Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1627

The isometric logratio transformation in compositional data analysis: a practical evaluation

Michael Greenacre and Eric Grunsky

January 2019

The isometric logratio transformation in compositional data analysis: a practical evaluation

Michael Greenacre[1] and Eric Grunsky[2]

[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

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.

Abstract: The isometric logratio transformation has been promoted by several authors as the theoretically correct way to contrast groups of parts in a compositional data set. But this transformation has only attractive theoretical properties, the practical benefits of which are questionable. A simple counter-example demonstrates the dangers of using the isometric logratio as a univariate response variable in practice. The study is then extended to a real geochemical data set, where the practical value of isometric logratios is further investigated. When groups of parts are 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. A reduced set of simple logratios of pairs of parts, possibly involving prescribed amalgamations, is adequate in accounting for the variance in a compositional data set, and highlights which parts are driving the data structure.

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

1 Introduction

In the approach to compositional data analysis by Aitchison (1986) based on a data set 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 a general set of values denoted by $x_1, x_2, ..., x_J$, there are $\frac{1}{2} J(J-1)$ unique logratios, a specific example of which is the set of $J-1$ additive logratios (ALRs):

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

where each of the parts x_i, except the last one, is ratioed with respect to the last one. Since the parts can be reordered so that any part is the last one, this gives a total of 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 and only set of J CLRs is defined as:

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

The CLRs 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), but otherwise have no useful interpretation *per se* as representing the parts in univariate analysis.

The isometric logratio (ILR), defined by Egozcue et al. (2003), has been promoted by several authors as the correct way, from a theoretical viewpoint, to transform a compositional data set, to a set of $J-1$ variables, 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 parts, denoted by J_1 and J_2, by defining the logratio of their respective geometric means, with a scaling factor (Egozcue and Pawlowsky-Glahn 2005):

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

where $|J_1|$ and $|J_2|$ denote the number of numerator parts and the denominator parts respectively, that is the cardinalities of the sets J_1 and J_2. So-called ILR "balances" have the property that they are orthonormal, which means they have lengths equal to 1 and scalar products between distinct pairs equal to 0. However, the number of possible sets of these "balances" skyrockets for higher-dimensional problems, equal to $(2J–2)!/(2^{J-1}(J–1)!)$ (see Section 3.1.5) and their choice presents serious practical difficulties.

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 always 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 $J_2 = \{j+1, j+2, ..., J\}$ (see, for example, Hron et al. 2017). Since any set of $J-1$ PLRs depends on the order of the parts, there are again very many such sets of PLRs possible owing to the multitude of permutations of the parts: the number of different sets is equal to $J!/2$ since the last two parts can be in any order.

A PLR, with its single part in the numerator, has the advantage of being able to be expressed and interpreted as an average of pairwise logratios. For example, the first PLR is, apart from the scalar multiplier, equal to $[\log(x_1/x_2) + \log(x_1/x_3)+\cdots\log(x_1/x_J)]/(J-1)$. 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, for example $\log(x_1/x_1)$ for the first CLR, which is equal to 0.

Other ILRs with the same number of parts in the numerator and denominator can also be expressed as averages of pairwise logratios, but the expression is not unique; for example, $\log[(x_1x_2)^{1/2}/(x_3x_4)^{1/2}] = \frac{1}{2}[\log(x_1/x_3)+\log(x_2/x_4)] = \frac{1}{2}[\log(x_1/x_4)+\log(x_2/x_3)]$. When ILRs have more than two and different numbers of parts in the numerator and denominator, no simple expression in terms of pairwise logratios is possible.

Logratios of amalgamations of parts have not been widely used, although – paradoxically – parts, as defined in ILRs or PLRs, are often themselves equivalent to amalgamations. Denoted here by SLR (standing for "summated logratio"), an amalgamation logratio (or "amalgamation balance") is more simply defined 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, just like any other pairwise logratio, and sets of SLRs do not pretend to have any mathematical properties such as orthonormality, nor do they need to. They are simple and easily understandable and interpretable transformations of the compositional data that have specific substantive meaning and value to the researcher. In the approach taken here, the very many number of possible ways parts can be amalgamated is not an issue, since amalgamations and thus amalgamation balances will be defined on substantive

grounds by the specialist who has knowledge of the compositional data and the particular objective of the study. Moreover, amalgamations are often performed in the simplex based on the understanding of the stoichiometric balances.

In this investigation the following questions are considered:

1. What is the interpretation of an ILR? Having defined an ILR transformation on some compositional parts, is it clear what its values are measuring?

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. Section 4 concludes with a discussion and overall conclusion. 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 of the Compositional Data Analysis (CoDa) Association about the use of ILRs[*]. To give the example a context, it is supposed that a researcher is investigating the relationship between the monthly consumption of alcoholic spirits, beer and wine in 50 groups of consumers. The data are proportions and form a three-part composition and the researcher is studying the patterns of consumption, in particular whether the proportion of spirits consumed depends on the relative consumption of beer compared to wine. The data are provided as supplementary online material.

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 in 87 samples of glacial sediment in the Aar Massif, Switzerland (Tolosana-Delgado and Eynatten 2010). The same data have been analysed by van den Boogaart and Tolosana-Delgado (2011) and Martín-Fernández et al. (2018). The oxides are Al_2O_3, CaO, Fe_2O_3t, K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2 and TiO_2. 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 multivariate data set in a meaningful and interpretable way. However, apart from understanding the structure of the parts there is interest in the following three groupings of oxides based on geochemical considerations:

Mafic: MgO, Fe_2O_3, MnO

Felsic: Na_2O, SiO_2, Al_2O_3, K_2O

Carbonate: CaO, P_2O_5

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

Soils, sediments, igneous and metamorphic rocks are comprised 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.

2.3 Methods

Apart from some standard summary statistics and statistical methods such as linear regression, the approach focuses on the analysis of logratios of parts or of amalgamated parts, compared to the use of isometric logratios. Comparisons are made in terms of (i) measurement, substantive meaning and interpretation, (ii) logratio variance explained, (iii) identification of parts 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 investigated, namely what each logratio is actually measuring. Their meaning and interpretation are judged relative to the objectives of the particular study, and it is investigated whether the logratios serve the purpose for which they are intended and whether they have a clear 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 the average of the variances of the CLRs, equivalently the sum or the average of all pairwise logratios (the average option is taken here, as in Greenacre (2018a, b), although it makes no difference to the results when percentages of variance explained are calculated).

