A comparison of amalgamation and isometric logratios in compositional data analysis

Michael Greenacre[1], Eric Grunsky[2] and John Bacon-Shone[3]

[1](Corresponding author)

Department of Economics and Business, Universitat Pompeu Fabra,

and Barcelona Graduate School of Economics,

Ramon Trias Fargas 25-27, Barcelona

08005 Spain

Email: michael.greenacre@upf.edu

Telephone (mobile): +34 680232383

Orcid: https://orcid.org/0000-0002-0054-3131

[2] Dept of Earth & Environmental Sciences

200 University Ave. W

Waterloo, Ontario, Canada N2L 3G1

Email: egrunsky@gmail.com

Orcid: https://orcid.org/0000-0003-4521-163X

[3]Social Sciences Research Centre

The University of Hong Kong

Pokfulam Road

Hong Kong

Email: johnbs@hku.hk

Orcid:https://orcid.org/0000-0002-9827-1815

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, constructed the hierarchical thought example and suggested "what would JA do"

Abstract: The isometric logratio transformation 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 these properties are not a prerequisite for good practice in compositional data analysis. Some simple artificial examples demonstrate the weaknesses of using the isometric logratio in practice as a univariate response or explanatory variable. The source of the weakness is the use of the geometric mean, an averaging measure that is counter-intuitive to the combining of parts and highly affected by parts with small relative values. When a comparison between two groups of parts is required in practical applications, preferably based on substantive knowledge, it is demonstrated that logratios of amalgamations serve as a simpler, more intuitive and more interpretable alternative to isometric logratios. The study is then extended to a real geochemical data set, which is transformed to a new set of variables called isometric logratio "balances". A simpler alternative approach, using a reduced set of pairwise logratios of parts, optionally involving prescribed amalgamations, which also has an exact inverse, is close to optimal in accounting for the variance in a compositional data set. This approach highlights which compositional parts are driving the data structure, using variables that are easy to interpret and map well to research-driven objectives.

Keywords: amalgamation; balance; compositional data; geometric mean; logratio transformation; logratio analysis; logratio distance; multivariate analysis; ratios; subcompositional coherence; univariate statistics.

Highlights:

- Isometric logratios, which use geometric means, have difficulties in interpretation
- Logratios of amalgamated parts are a preferable alternative in practice
- Simpler logratio transformations can explain the structure of compositional data

1 Introduction

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. The simplest examples are the log-transformed ratios of two parts of a composition, or pairwise logratios, which have been used since the earliest work of Aitchison. For a J-part composition, with values denoted by $x_1, x_2, ..., x_J$, there are $½J(J-1)$ unique pairwise logratios.

The simplest set of logratios that includes all compositional parts is the set of additive logratios (ALRs), where a specific reference part is contrasted with all the other parts (here the reference is chosen as the last part):

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

Since any part can be a reference part, there are exactly J possible sets of ALRs.

The centered logratio (CLR) is the logratio between a part and the geometric mean of all the parts. The complete set of J CLRs is defined as:

$$\mathrm{CLR}(j) = \log\frac{x_j}{\left(\prod_j x_j\right)^{1/J}} \qquad j = 1,..., J. \tag{2}$$

The CLRs are not intended to be used as alternative variables to represent the parts in each numerator, but serve a very useful computational purpose, in that the set of J CLRs provides a computational shortcut to analyzing the complete set of pairwise logratios (see, for example, Aitchison and Greenacre 2002, Appendix A).

The isometric logratio (ILR), defined by Egozcue et al. (2003), has been promoted by several authors as the mathematically correct way express a compositional data set with respect to a new set of $J-1$ coordinates, called ILR "balances", after which these ILRs are used in data analysis, modeling, and multivariate methods such as clustering and dimension reduction – see, for example, Egozcue and Pawlowsky-Glahn (2006), Mateu-Figueras, Pawlowsky-Glahn and Egozcue (2011), van den Boogaart and Tolosona-Delgado (2013), Buccianti (2015), Hron et al. (2017), Morton et al.

(2017), Washburne et al. (2017) and Martín-Fernández et al. (2018). A single ILR contrasts two subsets of different parts, denoted by J_1 and J_2, by defining the logratio of their respective geometric means, with a scaling factor (Egozcue et al. 2003):

$$\mathrm{ILR}(J_1, J_2) = \sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}} \log \frac{\left(\prod_{j\in J_1} x_j\right)^{1/|J_1|}}{\left(\prod_{j\in J_2} x_j\right)^{1/|J_2|}} \tag{3}$$

where $|J_1|$ and $|J_2|$ denote the number of numerator parts and denominator parts respectively.

A special case of a set of ILRs is a set of pivot logratios (PLRs), which are a succession of ILRs 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:

$$\mathrm{PLR}(j) = \sqrt{\frac{|J_2|}{1+|J_2|}} \log \frac{x_j}{\left(\prod_{j\in J_2} x_j\right)^{1/|J_2|}} \tag{4}$$

where $j = 1,..., J-1$ and J_2 is the set of parts $J_2 = \{j+1, j+2, ..., J\}$ (Hron et al. 2017). A PLR, with its single part in the numerator, has the advantage of being able to be expressed as an average of pairwise logratios. For example, the first PLR is, apart from the scalar multiplier, equal to $[\log(x_1/x_2) + \log(x_1/x_3)+\cdots+\log(x_1/x_J)]/(J-1)$. Thus the first PLR is the negative of the average of the ALRs defined in (1), where x_1 is the reference part in the denominator. Notice that these "first PLRs" (i.e., the J PLRs that each have a different part in the numerator and all the $J-1$ others in the denominator) are proportional to the set of CLRs, since the CLRs just have an extra logratio equal to 0 (e.g. $\log(x_1/x_1)$ for the first CLR) (Hron et al. 2017). Similarly, they are -1 times the averages of the respective sets of ALRs where each x_j in turn is the reference part. When it comes to choosing a set of ILRs, the PLRs are the "first option" of Filzmoser, Hron and Templ (2018, page 35).

Logratios of amalgamations of parts have not been widely used, although – paradoxically – parts used in compositional data analysis are often defined as amalgamations themselves. Denoted here by SLR (standing for "summated logratio"), an amalgamation logratio (or "amalgamation balance") is more simply defined, again for two different subsets of parts, as:

$$\mathrm{SLR}(J_1, J_2) = \log \frac{\sum_{j \in J_1} x_j}{\sum_{j \in J_2} x_j} \tag{5}$$

Notice that a SLR is a logratio, without any scaling factor, is just like any other pairwise logratio. Amalgamations are often performed in chemistry based on the understanding of the stoichiometric balances. Just like all the other logratio transformations (1)–(4) above, a set of $J-1$ SLRs that involve all the parts can be inverted back to the original J part values, in this case by solving a set of linear equations – computational details are given in the Appendix.

In this study the following questions are considered:

1. What is the interpretation of an ILR? Is it clear what its values are measuring in practice?

2. What are the advantages of the ILR transformation? Are these advantages of practical worth?

3. What are the disadvantages of the ILR transformation? Do these disadvantages have practical repercussions?

4. Are ratios that involve amalgamations of parts a viable alternative to ILRs? And what are the advantages and disadvantages of such amalgamation balances in practice?

Two sets of data are used in order to answer these questions: first, a small artificially constructed data set, and second, a typical geochemical data set. Section 2 describes these data sets as well as the methodology followed and software used. Section 3 gives the results for each of the data sets. Sections 4 and 5 follow with a discussion and overall conclusion about the above-mentioned questions. Supplementary material is supplied, including data, additional tables and R code.

