. Apart from subcompositional coherence, SLR balances have none of the other elegant mathematical properties of ILR balances. An SLR balance contains no information about the part values within the numerator and denominator amalgamations – to understand those parts of the variability, other LRs or SLR balances need to be identified and interpreted. Amalgamations do impose a model as determined by the researcher, which is a limitation. However, the researcher can use different amalgamations to examine different possible meaningful processes, which follows the true intent of scientific inquiry. Based on the dataset considered in this paper, there is potential value in SLR balances as an additional means for examining and analysing compositional data.

Computer code availability

An R script for the analyses in this article can be downloaded from https://github.com/michaelgreenacre/CODAinPractice

Extensive use is made of the **easyCODA** package in R (R Core Team, 2019), which accompanies the book by Greenacre (2018). Version 0.32 of **easyCODA** on CRAN (Greenacre, 2018) was used in the analyses presented here, but the latest version 0.34 is already available on CRAN, containing new features such as clustering by amalgamation. The latest version is always available on R-Forge using the command:

```
install.packages(“easyCODA”, repos = “http://R-Forge.R-project.org”)
```

The **easyCODA** package depends on the **ca** package (Nenadić and Greenacre, 2007) and the **vegan** package (Oksanen et al., 2015). For example, the **vegan** function **protest()** computes the Procrustes correlation between two configurations of samples.

In the **ILR()** and **PLR()** functions in the **easyCODA** package, compared to Eqns (4) and (5) of Section 1, part weights are used rather than counts. Since all parts are considered equally weighted in the present study, they receive equal weights $1/J$, and the computations of ILRs and PLRs in **easyCODA** differ by a simple constant scaling factor, being the original definitions (4) and (5) divided by the square root of J. See Greenacre and Lewi (2009) for the justification of using unequal weights in compositional data analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Raimon Tolosana-Delgado for supplying the Aar Massif data set, which is used in the book by van den Boogaart and Tolosana-Delgado (2013) and which has recently been incorporated in the R package **compositions** (van den Boogaart et al., 2019). The stimulating and open debate within the CoDA community and the refinement of the developing field of compositional data analysis have been an inspiration for this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2020.104621.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman & Hall, London. Reprinted in 2003 with additional material by Blackburn Press.

Aitchison, J., 2008. The Single Principle of Compositional Data Analysis, Continuing Fallacies, Confusions and Misunderstandings and Some Suggested Remedies. Keynote address, CODAWORK08. https://core.ac.uk/download/pdf/132548276.pdf. (Accessed 31 December 2018).

Aitchison, J., Greenacre, M.J., 2002. Biplots for compositional data. J R Stat Soc Ser C (Appl Stat) 51, 375–392.

de Caritat, P., Grunsky, E., 2013. Defining element associations and inferring geological processes from total element concentrations in Australia catchment outlet sediments: multivariate analysis of continental-scale geochemical data. Appl. Geochem. 33, 104–126.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

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

Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences: from Theory to Practice, vol. 264. Geological Society, London, Special Publications, pp. 67–77.

Egozcue, J.J., Lovell, D., Pawlowsky-Glan, V., 2013. Testing compositional association. In: Proceedings of the 5th International Workshop on Compositional Data Analysis. Vorau, Austria, pp. 28–36.

Fačevicová, K., Hron, K., Todorov, V., Templ, M., 2016. Compositional tables analysis in coordinates. Scand. J. Stat. 43, 962–977.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis. Springer, New York.

Gower, J.C., Dijksterhuis, G.B., 2004. Procrustes Problems. Oxford University Press, Oxford.

Graeve, M., Greenacre, M., 2020. The selection and analysis of fatty acid ratios: a new approach for the univariate and multivariate analysis of fatty acid trophic markers in marine organisms. Limnol Oceanogr. Methods 18 (5), 196–210. https://doi.org/10.1002/lom3.10360.

Greenacre, M., 2009. Power transformations in correspondence analysis. Comput. Stat. Data Anal. 53, 3107–3116.