Given any explanatory variables, the amount of the total logratio variance that is explained by these variables 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 (Wollenberg 1994), which is 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 objective is rather to quantify how much variance can be explained by a subset of the logratios themselves, and to compare with the corresponding results for ILRs. The rank of a J-part compositional data set is equal to $J-1$ and a subset of $J-1$ "independent" pairwise logratios explains 100 % of the total logratio variance – see, for example, 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 (1987), proposed a stepwise process where logratios are selected that explain a maximum part of the compositional data variability at each step. Identifying such a subset implies identifying a subcompostion of parts that are the main drivers of the patterns in the data. Amalgamations that are pre-defined by the practitioner and thus knowledge-driven groupings of the parts, are included as candidates for forming 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, as described more fully by Greenacre (2018a,b). The percentages of variance explained are computed to show how well chosen 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 $\frac{1}{2}J(J-1)$ pairwise logratios or equivalently between the J CLRs, or between a set of ILR or PLR balances. This logratio space of the samples is sometimes referred to as the "Aitchison geometry". In order to see how closely this multivariate structure can be approximated by a smaller set of logratios or by logratios of 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, usually of dimension two for ease of interpretation. The subspace explains a maximum amount of logratio variance. LRA is equivalent to the PCA of the full set of CLRs, where the resultant biplot shows the J parts with the interpretation focusing on the $\frac{1}{2}J(J-1)$ 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 in the full $(J-1)$-dimensional space. LRA is thus also equivalent to the PCA of the matrix of pairwise logratios. When a reduced subset of logratios is selected, possibly including logratios of amalgamations, their 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 0.29 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")
```

In particular, the functions **LR()** (for computing all pairwise logratios), **ILR()** (for computing a single ILR), **CLR()** (for computing the set of CLRs), **PLR()** (for computing a specified set of PLRs), **LR.VAR()** (for computing the total and individual logratio variances), **STEP()** (for the stepwise selection of logratios) , **PCA()** (for computing the PCA of a set of logratios), and **LRA()** (for computing logratio analysis, i.e. the PCA of the CLRs), as well as several associated plotting functions. 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()** is used to compute the Procrustes correlation between the exact logratio geometry and another one based on chosen logratios. Ternary plots are drawn using function **TernaryPlot()** in the **Ternary** package (Smith 2017).

A minor 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. Instead of the cardinality $|J_1|$, for example, of subset J_1, the combined weights of subset J_1 are used. In this paper equal weighting is considered throughout, and the important weighting issue is avoided since it is not important to the present discussion. Since all parts are considered equally weighted here, 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 as a counter-example to using ILRs

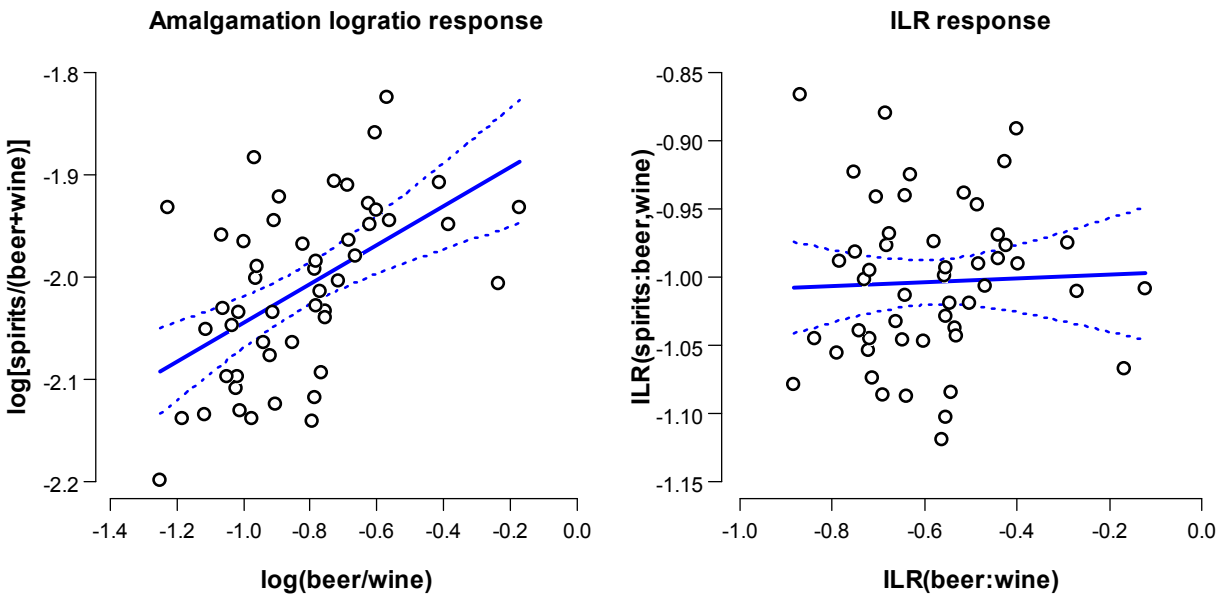
3.1.1 A simple regression analysis according to the study's objective

Since the researcher is principally interested in the proportion of spirits as a response variable, 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 consumed) against the logratio of beer/wine (Figure 1a). The linear regression turns out to be highly significant ($p <$ 0.0001). The effect size in a 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 The ILR alternative

As an alternative, the researcher uses isometric logratios for both variables, namely $\sqrt{2/3}\log(\text{spirits}/(\text{beer}\times\text{wine})^{1/2})$ versus $\sqrt{1/2}\log(\text{beer}/\text{wine})$, the latter logratio being the simple logratio used before along with the scaling constant that is inherent in the ILR definition. In this case the relationship, plotted in Fig. 1b, is no longer significant ($p = 0.79$) and 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.

255 (a) (b)



256

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

263 *3.1.3 The original data in a ternary plot*

264 In an attempt to understand which of the two analyses is reflecting the true situation, the data are
265 visualized in a ternary plot (Fig. 2a). It is clear that, as the sample points are moving from left to
266 right, for increasing beer/wine ratio, there is an increase in the proportion of spirits. This visual
267 display of the original compositional values corroborates the result of the first analysis. From
268 Fig. 1a, over the range of the explanatory variable, the response has predicted a mean value
269 going from –2.1 to –1.9, i.e. from 0.122 to 0.150 in the ratio spirits/ (beer+wine), which
270 corresponds to a change in the proportion of spirits from 0.109 to 0.130. Fig. 2b shows an
271 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.

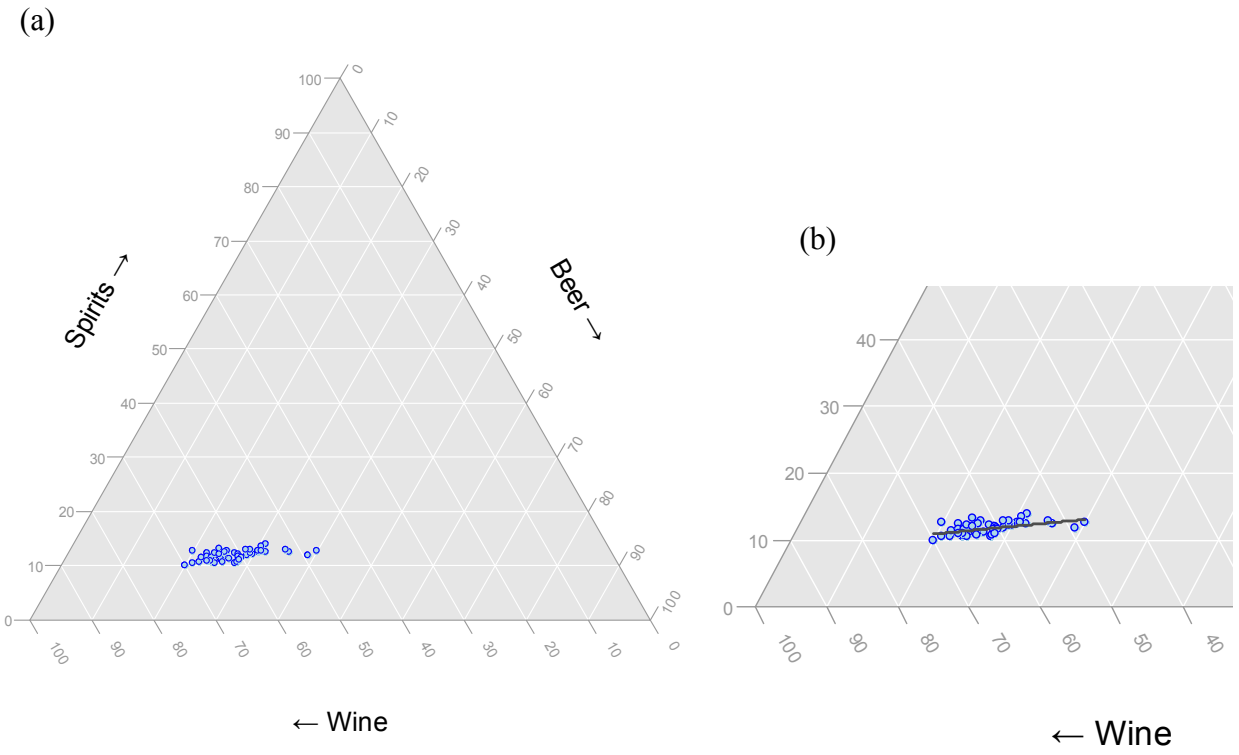


Figure 2. (a) The ternary plot of the three-part compositional data of Table 1. (b) An enlargement of part of the scatterplot in (a), showing the regression line in Fig. 1a transformed back into ternary coordinates as a monotonically increasing curve.

3.1.4 Fundamental difficulties with the interpretation of isometric logratios

Why does the ILR conceal what is obvious in the ternary plot? The reason is that the ILR does not truly contrast the proportion of spirits against that of the combination of beer and wine (or "non-spirits" in this case). Its value depends as well 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. $\sqrt{\text{beer}\times\text{wine}}$, can vary as a function of the ratio beer/wine, for a fixed value of 0.88 of the sum beer+wine.