2 Material and methods

2.1 Data set 1: a three-part artificial data set

This artificial data set was provided by Martín-Fernández (2018) during an online debate about the use of ILRs[*]. The data (see Supplementary Material 2) form a three-part composition of the consumption of wine, beer and spirits, and the objective is to investigate whether the proportion of consumed spirits depends on the relative consumption of beer compared to wine. The contextual use of wine, beer and spirits is chosen only because these are familiar products, but they could just as well be carbohydrate, protein and fat in foodstuffs or any three chemical compositional parts of one's choice.

2.2 Data set 2: the Aar Massif data, a typical geochemical 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 Eynatten 2010), analysed by van den Boogaart and Tolosana-Delgado (2011) and Martín-Fernández et al. (2018). 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: CaO, P_2O_5

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 and define the stoichiometry of the mineral. 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.

[*] https://www.coda-association.org/en/coda-info/coda-letters/debate-1-2017june/

2.3 Methods

Apart from some standard statistical methods, the approach focuses on the analysis of logratios of parts or of amalgamated parts, compared to the use of ILRs. 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 logratios.

2.3.1 Measurement, substantive meaning and interpretation

Here the scales of the particular logratios are examined, namely what each logratio is actually measuring. Their meaning and interpretation are judged relative to the research question of the particular study, and it is investigated whether the logratios have a clear and unambiguous interpretation.

2.3.2 Explained logratio variance

The total logratio variance in a compositional data set quantifies the data content and is equal to the sum (or average) of the variances of the CLRs, equivalently the sum (or average) of all pairwise logratios (the average option is taken here, as in Greenacre (2018a, b), although it makes no difference to the eventual percentages of explained variance). The advantage of using the average of the variances is that the variance does not increase with the number of parts in the study. Additionally, variables that are not deemed relevant to the research question, or do not contribute to the structure of the data can be dropped. This is, in effect, a subcomposition, but one that is directly related to the investigation of structure or a research question. Given any explanatory variables, the amount of the total logratio variance explained can be computed by regressing each of the J CLRs on these variables, obtaining the parts of variance explained in each case, summing these J explained parts and then expressing that sum relative to the total variance. This set of regressions is embodied in the method of redundancy analysis (van den Wollenberg 1994), which can be used to obtain the percentage of explained variance in a simple matrix computation.

The present application uses explanatory variables in the form of pairwise logratios, ILRs or SLRs, so that the approach involves quantifying how much variance can be explained by a subset of the logratios themselves, and comparing with the corresponding results for ILRs – see Greenacre (2018b).

2.3.3 Selecting logratios to identify parts that explain data structure

To find a subset of logratios, Greenacre (2018b), inspired by Krzanowski's (1987) work on variable selection, proposed a stepwise process where logratios are selected that explain a maximum part of the total logratio variance at each step. Identifying such a subset of pairwise logratios implies identifying a subcomposition of parts (i.e. those used in the logratios) 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, should be included as candidates for creating logratios.

The stepwise procedure starts by first finding the logratio 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 (2018a,b). The percentages of variance explained show how well these sets of logratios can serve as alternative variables to represent the compositional data set.

2.3.4 Procrustes correlation

The samples can be displayed exactly in a $(J-1)$-dimensional Euclidean space, where their interpoint distances match the Euclidean distances either between the $½J(J-1)$ pairwise logratios or equivalently between the J CLRs, or between a set of ILR or PLR balances. In order to see how closely this multivariate structure of the samples 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 (2018b, Appendix) for the mathematical definition.

2.3.5 Principal component analysis of logratios and logratio analysis

In order to visualize the structure of compositional data, logratio analysis (LRA) is used (Aitchison and Greenacre 2002, Greenacre 2010, 2018a,b) to reduce the dimensionality of the data, projecting them onto a subspace that explains the maximum amount of logratio variance, usually of dimension two for ease of interpretation. LRA is the PCA of the full set of CLRs, where the resultant biplot shows the J parts with the interpretation focusing on the $½J(J-1)$ links connecting pairs of parts. These links represent the respective pairwise logratios, while the positions of the samples are such that their interpoint distances approximate the true logratio distances. LRA is thus also equivalent to the PCA of the matrix of pairwise logratios. When a reduced subset of logratios is selected, its structure will be visualized and interpreted using PCA.

2.4 Software

Extensive use is made of the **easyCODA** package in R (R core team 2018), which accompanies the book by Greenacre (2018a). Version 0.31 of **easyCODA** was used in the analyses presented here. The package can be installed from CRAN but 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. Ternary plots are drawn using function **TernaryPlot()** in the **Ternary** package (Smith 2017).

A difference that should be mentioned in the **ILR()** and **PLR()** functions in the **easyCODA** package, compared to Eqns (3) and (4) of Section 1, is that part weights are used rather than counts. Since all parts are considered equally weighted in the present study, they each receive weight 1/*J*, and the computations of ILRs and PLRs in **easyCODA** differ by a simple constant scaling factor, being the original definitions (3) and (4) divided by the square root of *J*. See Greenacre and Lewi (2009) for the justification of using unequal weights in compositional data analysis.

3 Results

3.1 Artificial example using data set 1 showing the weaknesses in using ILRs

3.1.1 Simple regression analysis satisfying the research question

Since interest is focused on the proportion of spirits as a response variable and wants to use logratios, the researcher plots the amalgamation logratio of spirits/(beer+wine) (i.e., spirits/(1-spirits) in this case, which is monotonically related to the proportion of spirits) against the logratio of beer/wine (Figure 1a). The linear regression is highly significant ($p <$ 0.0001). The effect size in this log-log relationship is expressed in terms of percentage changes in both the independent and dependent variables and the slope of 0.189 in Fig. 1a translates to an estimated increase of 1.82 % in the spirits/(beer+wine) ratio for every 10 % increase in the beer/wine ratio.

3.1.2 ILR alternative

As an alternative, the researcher uses isometric logratios for both variables, namely $\sqrt{2/3}\,\log(\mathrm{spirits}/(\mathrm{beer}\times\mathrm{wine})^{1/2})$ versus $\sqrt{1/2}\,\log(\mathrm{beer/wine})$, the latter logratio being the simple logratio used before with the scaling constant inherent in the ILR definition. Now the relationship, plotted in Fig. 1b, is no longer significant ($p = 0.79$). This presents the researcher with a dilemma, since different results are obtained depending on whether the amalgamation or isometric logratio is used as the response variable.

3.1.3 Original data in a ternary plot

In an attempt to understand which of the two analyses is reflecting the true situation, the data are visualized in a ternary plot (Fig. 2a). It is clear that, as the sample points are moving from left to right, for increasing beer/wine ratio, there is an increase in the proportion of spirits, corroborating the result of the first analysis. Fig. 2b shows an enlargement of the points in the ternary plot, added to which is the fitted model in Fig. 1a back-transformed to ternary space as a curve. The ascent of the curve is clear as the proportion of spirits rises with increasing beer to wine ratio

3.1.4 Fundamental difficulties with the interpretation of isometric logratios in practice

The ILR obscures what is obvious in the ternary plot because it does not truly contrast the proportion of spirits against that of the combination of beer and wine. Its value also depends on the relative values of beer and wine, which affect the geometric mean in the denominator of the ratio. The present example has values of (beer+wine) on average 0.88. Fig. 3 shows how much the geometric mean of beer and wine, i.e. $\sqrt{\mathrm{beer}\times\mathrm{wine})}$, can vary as a function of the ratio beer/wine, for this fixed value of 0.88 of the sum beer+wine.