Greenacre, M., 2010a. Logratio analysis is a limiting case of correspondence analysis. Math. Geosci. 42, 129–134.

Greenacre, M., 2010b. Biplots in Practice. BBVA Foundation, Bilbao. Free download at www.multivariatestatistics.org.

Greenacre, M., 2011. Measuring subcompositional incoherence. Math. Geosci. 43, 681–693.

Greenacre, M., 2013. Contribution biplots. J. Comput. Graph Stat. 22, 107–122.

Greenacre, M., 2018. Compositional Data Analysis in Practice. Chapman & Hall/CRC, Boca Raton, Florida.

Greenacre, M., 2019. Variable selection in compositional data analysis, using pairwise logratios. Math. Geosci. 51, 649–682.

Greenacre, M., 2020. Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. Appl Comput Geosci 5. https://doi.org/10.1016/j.acags.2019.100017.

Greenacre, M.J., Lewi, P.J., 2009. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J. Classif. 26, 29–64.

Grunsky, E.C., Kjarsgaard, B.A., Egozcue, J.J., Pawlowsky-Glahn, V., Thio-Henestrosa, S., 2008. Studies in stoichiometry with compositional data. In: The 3rd Compositional Data Analysis Workshop. University of Girona, Girona, Spain, pp. 27–30. May, 2008, CD-ROM publication.

Grunsky, E.C., Bacon-Shone, J., 2011. The stoichiometry of mineral compositions. In: Proceedings of 2011 Compositional Data Analysis Workshop. https://www.recercat.cat/handle/2072/324114. (Accessed 25 May 2019).

Grunsky, E.C., Kjarsgaard, B.A., 2016. Recognizing and validating structural processes in geochemical data. In: Martín-Fernández, J.A., Thio-Henestrosa, S. (Eds.), Compositional Data Analysis, Springer Proceedings in Mathematics and Statistics, vol. 187, pp. 85–116. https://doi.org/10.1007/978-3-319-44811-4_7, 209pp.

Hron, K., Filzmoser, P., de Caritat, P., Fišerová, E., Gardlo, A., 2017. Weighted pivot coordinates for compositional data and their application to geochemical mapping. Math. Geosci. 49, 777–796.

Krzanowski, W.J., 1987. Selection of variables to preserve multivariate data structure, using principal components. Appl Statist 36, 22–33.

Kynčlova, P., Hron, K., Filzmoser, P., 2017. Correlation between compositional parts based on symmetric balances. Math. Geosci. 49, 777–796.

Legendre, P., Legendre, L., 2012. Numerical Ecology, third ed. Elsevier, Amsterdam.

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2018. Advances in principal balances for compositional data. Math. Geosci. 50, 273–298.

Mateu-Figueras, G., Daunis-i-Estadella, J., 2008. Compositional amalgamations and balances: a critical approach. In: Daunis-i-Estella, J., Martín-Fernández, J.A. (Eds.), Proceedings of 3rd Compositional Data Analysis Workshop. https://dugi-doc.udg.edu/handle/10256/738. (Accessed 12 May 2019).

Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J., 2011. The principle of working on coordinates. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications. Wiley, Chichester.

Nenadić, O., Greenacre, M., 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. J. Stat. Software 20 (3). http://www.jstatsoft.org/v20/i03/. (Accessed 31 December 2018).

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H., 2015. Vegan: Community Ecology Package. R Package Version 2.3-2. https://CRAN.R-project.org/package=vegan. (Accessed 31 December 2018).

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2007. Lecture Notes on Compositional Data Analysis. http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1. (Accessed 26 December 2019).

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. Wiley, UK.

R Core Team, 2019. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Tolosana-Delgado, R., von Eynatten, H., 2010. Simplifying compositional multiple regression: application to grain size controls on sediment geochemistry. Comput. Geosci. 36, 577–589.

van den Boogaart, K.G., Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer-Verlag, Berlin.

van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2019. Compositions: compositional Data Analysis. R Package Version, 1, pp. 40–43.

van den Wollenberg, A.L., 1977. Redundancy analysis – an alternative for canonical analysis. Psychometrika 42, 207–219.