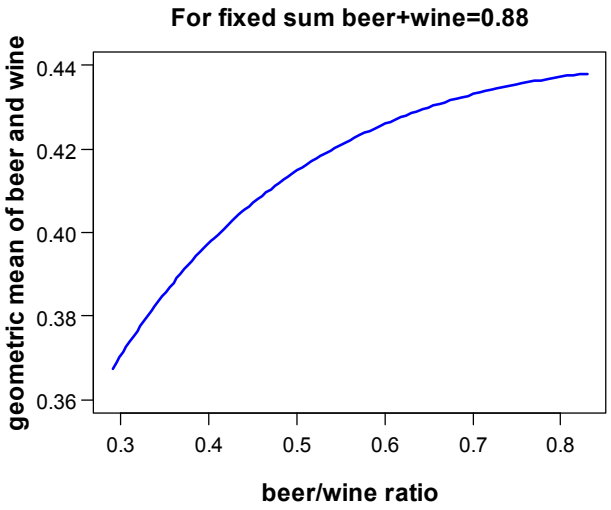


Figure 3. The changing value of the geometric mean according to the ratio of beer to wine, for a fixed value of the sum beer+wine = 0.88. The range of the beer/wine ratio is that found in the data set.

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.

In the online debate referred to before, a data set is given that exhibits the reverse phenomenon, namely where there is no relationship between spirits and the beer/wine ratio, by construction of the data, and where the use of the ILR as a response variable actually creates a statistically significant relationship where none exists – see Greenacre (2018d).

In summary so far, to understand the value of an ILR, it is necessary to understand the relative values of all the parts that constitute the geometric means in its definition. It does not involve simple groupings of the parts, and should in no way be construed as a type of amalgamation of the parts. Its interpretation is already complicated when just two parts are involved in a

geometric mean, as demonstrated in this simple example; when there are many parts, as in most real-life applications, an ILR is a variable with a very complex interpretation. Thinking of it simply as a ratio between two groupings of parts is erroneous and values defined by the ILR do not reflect values that have a clear meaning relative to the definition of the original parts.

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, involving the following 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}, each of which can be represented as a dendrogram (Fig. 4). Any one of these serves the purpose for which ILRs are intended, but it is the first one in Fig. 4a that was chosen to be used in Fig. 1, because of the nature of the researcher's objective.

Notice that the dendrogram in Fig. 4a is drawn differently from the other two, because the contrast between spirits and wine&beer together is not as great as the contrast between wine and beer, whereas it is the opposite in Figs 4b and 4c, where wine has a greater contrast with spirits&beer, and similarly beer versus spirits&wine.

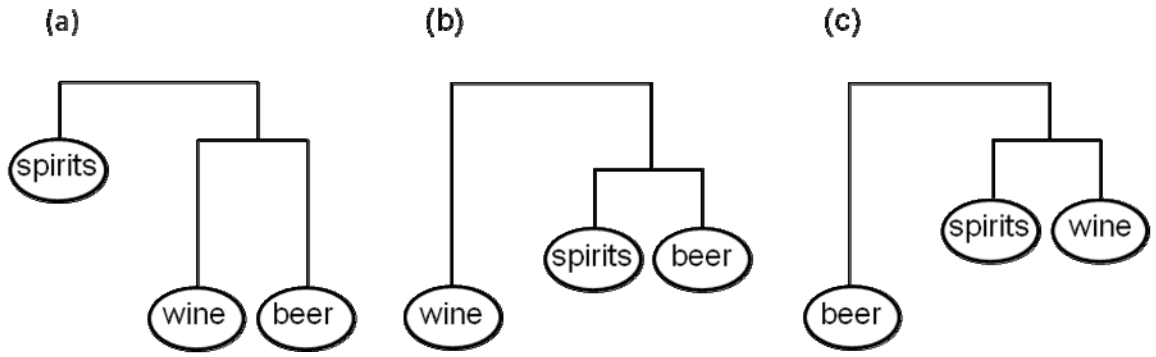


Figure 4. The three sets of contrasts represented graphically. The first one (a) is the one used in the present study, contrasting spirits vs. wine&beer and then wine vs. beer.

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 Fig. 4b as the set of principal balances, because the contrast between wine and spirits&beer is the highest between two subsets of the three parts. But the researcher would not be interested in such a split, since the objective 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 possible dendrograms. However, the number of possibilities becomes astronomical for higher-dimensional problems, since the number of possible dendrograms for a J-part compositional problem is 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 $m=10$, as in the forthcoming geochemical example in Sect. 3.2, which with its 10 parts is 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.

3.2 Isometric and amalgamation logratios in geochemistry

3.2.1 Influence of rare parts

Suppose for sake of illustration that the ratio between the subset SiO_2, Na_2O and MnO relative to P_2O_5 was of interest, and the ILR computed: $\sqrt{3/4}\log\left((SiO_2 \times Na_2O \times MnO)^{1/3}/P_2O_5\right)$. The oxide MnO happened to be the rarest of the parts, measured on average as 0.06 % and with a range across the samples from 0.02 % to 0.24 %. Because rare parts often incur large relative errors, an alternative analysis of the data set excluding MnO was considered, using the corresponding nine-part subcomposition. The ILR of interest was then $\sqrt{2/3}\log\left((SiO_2 \times Na_2O)^{1/2}/P_2O_5\right)$. Fig. 5a compares the ILR including MnO (x-axis) , with a

range of 0.57–1.12 with the ILR where MnO was dropped (y-axis), with a range of 1.04–1.70. This large overall difference in values, which is due to the mere inclusion or exclusion of a very rare part, presents a dilemma to the practitioner, but the dramatic change in the patterns of the values is even more perplexing. It seems that using one or the other might lead to different conclusions when related to other variables of interest. This is similar to the problem described in Sect. 3.1 in that the origin of this difference is the geometric mean in each numerator of the two respective calculations, compared in Fig 5b. The very low value of MnO pulls down the geometric mean from 0.171 (without MnO) to 0.024 (with MnO) on average, but by varying amounts depending on its value and the values of SiO_2 and Na_2O. On the other hand, the amalgamation logratios are hardly affected, because there are only very tiny differences in the amalgamations in the respective numerators (Fig. 5c).

The same problem would be encountered in the CLRs – including or excluding a rare part such as MnO noticeably affects the values of the CLRs, so they are also not useful in univariate analysis.

(a) (b) (c)

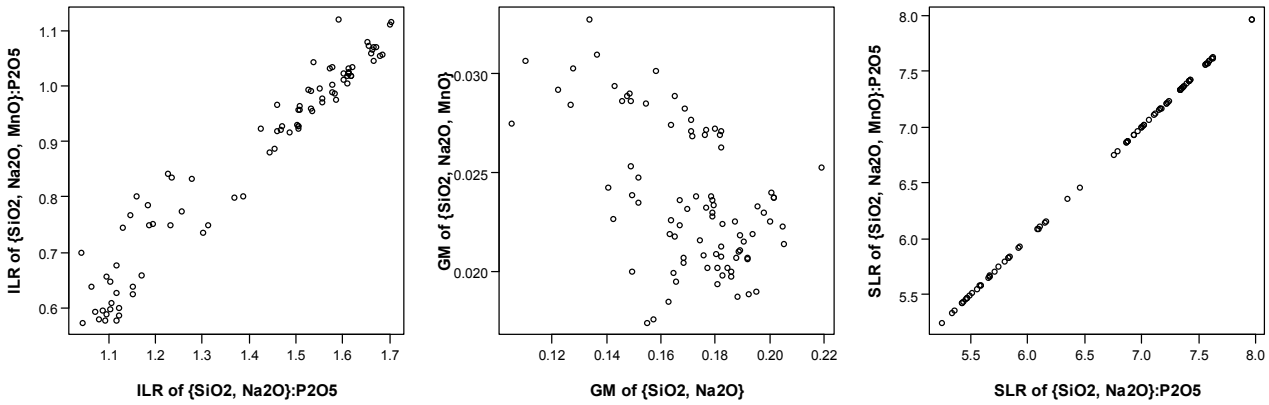


Figure 5. (a) ILR of SiO_2, Na_2O and MnO relative to P_2O_5 (y-axis) versus ILR of SiO_2 and Na_2O (i.e. without MnO) relative to P_2O_5 (x-axis) ; (b) Geometric means involved in the numerators of the two ILRs, respectively, in (a); (c) amalgamation

logratio (SLR) of SiO_2, Na_2O and MnO relative to P_2O_5 (y-axis) versus SLR of SiO_2 and Na_2O (i.e. without MnO) relative to P_2O_5 (x-axis)

3.2.2 Logratios of amalgamations as alternative

Rather than using ILRs, logratios of amalgamations (SLRs) can be used as variables that contrast parts or groups of parts with a simpler definition and interpretation. For example, the complete set of principal balances for the Aar Massif data set, as published by Martín-Fernandez et al. (2018) and reproduced in the supplementary material, was considered by Greenacre (2018b). Logratios using amalgamations instead of geometric means fulfilled the same role for all practical purposes (Greenacre 2018b, Appendix A.3, the relevant part of which is reproduced as supplementary material). These amalgamation balances explained 99.97 % of the variance in the data, only 0.03 % less than the theoretically expected 100% of the ILRs, which is one of their claimed benefits.