Thus, for any fixed value of the amalgamation beer+wine, the value of the geometric mean in the denominator of the ILR $\sqrt{2/3}\,\log(\text{spirits}/\sqrt{\text{beer}\times\text{wine}})$ changes depending on the ratio beer/wine. This additional source of variation in the ILR value has effectively nullified the relationship between spirits and the ratio beer/wine, a relationship that clearly exists and which is statistically significant.

This simple three-part example shows that the geometric means in an ILR do not involve simple groupings of the parts. When there are many parts, as in most real-life applications, an ILR is a variable with a very complex interpretation. Thinking of it superficially as a ratio between two groupings of parts can be erroneous.

3.1.5 Selecting a set of ILR "balances"

In this example, there are only three possible sets of ILR "balances", each consisting of two contrasts: {spirits vs. wine&beer, and wine vs. beer}, {wine vs. spirits&beer, and spirits vs. beer} and {beer vs. spirits&wine, and spirits vs. wine}. Any one of these serves the purpose for which ILRs are intended, but it is the first one that was chosen to be used in Fig. 1, because of the research question. Martín-Fernández et al. (2018) describe a recursive partitioning algorithm for choosing a set of "principal balances" where, starting from the full set of parts, an optimal split is found which engenders the greatest contrast. This algorithm would favor the contrast between wine and spirits&beer as the first "principal balance". But the researcher would not be interested in such a split, since the research question is to compare spirits consumption with beer&wine consumption. An automatic choice is of no use in this case, where the choice should be decided by the practitioner.

Furthermore, the enumeration of the possible sets of ILRs is trivial in this three-part problem, where there are only three possibilities, each of which is representable as a dendrogram. But the number of combinations becomes astronomical for higher-dimensional problems, equal to $(2J–2)!/(2^{J–1}(J–1)!)$ (Murtagh 1984, Bóna 2006) – this is equal to 3 when $J = 3$, as above, but is equal to 34 459 425 when $J = 10$, as in the forthcoming geochemical example in Sect. 3.2, a data set of quite modest dimensionality. Martín-Fernández et al. (2018) admit that their exhaustive search algorithm is feasible computationally up to 15 parts, which is less than many geochemical data sets.

3.1.6 Same situation but with many parts

A hypothetical situation is when the collected data consist of the consumption broken down into 10 brands of spirits, 20 brands of beer and 30 brands of wine for a total of 60 parts. This dataset is intended to address the same research objective as for data set 1, but is problematic for compositional data analysis because most rows contain at least 50 zeros as households typically consume at most 10 brands of alcohol. The logratios of amalgamations would provide results matching those for data set 1 as brands are amalgamated within each of the three types of alcoholic drink. Conversely the ILR approach to this dataset requires the use of zero replacement for at least 50 of the 60 parts, making the geometric means of the three types of alcoholic drinks depend heavily on the specific zero replacement strategy used, which is unrelated to the stated objective. The geometric means will also depend heavily on the degree of brand loyalty, as this affects the number of zeroes and hence the geometric means, again not related to the stated objective. Hence the theoretical advantage of ILR is wholly negated by its requirement of large-scale zero replacement and introduction of unrelated variability, both distorting the analysis to map poorly onto the stated objective.

3.2 Isometric and amalgamation logratios in geochemistry, using data set 2

3.2.1 Ratio selection, including ratios of amalgamations

For the Aar Massif data, the three amalgamations of Mafic, Felsic and Carbonate (see Sect. 2.1) were used to form ratios with the oxides or with other amalgamations in the search for the set of logratios that maximized the explained variance of the compositional data set. Table 1 shows the selected ratios, their cumulative explained variance, and the Procrustes correlations of the sample configurations with the exact sample configuration. Fig. 4 shows a graph of the solution, which involves Felsic and Carbonate but not Mafic. The explained variance was only 0.003% short of 100% (it was equal to 99.997%, rounded to 100.0 in Table 1), with a Procrustes correlation between their geometry and the exact logratio geometry of 0.993. An even smaller set of ratios can be considered seeing that already with only four ratios more than 95% of the logratio variance is explained, with a Procrustes correlation of 0.976.

Fig. 5 compares the original geometry of the compositional data set, using 45 pairwise logratios, with that of the reduced set of nine logratios, using PCA. The similarity in the configurations of the samples is clear, due to the Procrustes correlation that is almost 1.

3.2.2 Knowledge-driven intervention in the stepwise process

The completely automatic stepwise process, giving the results in Table 1, Fig. 4 and Fig. 5b, chooses the logratio that gives the highest additional explained logratio variance at each step. In fact, there are several logratios 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 a logratio 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. 4), but its components MgO, Fe_2O_3 and MnO are clearly aligned in Fig. 5a and opposing the Felsic parts Na_2O, SiO_2, Al_2O_3, K_2O. From the positions of MgO and Na_2O in Fig. 5a it is no surprise that MgO/Na_2O is the ratio of choice in the first step of the algorithm. This optimal logratio of a Mafic part with respect to a Felsic part has an optimal explained variance of 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 by which one can experiment with 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, being the fourth in the list of Table 2. After selecting this ratio as the first one, and then letting the stepwise process take its automatic course afterwards, a partially different selection of logratios is obtained, but still explaining 99.997% of the logratio variance, the same as before, and with a high Procrustes correlation of 0.990. The resulting PCA of the logratios is shown in Fig. 6, where the configuration of samples is practically identical to those in Fig. 5.

3.2.3 Comparison of best single ratios of different types

It is instructive to compare the best single ratios 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 can also be written as a logratio, involving powers of the parts. In descending order, the best ratios are:

• the first principal component:	71.2 %
• the first principal balance of Martín-Fernández et al. (2018):	70.7 %
• the first pairwise logratio of MgO/MnO in Table 1:	69.1 %
• the CLR of Na_2O:	68.6 %
• the "first" PLR of Na_2O versus the other oxides	68.6%

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

The single pairwise logratio of MgO/MnO, involving only two parts, compares very favorably with the others, all of which involve the complete set of 10 parts. This ratio, found with minimal computational effort, explains only 1.6 percentage points less than the first principal balance, which involves an exhaustive and costly search algorithm to find the optimal ILR. This good behavior of simple pairwise logratios has been found in different applications, for example Greenacre (2018a,b), Graeve and Greenacre (2018).

4. Discussion

Various articles and books on compositional data analysis maintain that using ILRs, or at least transformations to orthonormal coordinates, is mandatory for further statistical analysis. For example, Mateu-Figueras, Pawlowsky-Glahn and Egozcue (2011) require the use of coordinates using an orthonormal basis based on ILRs, but admit that "it is not obvious how to determine which basis is the most appropriate for any given problem". Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2015) state categorically that "compositions are represented by orthonormal coordinates, which live in a real Euclidean space". Fačevicová et al. (2016) say that "compositional vectors need to be expressed in orthonormal coordinates, thereby allowing further processing using standard statistical tools". Kynčlová, Hron and Filzmoser (2017) say that compositional data should be expressed with respect to orthonormal coordinates that "guarantee isometry between the Aitchison geometry [of the simplex] and the real space". Filzmoser, Hron and Templ (2018, page 35) say that "isometric logratio coordinates, real coordinates with respect to an orthonormal basis in the Aitchison geometry, are preferable". By contrast, it is remarkable that Aitchison himself expresses disagreement about this requirement. He refers to the "fallacy" of using orthonormal coordinates, saying that "given the elegance of the algebraic-geometric (Hilbert space) structure of the simplex 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).

While the above statements generally refer to a new set of coordinates for compositional data, several authors have substituted existing practice of using single ratios of amalgamations of parts with the use of single ILR equivalents, with unclear justification and interpretation of this alternative. For example, Buccianti (2015) rejects well-established variables and scatterplots in water chemistry saying that "it is not possible to apply statistical analysis correctly (...) on diagrams with coordinates given by molar ratios, since they represent non-orthogonal directions, thus limiting the modelling phase." Consequently, Buccianti (2015) revises a "classical diagram from a compositional data analysis perspective", namely the Gibbs diagram with logarithm of total dissolved solids (TDS, an amalgamation) on the vertical y-axis. The vertical axis is substituted with the logarithm of the geometric mean of the eight dissolved solids versus the actual amalgamation of all the other components, questionably called an ILR "balance". With this "revision" the approach is stated as now being "coherent with the nature of compositional data, thus obtaining a simple tool to be used in a statistical sense, going beyond the descriptive approach" (Buccianti 2015, page 94). In fact, the geometric mean of the 8 TDS parts is minuscule compared to the other water components, hence the value of this "balance" is, for all practical purposes, almost exactly proportional to log(TDS) used in the "classical diagram". The use of this "balance" implies some benefit over the Gibbs diagram, but it only adds an unnecessary complication to the interpretation. In reality, using molar ratios is necessary if we are to correctly identify any stoichiometric integer linear constraints (which are only integer linear constraints in molar ratios), which in turn is essential if we are to ensure that our models make scientific sense (Grunsky and Bacon-Shone, 2011).

Another example is by Morton et al. (2017), who relate a sparse 88×116 data matrix of counts of 116 microbial species in 88 soil samples to several environmental variables. A measure of counts should not be considered as compositional as there is no closure effect and a compositional approach may not reveal the true relationships derived from count data. One analysis consists of computing an ILR contrasting the abundances of 86 species with those of the other 30 and showing a scatterplot of this ILR versus the pH of the samples, with a clear negative correlation of −0.91. Using the much simpler logratio of the two respective amalgamations gives an almost identical result when plotted against pH and the correlation is −0.94 – see Greenacre (2018c). The ILR depends on the relative values of all 86 species in the numerator and all 30 species in the denominator, making it difficult, if not impossible, to give it an unambiguous interpretation. The authors also interpret this "balance" as a ratio of summed values, saying that in a comparison of a two groups of "Blue" and "Red" individuals (signifying their two groups of parts) the balance becomes "increasingly negative, since there are more Blue individuals than Red individuals".

Everything said above applies to the pivot logratios, defined in (4). Pivot balances depend on the ordering of the parts, so there are fewer sets of them compared to ILR balances: for J = 10 there are $J!/2$ = 10!/2 = 1 814 400 possible sets of pivot balances, but still too many for con-sideration in practice.

There are two main professed benefits of ILRs: first, the definition of a new set of orthonormal coordinates for the data, which preserve the Aitchison geometry; and second, their role in grouping the parts. While the Aitchison geometry provides a useful theoretical reference, neither of these benefits are required for good practice, and can impose unrealistic limitations in practical applications that distort the answers to meaningful research objectives.

Concerning the geometry, a set of J–1 ILR balances needs to be defined in order to provide a new set of coordinates for a J-part compositional data set (see Sect. 2.4). These 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, and this property is the principal justification for their definition.

However, it is not necessary to satisfy perfectly this mathematical exact reproduction of the sample space by the ILR coordinates. PCA of the CLRs, called logratio analysis or LRA (Greenacre 2010, 2018) and used in Fig. 6a, is specifically used to separate non-random from random variation in a compositional data set, after which the non-random part on the major principal dimensions is interpreted, discarding the minor dimensions. These lesser components may represent either random effects or under-sampled processes (Grunsky and Kjarsgaard 2016). So it seems perfectly acceptable that some non-informative variability in the compositional data set be removed by appropriate and meaningful transformations rather than using the complicated ILRs which aim at a perfect re-expression of the original data. Selecting key logratios and possible ratios of amalgamations that explain almost 100% of the logratio variance and which closely approximate the logratio geometry presents an easier alternative for the practitioner.

Mateu-Figueras and Daunis-i-Estadella (2008) make a comparison between logratios based on amalgamations and ILRs 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 amalgamation logratio is found to be non-significant between groups, whereas the mean of the ILR is significant. This is reminiscent of the contradictory results found in Section 3.1. The interpretation of the amalgamation logratio is clear and unambiguous, but what the ILR measures is not, being affected by several pairwise logratios. One of the parts has much lower proportions than the others and thus radically affecting the geometric mean where it is included. The significant group difference in the case of the ILR is not possible to explain unless further investigation is conducted.

Concerning the claimed benefit, the ILR balances are promoted as being "easily interpreted in terms of grouped parts of a composition" (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2015, p. 38). This statement is speculative − Sect. 3.1 has already shown the simplest of examples to refute that they are "easily interpreted". Washburne et al. (2017) say that "the balances in a rooted ILR transform ... can be intuited as the average difference between taxa in two groups". But any claim or suggestion that ILRs are contrasting groups of parts in the sense of amalgamating them (or averaging them) should be viewed with the utmost caution.

Introducing the term "balance" (Egozcue and Pawlowsky-Glahn 2005) creates the impression that an ILR is "balancing" two groups of parts, as if there were two sets of weights. In fact, Pawlowsky-Glahn, Monreal-Pawlowsky and Egozcue (2015, Figure 4) depict it as a physical balance with parts on the left and on the right and say that "when the mean balance is placed at the left side (...) it points out that the parts on the left have greater proportions than the parts placed at the right: it works like a lever in equilibrium i.e. a balance in the plain sense." The use of "greater proportions" suggests amalgamations of proportions, not geometric means of them. Similarly, Morton et al. (2017, page 2) draw a "balance" reminiscent of a child's see-saw, implying that the opposite sides of the see-saw are weighted down by "multiple species on the left end of the balance and multiple species on the right end of the balance" (Morton et al. 2017, page 5).

Buccianti, Nisi and Raco (2015) say that in an ILR the "ratio measures the relative weight of each group and the logarithm provides the appropriate scale, and a positive balance means that, in (geometric) mean, the group of parts of the numerator has more weight in the composition than the group of the denominator (and conversely for negative balances)." The essential detail here is to understand exactly what is being measured by the weights of the groups of parts in the numerator and the denominator "in (geometric) mean".

Pawlowsky-Glahn et al. (2015, p. 41) create the misleading impression that an ILR is indeed interpretable as a logratio of amalgamated parts. They give an election example of an ILR balance contrasting four leftwing parties with two rightwing parties:

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

saying 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". This is incorrectly interpreting the ILR balance as if it is the ratio of sums, not of geometric means. Even the simplest 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.

In summary, explanations in the literature as to what the ILR really measures are either misleading or vague. Another example is in the interpretation of the ILR special case, the pivot logratio: Filzmoser et al. (2018) say that if the PLR = 0, it indicates "a balanced state between [the numerator part] and an average behavior of the other parts in the given composition", where the meaning of "average behavior" of several parts, some of which could be common and some rare, is vague and unclear.