Figure 6 shows the principal component analysis of the ILR balances and of the amalgamation balances. There is a strong similarity between the two results, with the Procrustes correlation between the two sample configurations equal to 0.97 in their respective full nine-dimensional spaces, and 0.96 in the two-dimensional spaces of Fig. 6. The advantage of the amalgamation balances is that it is clear what is in each numerator and each denominator, simple sums of parts, as opposed to geometric means which have been shown to depend on the relative values of the parts within each geometric mean. In other words, for the variables O1 to O9 in Fig. 6a it is difficult to pin down exactly what they are measuring, since they are subject to many sources of variation in the geometric means. In Fig. 6b, by contrast, the corresponding amalgamation balances O1 to O9 involve groupings of parts that are simple and clearly comprehensible, just like pairwise logratios, and with a more intuitive interpretation.

(a) (b)

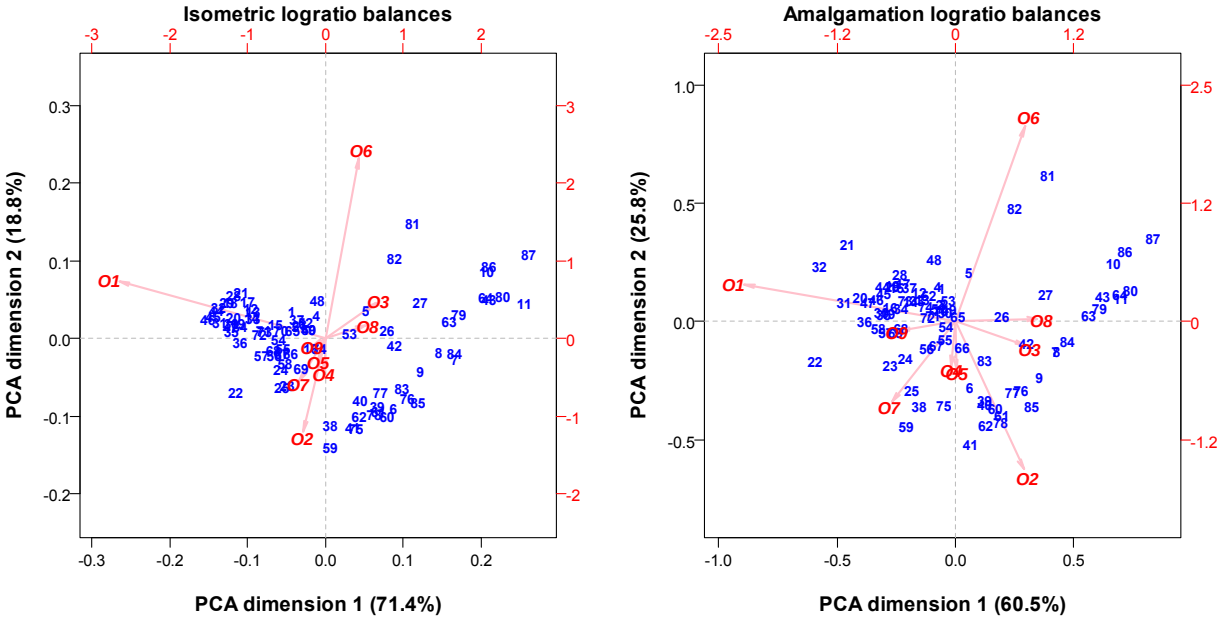


Figure 6. PCA of (a) ILR balances; and (b) SLR balances using same partitioning of parts as for the ILRs. The labels O1 to O9 of the logratios by Martín-Fernandez et al. (2018) are used – see supplementary material for their definitions.

3.2.3 Ratio selection, including ratios of amalgamations

The three amalgamations of Mafic, Felsic and Carbonate (see Section 2.1) were created by summing their compositional values. These amalgamations were allowed 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. The results of the stepwise search are given in Table 1, showing the ratios, their cumulative explained variance, and the Procrustes correlations of the sample configurations with the exact sample configuration. Fig. 7 shows a graph of the solution.

The search was restricted to a set of nine ratios, which is the dimensionality of these 10-part compositional data. The set of logratios involved the prescribed amalgamations of Felsic and Carbonate, and they even appeared together in a ratio. The explained variance was only 0.003% short of 100% (it was equal to 99.997%, rounded to 100.0 in Table 1), which shows that this set

can effectively replace the compositional data set, with a Procrustes correlation of their geometry compared to the exact logratio geometry of 0.993. An even smaller set of ratios can be considered seeing that already with just four ratios more than 95% of the logratio variance is explained, with a Procrustes correlation of 0.976.

	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.

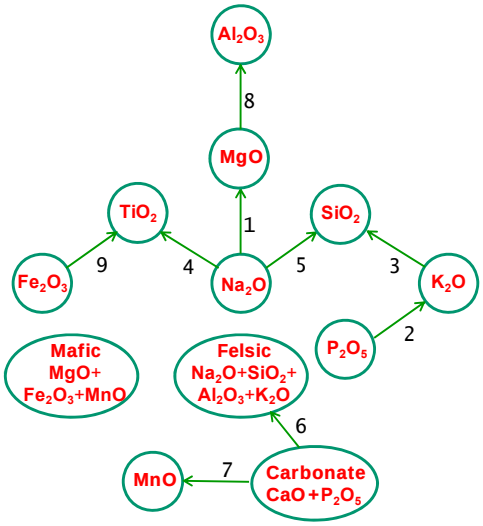


Figure 7: Graph of the ratios in Table 1. The arrows point to the numerator of each ratio. The numbers refer to the ordering of the steps in Table 1. The Mafic amalgamation does not enter into any ratio.

To compare the original geometry of the compositional data set, using 45 pairwise logratios, with that of the reduced set of nine logratios, Fig. 8 shows the logratio analysis of the data set (i.e., the PCA of the centred logratios) and the PCA of the nine chosen logratios – the relative positions of the samples are almost identical. Notice also the similarity with the sample configurations in Fig. 6, apart from the inversion of the second axis. The difference here is that the variables defining the biplot arrows in Fig. 8b are well understood and clearly interpretable, whereas in Fig. 6a the loadings on the axes are difficult to interpret.

(a) (b)

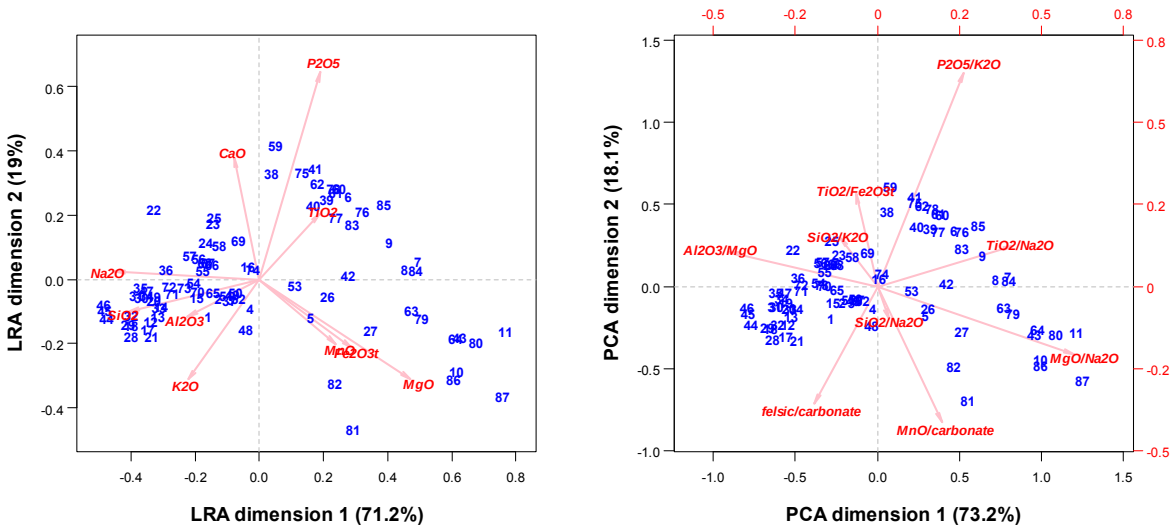


Figure 8: (a) Logratio analysis (LRA) of the Aar massif data set; (b) PCA of the nine selected logratios. The contribution biplot scaling is used.

Notice that in Fig. 8b any ratio can be inverted, in which case the biplot arrow will be reflected with respect to the origin. It might be preferred to show, for example the inverted logratio of Al_2O_3/MgO, i.e. MgO/Al_2O_3, aligned with MgO/Na_2O, which would enhance the recognition of clay minerals in the Aar data.

3.2.4 Knowledge-driven intervention in the stepwise process

The completely automatic stepwise process, giving the results in Table 1, Fig. 7 and Fig. 8b, 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 leaves the opportunity open 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, as shown in Table 1 and Fig. 7, but its components MgO, Fe_2O_3 and MnO are clearly aligned in Fig. 8a and opposing the Felsic parts Na_2O, SiO_2, Al_2O_3, K_2O. From the positions of MgO and Na_2O in Fig. 8a 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).

	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.

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 might be 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 Procrustes correlation of 0.990, compared to 0.993

before. The resulting PCA of the logratios is shown in Fig. 9, where the configuration of samples is practically identical to those in Fig. 8.

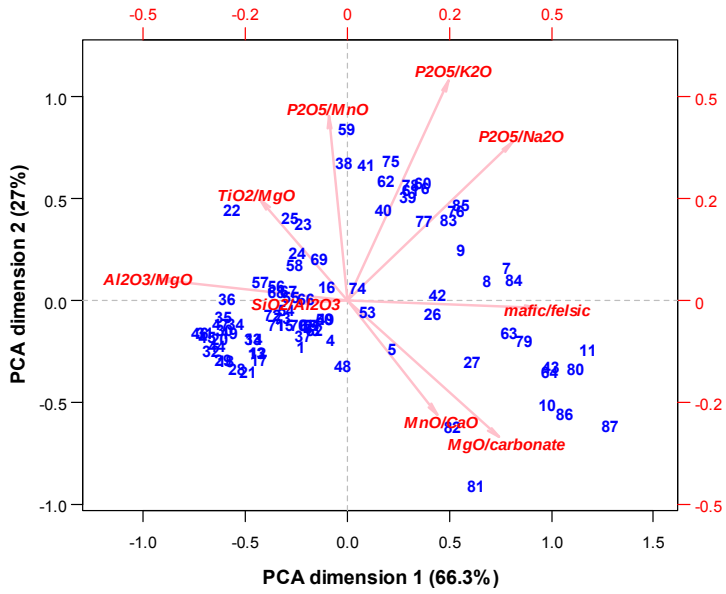


Figure 9: PCA biplot of the 9 logratios selected after the Mafic/Felsic logratio is chosen at the first step

3.2.5 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 logratio variance explained. 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-Fernandez 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 PLR have the same explanatory power because they only differ by a scaling factor.

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

509 **4. Discussion**

510 Various articles on compositional data analysis have established a condition that using ILRs, or
511 at least transformations to orthonormal coordinates, is mandatory for further statistical analysis.
512 For example, Fačevicová et al. (2016) say that compositional vectors need to be expressed in
513 orthonormal coordinates, thereby allowing further processing using standard statistical tools.
514 Kynčlová, Hron and Filzmoser (2017) say that compositional data should be expressed with
515 respect to orthonormal coordinates that "guarantee isometry between the Aitchison geometry and
516 the real space". Mateu-Figueras, Pawlowsky-Glahn and Egozcue (2011) insist on using
517 coordinates with respect to an orthonormal basis based on ILRs but admit that "it is not obvious