When it comes to combining parts, the specialist has knowledge about the possible models that the empirical relationships might reveal and amalgamations rely on this knowledge. Moreover, amalgamations can be applied if there are problems with the number of degrees of freedom and 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.

In the book by Pawlowsky-Glahn et al. (2015), amalgamations are ruled out, where they specifically state that "amalgamation is incompatible with the techniques presented in this book". But then the same authors demonstrate the use of amalgamations in the form of a residual part: "a fill-up or residual value is equivalent to using an amalgamated composition" and "if only some parts of the composition are available, a fill-up or residual value can be defined". This is paradoxical: amalgamations are acceptable if they fix up a data problem, but in general they are not acceptable. Some practitioners create a “fill variable” that represents the difference between the total sum of the proportions that add up to less than the theoretical constant sum. This “fill variable” can represent many different processes because the variable does not define specific variables that account for processes and structure in the data. There is no added advantage in creating this variable and it is better to treat the data as a subcomposition. In reality, true complete compositions do not exist. Limitations on instrument measurements or errors in measurement will always result in an incomplete composition. Thus, in practice, all compositions are subcompositions and the data analyst will discover and test structure in the data based on these subcompositions.

A criticism repeatedly raised about using amalgamations is that they are not linear in the simplex (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. To amalgamate through ILR or some other multiplicative measure will not represent anything that is stoichiometrically meaningful.

The imposition of this mathematical condition again restricts the practitioner from using alternatives that make perfect substantive sense in practical applications. 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 should not impede their use John Aitchison himself said that "it is not that such structure (referring to ILRs and the orthonormal basis property) 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). The drawback of the interpretability of ILRs has been expressed, for example, by van den Boogaart and Tolosana-Delgado (2013, page 45): "the strongest difficulty with the ilr-transformed values or any orthonormal coordinates [is that] each coordinate might involve many parts (potentially all), which makes it virtually impossible to interpret them in general... The generic ilr transformation is thus a perfect black box". Aitchison also proposed the use of amalgamations, which he defined in Aitchison, (1986, p. 267), as a practical way of dealing with the problem of grouping of parts, especially when parts form hierarchies (Aitchison 2008, Sect. 6.3). These ideas are faithfully implemented in the present paper, demonstrating that amalgamations function well and thus supporting Aitchison's viewpoint.

Amalgamations can be included in the logratio search process to find a small set of interpretable variables that effectively replace the complete set of logratios. The practitioner can intervene in the stepwise process, as demonstrated in a study of fatty acid compositions by Graeve and Greenacre (2018) and in Sect. 3.2.4. 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 would never be defined by biochemists as geometric means.

An additional claimed benefit of the ILR transformation 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 analysis and computation of Mahalanobis distances. But this is not an additional benefit, because the generalized inverse can be used directly on the singular covariance matrix of the J CLRs, for example, with identical results. Moreover, any set of additive logratios (ALRs), or – more generally – any linearly independent set of $J-1$

pairwise logratios, has a nonsingular covariance matrix and induces the same Mahalanobis distances as those obtained using ILRs and serves as an equivalent set of independent logratio variables in multiple regression.

5. Conclusion

Our overall conclusion is that isometric logratios (logratios of geometric means), while theoretically attractive, present important barriers in the practice of compositional data analysis using meaningful research objectives and can be substituted by simple logratios and logratios of amalgamations, which have a clearer and unambiguous interpretation without any significant disadvantages. The responses to the specific questions posed in Section 1 are as follows.

1. *Interpretation of ILRs*: Their interpretation is not clear, nor is it clear what they are measuring, since they depend on the relative values of the parts in the geometric means. An ILR should not be interpreted as the ratio of amalgamations of parts.

2. *Advantages of ILRs*: A full set of so-called ILR balances forms an orthonormal basis of the compositional data vectors, which is a notable mathematical property for reference, but this property is not a requirement for good practice in compositional data analysis. The full set of ILR balances has a non-singular covariance structure that makes it useful for methods that require inversion of the covariance matrix, although, as stated above, a generalized inverse can be used when using the CLRs and there are other alternatives. For methods, such as classification, where the inverse of the covariance matrix is required, it does not matter which transform is used. They all yield the same results.

3. *Disadvantages of ILRs*: Single ILRs have no inherent value as summary variables, nor as responses or explanatory variables in modeling, where their relationships with other variables can be misleading. The changing of basis is of no real value in practice, because the ILRs are so difficult to interpret. ILRs have serious weaknesses with datasets containing many zeroes when the objectives can be expressed in terms of amalgamations; as they require using geometric means after zero replacements, which will often distort the analysis, instead of accommodating the sensible approach of amalgamation.

4. *Alternative use of amalgamations*: Amalgamating parts is a straightforward and understandable way of combining parts in all applications of compositional data analysis, including geochemical applications. Logratios involving amalgamations are just like simple logratios and can contribute, along with simple logratios of single parts, to forming a set of transformations that represents the quasi-totality of the logratio variance. The criticism that they are nonlinear transformations of the parts is of no consequence to the practice of compositional data analysis. 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..

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), as well as by Martín-Fernández et al. (2018). 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.

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. URL https://core.ac.uk/download/pdf/132548276.pdf (last accessed 31 December 2018)

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

Bóna M (2006) A walk through combinatorics: an introduction to enumeration and graph theory. Second Edition. World Scientific Publishing, Singapore

Buccianti, A., Nisi, B. and Raco, B. (2015). From univariate background (baseline) values towards the concept of compositional background (baseline) values. In Thió-Henestrosa S and Martín Fernández JA (eds), Proceedings of the 6th International Workshop on Compositional Data Analysis, chapter 5. URL https://upcommons.upc.edu/bitstream/handle/2117/81949/ProceedingsBook.pdf (last accessed 16 April 2019)

Buccianti A (2015) The FOREGS repository: Modelling variability in stream water on a continental scale revising classical diagrams from CoDA (compositional data analysis) perspective. J Geochem Expl 154:94-104

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

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

Egozcue JJ, 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. Geological Society, London, Special Publications, 264, 67–77

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

Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford University Press, Oxford.

Graeve M, Greenacre MJ (2018) 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. In review at Limnology and Oceanography: Methods.

Greenacre M (2010) Logratio analysis is a limiting case of correspondence analysis. Math Geosc 42:129–134

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

Greenacre M (2018b) Variable selection in compositional data analysis, using pairwise logratios. Math Geosc doi: 10.1007/s11004-018-9754-x

Greenacre M (2018c) Rejoinder to Jamie Morton. CoDa Association Debate#1. URL: https://docs.gestionaweb.cat/1387/10-rejoinder-to-jmorton-greenacre.pdf (last accessed 31 December 2018)

Greenacre M (2018d) A tale of two logratios. CoDa Association Debate#1. URL: https://docs.gestionaweb.cat/1387/11-a-tale-of-two-logratios-mgreenacre.pdf (last accessed 31 December 2018)

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

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

Grunsky EC, Kjarsgaard BA (2016) Recognizing and validating structural processes in geochemical data. In Compositional Data Analysis, JA Martín-Fernández and S Thio-Henestrosa (eds), Springer Proceedings in Mathematics and Statistics, 187. 85-116, 209pp., doi: 10.1007/978-3-319-44811-4_7

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 Geosc 49: 777–796

Krzanowski WJ (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 Geosc 49: 777–796

Legendre P, Legendre L (2012) Numerical ecology. Third edition. Elsevier, Amsterdam

Martín-Fernández JA (2018) Reply to “A Tale of Two Logratios: Rebuttal to Martin” by M. Greenacre. CoDa Association Debate#1. URL: https://docs.gestionaweb.cat/1387/15-reply-to-mg-rebuttal-jamartin.pdf (last accessed 31 December 2018)

Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, 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 and Martín-Fernández JA (eds) Proceedings of 3rd Compositional Data Analysis Workshop. URL: https://dugi-doc.udg.edu/handle/10256/738 (last accessed 12 May 2019)

Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ (2011). The principle of working on coordinates. In Pawlowsky-Glahn V and Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, Chichester.