518 how to determine which basis is the most appropriate for any given problem".

519 With this strict imposition in mind, several authors have substituted existing practice of using the
520 values of compositional parts or amalgamations of parts in an analysis with the use of single
521 ILR equivalents, with unclear justification. For example, Buccianti (2015) revises a "classical
522 diagram from a compositional data analysis perspective" in a study of water samples. The
523 "classical diagram" is the Gibbs diagram with logarithm of total dissolved solids (TDS, an
524 amalgamation) on the vertical y-axis and, for example, the ratio (Na+Ca)/(Na+K+Ca) or
525 Na/(Na+Ca), on a linear scale on the horizontal axis. To revise this from a compositional data
526 analysis perspective, Buccianti (2015) uses on the y-axis the logarithm of the "balance" of the
527 eight dissolved solids versus the amalgamation of all the other components, i.e. the

528 log(geometric mean of the 8 TDS components/sum of all the other parts), and the "balance"
529 log(Na/Ca) (which is a simple logratio) on the x-axis. Both of these are questionably called ILR
530 balances, especially the first one that has an amalgamation, and not a geometric mean, in the
531 denominator. The justification for this "revision" is unclear, except that now the approach is
532 stated as now being "coherent with the nature of compositional data, thus obtaining a simple tool
533 to be used in a statistical sense, going beyond the descriptive approach" (Buccianti 2015). No
534 attention is drawn to the fact that the value of the geometric mean of the 8 TDS components is
535 minuscule compared to the other water components, and that the value of this "balance" is, for
536 all practical purposes, almost exactly proportional to log(TDS) used in the "classical diagram".
537 The sophistication of using the "balance" implies that there is some benefit over the "classical
538 diagram", but it is difficult to see this benefit and adds an unnecessary complication in the plot's
539 interpretation.

540 Similarly, ILRs have entered the worlds of microbiology and "omics" (e.g. genomics,
541 proteomics or metabolomics) as well, with unclear justification. An example is Morton et al.
542 (2017), where a sparse 88×116 data matrix of counts of 116 microbial species in 88 soil samples
543 is related to several environmental variables. One analysis consists of computing an ILR
544 contrasting 86 species with the other 30, which is the logratio of the two respective geometric
545 means. Plotting this ILR against the pH of the samples, a scatterplot is obtained with a clear
546 negative correlation of −0.91. Using the same data, the much simpler logratio of the two
547 respective amalgamations can be computed, which is the logratio of the sum of the first set of 86
548 species values divided by the sum of the second set (in compositional proportional units, the
549 latter sum is just 1 minus the first). An almost identical pattern is observed when plotted against
550 pH and the correlation is −0.94 − see Greenacre (2018c). A practitioner might wonder why the
551 properties of the ILR balance are given so much prominence, whereas it is clearly just the ratio
552 between the totals of the two sets of species that seems to be operative. Added to this is the

553 difficulty that one would have in explaining what the ILR actually measures, since it depends on
554 the relative values of all 86 species in the numerator and all 30 species in the denominator,
555 making it difficult, if not impossible, to give it a clear interpretation. This leaves the
556 justification of using the ILR as a univariate statistic in considerable doubt.

557 Everything said above applies to the pivot logratios, defined in (4), that have also entered the
558 compositional data analysis literature (see, for example, Hron et al. 2017). These are a special
559 case of ILR balances with only one part in the numerator and the geometric mean of a set of
560 parts in the denominator: first part vs. the rest, the second part vs. all the others except the first
561 part, third part vs. all the others except the first and second parts, etc...). Pivot balances depend
562 on the ordering of the parts, so there are fewer sets of them compared to ILR balances: for $J = 10$
563 there are $J!/2 = 10!/2 = 1\,814\,400$ possible sets of pivot balances. The supposed advantage of a
564 pivot balance is that it is proportional to the average of pairwise logratios of the numerator part
565 with all the parts in the denominator (e.g., in the context of the simple example of Sect. 3.1, the
566 pivot logratio of Spirits versus Beer and Wine is the average of the logratios log(Spirits/Beer)
567 and log(Spirits/Wine). This again presents an interpretability problem: what does the average of
568 set of pairwise logratios measure, and how should a practitioner understand its meaning in the
569 context of the particular compositional data set being investigated?

570 There are mainly two professed benefits of ILRs: first, the definition of a new set of orthonormal
571 coordinates for the data; and second, their role in grouping the parts. Both these benefits are
572 debatable, and can impose unrealistic limitations in practical applications.

573 Concerning the geometry, in order to provide a new set of coordinates for a J-part compositional
574 data set, a set of $J-1$ ILR balances needs to be defined (see Sect. 2.4). These provide an
575 isometric transformation of the compositional data to a $(J-1)$-dimensional vector space defined
576 by the ILR coordinates. The logratio distances between the samples are identical to the
577 Euclidean distances between the ILR coordinates, and this property is the main justification for

their definition. The question that might be asked by the practitioner is: why is it necessary to reproduce the geometry of the samples exactly, using transformed variables that have a problematic interpretation? Or, putting it another way, is there a way of reproducing the geometry of the samples with a good approximation, using new variables that do have a clear and geochemically meaningful interpretation? As shown in the present study, the answer to the first question is that it is not necessary to reproduce the geometry with mathematical exactitude, and to the second question, yes, there is an easier and more useful way.