Morton J, Sanders J, Quinn RA et al. (2017) Balance trees reveal microbial niche differentiation. Ecol Evol Sci 2:e00162-16. https://doi.org/10.1128/mSystems.00162-16 (last accessed 31 December 2018)

Murtagh F (1984) Counting dendrograms: a survey. Discrete Appl Math 7: 191−199

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

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015). vegan: Community Ecology Package. R package version 2.3-2. URL https://CRAN.R-project.org/package=vegan (last accessed 31 December 2018)

Pawlowsky-Glahn V, Monreal-Pawlowsky T, Egozcue JJ (2015) Representation of species composition. In Thió-Henestrosa S and Martín Fernández JA (eds), Proceedings of the 6th International Workshop on Compositional Data Analysis, chapter 22. URL https://upcommons.upc.edu/bitstream/handle/2117/81949/ProceedingsBook.pdf (last accessed 16 April 2019)

R core team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Smith MR (2017) Ternary: An R Package for Creating Ternary Plots. Zenodo, DOI 10.5281/zenodo.1068996

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 KG, Tolosana-Delgado R (2013) Analyzing Compositional Data with R. Springer-Verlag, Berlin

van den Wollenberg AL (1977) Redundancy analysis – an alternative for canonical analysis. Psychometrika 42: 207–219

Washburne A, Silverman J, Leff JW, et al. (2017) Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ DOI 10.7717/peerj.2969

Appendix: Inversion of logratios that involve amalgamations

For example, consider the two amalgamation logratios for the W(ine), B(eer) and S(pirits) data set of Section 3.1:

$$y_1 = \log\left(\frac{S}{B+W}\right) \qquad y_2 = \log\left(\frac{B}{W}\right)$$

Then

$$e^{y_1} = \frac{S}{B+W} \qquad e^{y_2} = \frac{B}{W}$$

Multiplying out gives two linear equations, and the third equation is the condition that the parts sum to 1:

$$\begin{aligned} S - Be^{y_1} - We^{y_1} &= 0 \\ B - We^{y_2} &= 0 \\ S + B + W &= 1 \end{aligned}$$

Solving the following system thus gives the original parts *S, B* and *W*:

$$\begin{bmatrix} 1 & -e^{y_1} & -e^{y_1} \\ 0 & 1 & -e^{y_2} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} S \\ B \\ W \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

In general, for a set of $J-1$ amalgamation logratios, which should involve all J parts, set up a $(J-1)\times J$ matrix **A** describing the pattern of amalgamations in the numerator (using 1s) and in the denominator (using $-e^{y_i}$, where i is the row being coded and y_i is the value of the i-th logratio), all other elements being 0. Then add the last row a vector of 1s. The vector **b** ($J\times1$) consists of $J-1$ 0s and the last element 1. Solve the system of linear equations **Ax** = **b** for **x**.

For the set of ratios in Table 1, shown in the graph of Fig. 5, the set of equations to be solved for the following amalgamation ratios **MgO/Na2O, K2O/P2O5, SiO2/K2O, TiO2/Na2O, SiO2/Na2O, Felsic/Carbonate, MnO/Carbonate, Al2O3/MgO, TiO2/Fe2O3t**, where **Felsic = {SiO2, Al2O3, Na2O, K2O}, Carbonate = {CaO, P2O5}**, is the following:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -e^{y_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -e^{y_2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -e^{y_3} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -e^{y_4} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -e^{y_5} & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & -e^{y_6} & 1 & 1 & -e^{y_6} & 0 \\ 0 & 0 & 0 & 1 & 0 & -e^{y_7} & 0 & 0 & -e^{y_7} & 0 \\ 0 & 0 & 1 & 0 & -e^{y_8} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -e^{y_9} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} SiO2 \\ TiO2 \\ Al2O3 \\ MnO \\ MgO \\ CaO \\ Na2O \\ K2O \\ P2O5 \\ Fe2O3t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

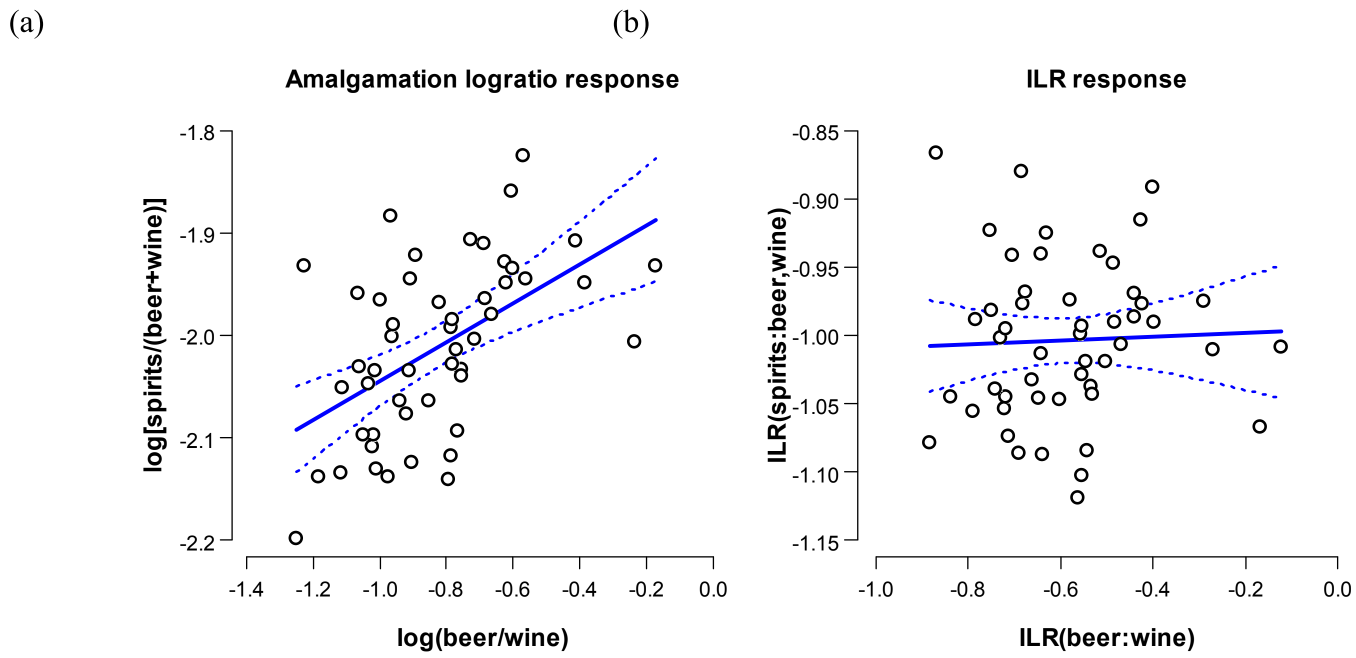

Figure 1 (a) Logratio of spirits relative to (beer+wine) plotted against the logratio of beer relative to wine. The regression shows a significant positive relationship (*p*<0.0001). (b) Isometric logratio of spirits relative to beer and wine plotted against the isometric logratio of beer relative to wine. The regression shows no significant relationship (*p*=0.79). Notice that ILR(beer:wine) = log(beer/wine) / √2 , so the x-axes differ only by this scaling factor.

.

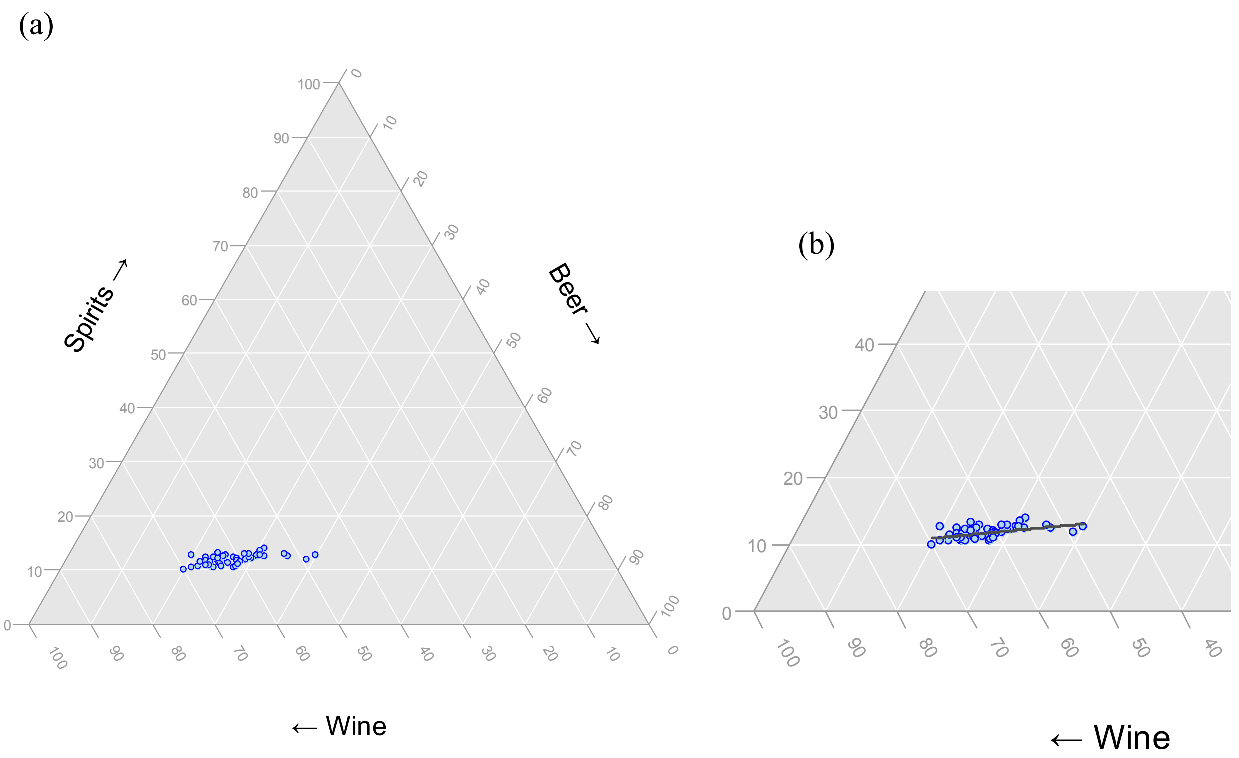

Figure 2. (a) The ternary plot of the three-part compositional data of Table 1. (b) An enlargement of part of the scatterplot in (a), showing the regression line in Fig. 1a transformed back into ternary coordinates as a monotonically increasing curve.

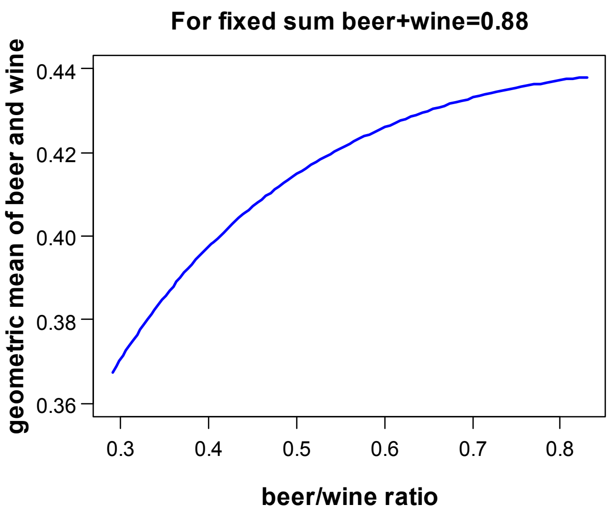

Figure 3. The changing value of the geometric mean according to the ratio of beer to wine, for a fixed value of the sum beer+wine = 0.88. The range of the beer/wine ratio is that found in the data set.

	RATIO	Cum % of var.expl.	Procrustes correlation
1.	MgO/Na2O	69.1	0.831
2.	K2O/P2O5	89.3	0.944
3.	SiO2/K2O	93.4	0.962
4.	TiO2/Na2O	96.6	0.976
5.	SiO2/Na2O	98.7	0.984
6.	Felsic/Carbonate	99.3	0.986
7.	MnO/Carbonate	99.8	0.989
8.	Al2O3/MgO	99.9	0.991
9.	TiO2/Fe2O3t	100.0	0.993

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

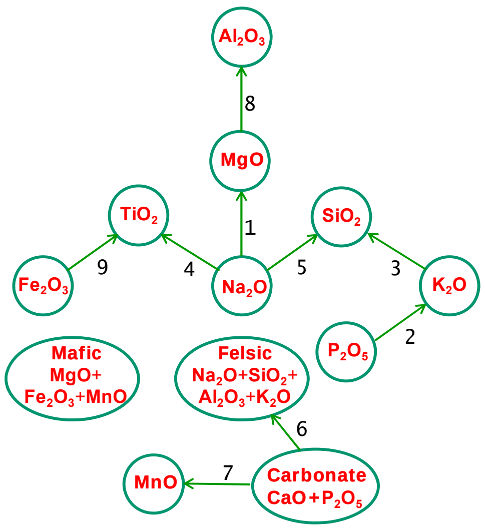

Figure 4: Graph of the ratios in Table 1. The arrows point to the numerator of each ratio. The numbers refer to the ordering of the steps in Table 1. The Mafic amalgamation does not enter into any ratio.