It is not necessary to satisfy the exactness of the isometric transformation to ILR coordinates because it is obvious that in practice the data themselves are not exact, but subject to all types of additional variability in the form of measurement error and random variation, so there is no reason to think that the exact geometry of the samples is the correct one. Hence, the quest to transform all the variability in the data set, including the random component, to another space in an exact manner seems to lack relevance. For example, the principal component analysis (PCA) of the CLRs, called logratio analysis or LRA (Greenacre 2010, 2018) and used in Fig. 8a, is frequently 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 initially by appropriate and meaningful transformations rather than using ILRs. Selecting key logratios and possible ratios of amalgamations presents a more justifiable alternative for the practitioner.

Concerning the second claimed benefit, the ILR balances are promoted as a meaningful grouping of the parts, for example that they are "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" as single variables. ILRs, although mathematically attractive, are complicated transformations of the data and it is not clear what their values are actually measuring. Several authors believe that ILRs are validly comparing groups of parts, for example 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". Any claim or suggestion that ILRs are contrasting groups of parts in the sense of amalgamating them (or averaging them, which is equivalent) should be viewed with the utmost skepticism.

Amalgamations of parts are a more intuitive and interpretable alternative to geometric means. The specialist has knowledge about the possible models that the empirical relationships might reveal. Amalgamations fundamentally rely on this knowledge. Alternatively, 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: "note that using 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".

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). The imposition of this mathematical condition restricts the practitioner from using alternatives that make perfect substantive sense in practical applications. As demonstrated in this study,

amalgamations can be used profitably to represent geochemical processes and their performance in the form of logratios can be gauged objectively by the variance accounted for in a compositional data set.

John Aitchison himself is quoted as saying 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. 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), and the associated log-contrasts as a more intuitive and practical way of dealing with the problem of grouping of parts (Aitchison 2003). These remarks are faithfully implemented in the present paper by demonstrating that amalgamations function well in compositional data analysis, 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. Certain amalgamations can even be forced into the selection, because of their important role in the context of the study – for example, 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 subsets of fatty acids would never be defined by biochemists as geometric means. Moreover, 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. It is often the case that a ratio which is not explaining the exact maximum variance, but slightly less, is a more meaningful and justifiable

651 ratio to select at a particular step, in which case this slightly sub-optimal ratio would be preferred
652 by the specialist.

653 An additional claimed benefit of the ILR transformation is that it reduces the J-part data set of
654 rank to $J-1$ to one of $J-1$ variables that are linearly independent, and whose covariance matrix
655 is easily inverted in matrix computations such as multiple regression analysis and computation of
656 Mahalanobis distances. But this is not an additional benefit, because the generalized inverse can
657 be used directly on the singular covariance matrix of the J CLRs, for example, with identical
658 results. Moreover, any set of additive logratios (ALRs) has a nonsingular covariance matrix and
659 induces the same Mahalanobis distances as those obtained using ILRs.

660 **5. Conclusion**

661 Our overall conclusion is that isometric logratios (logratios of geometric means) present
662 significant barriers in the practice of compositional data analysis and can be substituted by
663 simple logratios and logratios of amalgamations, which have a clearer and unambiguous
664 interpretation. The responses to the specific questions posed as objectives in Section 1 are as
665 follows.

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

669 2. *Advantages of ILRs*: A full set of so-called ILR balances forms an orthonormal basis of the
670 compositional data vectors. This is a notable mathematical property, but the practical
671 consequences are not interesting because of the interpretability problems of these
672 transformations, making the changing of basis of no real value in practice. The full set of ILR
673 balances has a nonsingular covariance structure that makes it useful for methods that require

674 inversion of the covariance matrix, although, as stated above, a generalized inverse can be
675 used in the case of CLRs.

676 3. *Disadvantages of ILRs*: Single ILRs have no inherent value as summary variables, nor as
677 responses or explanatory variables in a regression analysis – examples can easily be found
678 where their relationships with other variables are found to be misleading or counter-intuitive..

679 4. *Alternative use of amalgamations*: Amalgamating parts is a straightforward and
680 understandable way of combining parts in all applications of compositional data analysis,
681 including geochemical applications. Logratios of amalgamations are just like simple logratios
682 and thus easy to interpret and can contribute, along with simple logratios of single parts, to
683 forming a set of transformations that represents the quasi-totality of the variance in a
684 compositional data set. The criticism that they are nonlinear transformations of the parts is of