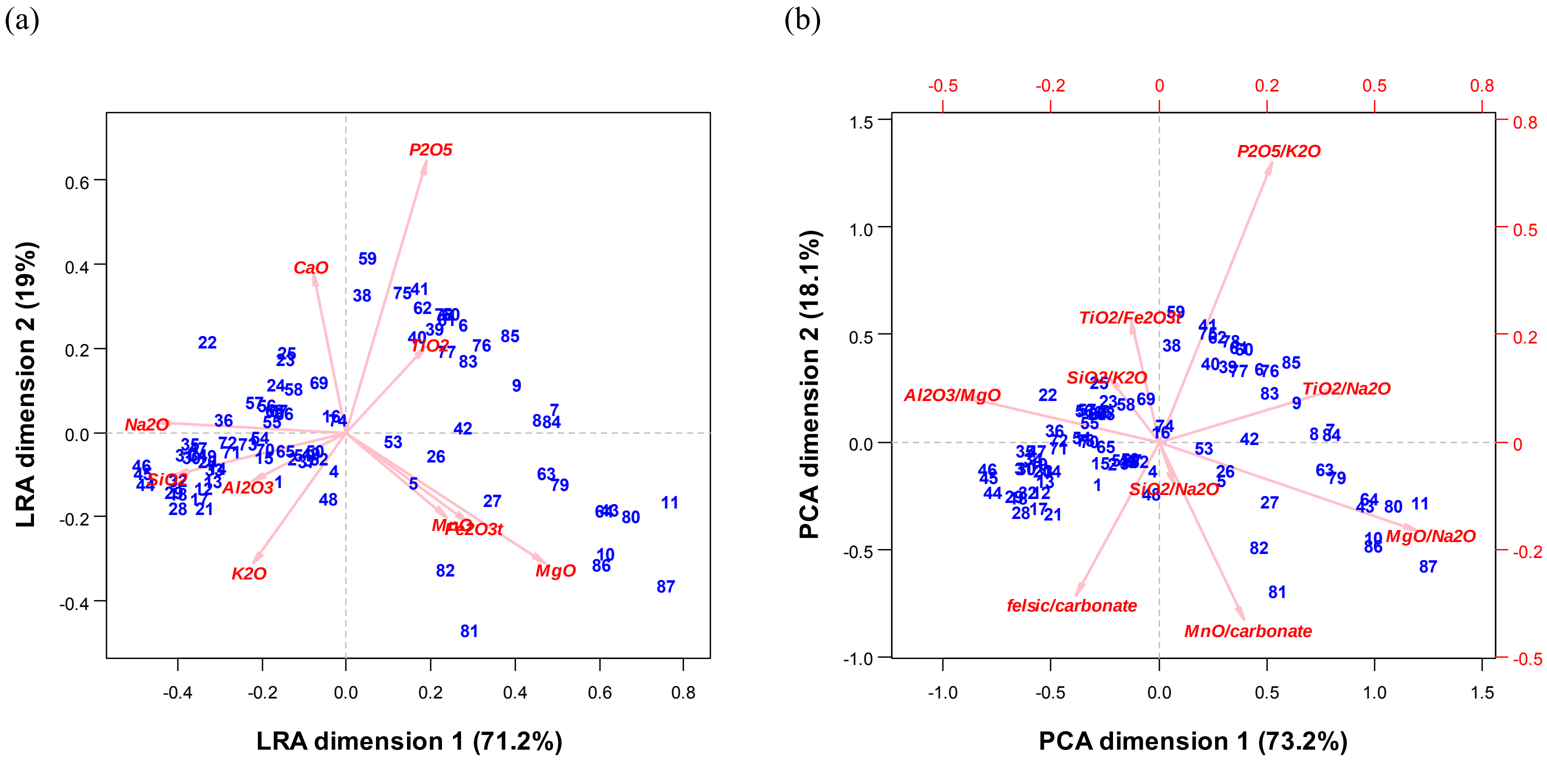

Figure 5: (a) Logratio analysis (LRA) of the Aar massif data set; (b) PCA of the nine selected logratios. The contribution biplot scaling is used.

	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

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.

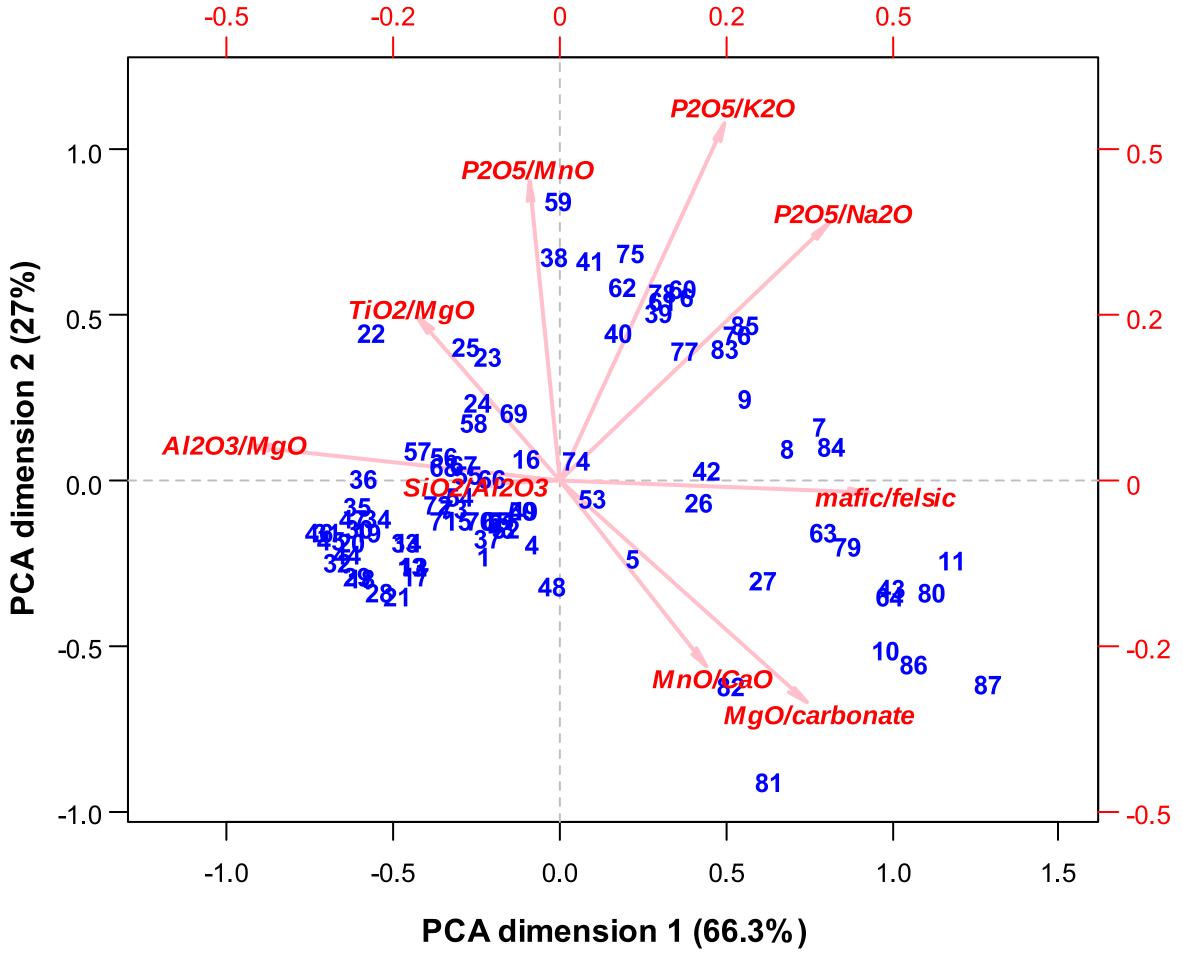

Figure 6: PCA biplot of the 9 logratios selected after the Mafic/Felsic logratio is chosen at the first step