685 no consequence to the practice of compositional data analysis. Amalgamations do impose a
686 model as determined by the researcher, which is a limitation. However, the researcher can use
687 different amalgamations to extract different processes.
688
689

690 **Acknowledgments**

691 The authors thank Raimon Tolosana-Delgado for supplying the Aar Massif data set, which is
692 used in the book by van den Boogaart and Tolosana-Delgado (2013), as well as by Martín-
693 Fernández et al. (2018). The stimulating and open debate within the CoDA community and the
694 refinement of the developing field of compositional data analysis have been an inspiration for
695 this manuscript.

696

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London. Reprinted in 2003 with additional material by Blackburn Press

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 (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 (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.

719 Greenacre M (2010) Logratio analysis is a limiting case of correspondence analysis. Math Geosc
720 42:129–134

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

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

725 Greenacre M (2018c) Rejoinder to Jamie Morton. CoDa Association Debate#1. URL:
726 https://docs.gestionaweb.cat/1387/10-rejoinder-to-jmorton-greenacre.pdf (last accessed 20
727 December 2018)

728 Greenacre M (2018d) A tale of two logratios. CoDa Association Debate#1. URL:
729 https://docs.gestionaweb.cat/1387/11-a-tale-of-two-logratios-mgreenacre.pdf (last accessed 20
730 December 2018)

731 Greenacre MJ, Lewi PJ (2009) Distributional equivalence and subcompositional coherence in
732 the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif
733 26: 29–64

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

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

740 Krzanowski WJ (1987) Selection of variables to preserve multivariate data structure, using
741 principal components. Appl Statist 36: 22–33

742 Kynčlova P, Hron K, Filzmoser P (2017) Correlation between compositional parts based on
743 symmetric balances. Math Geosc 49: 777–796

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

745 Martín-Fernández JA (2018) Reply to "A Tale of Two Logratios: Rebuttal to Martin" by M.
746 Greenacre. CoDa Association Debate#1. URL: https://docs.gestionaweb.cat/1387/15-reply-
747 to-mg-rebuttal-jamartin.pdf (last accessed 20 December 2018)

748 Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2018) Advances
749 in principal balances for compositional data. Math Geosci 50: 273–298

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

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

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

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

759 Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos
760 P, Stevens MHH, Wagner H (2015). vegan: Community Ecology Package. R package version
761 2.3-2. https://CRAN.R-project.org/package=vegan

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

764 Smith MR (2017) Ternary: An R Package for Creating Ternary Plots. Zenodo, DOI
765 10.5281/zenodo.1068996

766 Tolosana-Delgado R, von Eynatten H (2010) Simplifying compositional multiple regression:
767 application to grain size controls on sediment geochemistry. Comput Geosci 36:577–589
768 van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing Compositional Data with R.
769 Springer-Verlag, Berlin
770 Wasburne A, Silverman J, Leff JW, et al. (2017) Phylogenetic factorization of compositional
771 data yields lineage-level associations in microbiome datasets. PeerJ DOI 10.7717/peerj.2969
772 Wollenberg AL (1977) Redundancy analysis – an alternative for canonical analysis.
773 Psychometrika 42: 207–219

774 **SUPPLEMENTARY MATERIAL**

775

776 1. Simple three-part data set supplied by Martín-Fernández (2018)

777 2. Definition of parts in principal balances computed by Martín-Fernández et al. (2018, Table 3)

778 3. Cumulative explained variances of ILRs and corresponding SLRs in explaining variance of
779 the Aar Massif data, reproduced from Greenacre (2018b, Appendix A.3)

780 4. R script for computing the results

781

Simple three-part data set supplied by Martín-Fernández (2018)

	wine	beer	spirits
1	0.570	0.306	0.125
2	0.622	0.265	0.113
3	0.557	0.317	0.125
4	0.700	0.200	0.100
5	0.619	0.253	0.128
6	0.550	0.312	0.139
7	0.638	0.250	0.113
8	0.655	0.239	0.106
9	0.637	0.243	0.119
10	0.591	0.290	0.119
11	0.675	0.198	0.127
12	0.630	0.254	0.116
13	0.568	0.305	0.127
14	0.474	0.399	0.126
15	0.674	0.220	0.106
16	0.605	0.275	0.120
17	0.616	0.278	0.105
18	0.652	0.224	0.124
19	0.653	0.233	0.114
20	0.629	0.239	0.132
21	0.656	0.228	0.116
22	0.656	0.236	0.108
23	0.609	0.268	0.123
24	0.493	0.389	0.118
25	0.649	0.245	0.105
26	0.604	0.276	0.121
27	0.602	0.283	0.116
28	0.521	0.355	0.125
29	0.580	0.299	0.121
30	0.635	0.254	0.111
31	0.685	0.210	0.105
32	0.603	0.279	0.118
33	0.624	0.251	0.125
34	0.613	0.280	0.107
35	0.523	0.347	0.129
36	0.582	0.294	0.123
37	0.580	0.291	0.129
38	0.606	0.277	0.116
39	0.667	0.220	0.114
40	0.649	0.235	0.116
41	0.602	0.283	0.115
42	0.587	0.284	0.129
43	0.641	0.236	0.123
44	0.636	0.244	0.120
45	0.560	0.306	0.135
46	0.564	0.310	0.126
47	0.635	0.258	0.107
48	0.654	0.237	0.109
49	0.608	0.282	0.110
50	0.660	0.231	0.109

Definition of parts in principal balances computed by Martín-Fernández et al. (2018, Table 3), in the form of numerator parts | denominator parts

O1: Al_2O_3, K_2O, Na_2O, SiO_2 | Fe_2O_3t, MgO, MnO, P_2O_5

O6: Fe_2O_3t, MgO, MnO | P_2O_5

O3: Al_2O_3, K_2O | Na_2O, SiO_2

O2: CaO | Al_2O_3, K_2O, Na_2O, SiO_2, Fe_2O_3t, MgO, MnO, P_2O_5

O8: MgO | MnO

O7: TiO2 | Al_2O_3, K_2O, Na_2O, SiO_2

O5: Na_2O | SiO_2

O4: Al_2O_3 | K_2O

O9: Fe_2O_3t | MgO, MnO

Cumulative explained variances of ILRs and corresponding SLRs in explaining variance of the Aar Massif data, reproduced from Greenacre (2018b, Appendix A.3)

	ILR Principal Balances[2]	Amalgamation Balances[3]
PB	CumVar	CumVar
O1	0.7067	0.6901
O6	0.8940	0.8741
O3	0.9317	0.9292
O2	0.9507	0.9510
O8	0.9757	0.9729
O7	0.9944	0.9938
O5	0.9982	0.9966
O4	0.9994	0.9989
O9	1.0000	0.9997

R script for computing the results

```
### Install easyCODA package from CRAN in usual way or from R-Forge:

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

### (version 0.29 of easyCODA was used here)

### 3-part data set of wine, beer and spirits assumed in data frame 'wbs'

### Model and plot of amalgamation logratio vs. logratio, and ILR vs. logratio
##  transformations
#       alog = amalgamation logratio
#       plog = single logratio
#       ilog = isometric logratio

attach(wbs)
alog <- log(spirits/(beer+wine))
plog <- log(beer/wine)
# ILR "by hand" classical definition
ilog <- sqrt(2/3) * log(spirits / sqrt(beer*wine))  # using counts
# using function ILR in easyCODA weights are used
# so the result for option weight=FALSE (equal weighting) is
# the classical definition divided by sqrt(number of parts) = sqrt(3)
ILR(wbs, numer=3, denom=c(1,2), weight=FALSE)$LR
# compare above result with
ilog/sqrt(3)

##  Model with amalgamation logratio and plotting of Fig. 1a

mod1 <- lm(alog ~ plog)
summary(mod1)
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -1.85503    0.03796 -48.872  < 2e-16 ***
# plog         0.18915    0.04413   4.286 8.72e-05 ***

mod1.pred <- predict(mod1, type="response", se.fit=T,
                     newdata=data.frame(plog=seq(range(plog)[1], range(plog)[2],length=100)))

par(mar=c(4.2,4,3,1), font.lab=2, cex.lab=1.2, mgp=c(2.7,0.7,0), las=1, mfrow=c(1,2))
plot(plog, alog, type="n", bty="n", xaxt="n", yaxt="n", xlab="log(beer/wine)",ylab="log[spirits/(beer+wine)]",
     xlim=log(c(0.25,1)), ylim=c(-2.2,-1.8), las=1, main="Amalgamation logratio response")
axis(1)
axis(2)
segments(range(plog)[1], mod1.pred$fit[1], range(plog)[2], mod1.pred$fit[100], lwd = 3, col = "blue")
lines(seq(range(plog)[1], range(plog)[2],length=100), mod1.pred$fit+1.96*mod1.pred$se, lty = 3, lwd = 2,
      col = "blue")
lines(seq(range(plog)[1], range(plog)[2],length=100), mod1.pred$fit-1.96*mod1.pred$se, lty = 3, lwd = 2,
      col = "blue")
symbols(plog, alog, fg="black", bg="white", circles=rep(0.02, length(alog)), inches=F, add=T, lwd=2)

##  Model with isometric logratio and plotting of Fig. 1b
#   Notice that plog is now divided by sqrt(2) to be the ILR definition

ilog2 <- plog/sqrt(2)
mod2 <- lm(ilog ~ ilog2)
summary(mod2)
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -0.99574    0.03066 -32.473   <2e-16 ***
# ilog2        0.01366    0.05042   0.271    0.788
```

```
mod2.pred <- predict(mod2, type="response", se.fit=T,
                     newdata=data.frame(ilog2=seq(range(ilog2)[1], range(ilog2)[2],length=100)))

plot(ilog2, ilog, type="n", bty="n", xaxt="n", yaxt="n", xlab="ILR(beer:wine)", ylab="ILR(spirits:beer,wine)",
     xlim=c(-1.0, 0), ylim=c(-1.15,-0.85), las=1, main="ILR response")
axis(1)
axis(2)
segments(range(ilog2)[1], mod2.pred$fit[1], range(ilog2)[2], mod2.pred$fit[100], lwd = 3, col = "blue")
lines(seq(range(ilog2)[1], range(ilog2)[2],length=100), mod2.pred$fit+1.96*mod2.pred$se, lty = 3, lwd = 2,
      col = "blue")
lines(seq(range(ilog2)[1], range(ilog2)[2],length=100), mod2.pred$fit-1.96*mod2.pred$se, lty = 3, lwd = 2,
      col = "blue")
symbols(ilog2, ilog, fg="black", bg="white", circles=rep(0.014, length(ilog)), inches=F, add=T, lwd=2)


### Ternary plot (Figure 2)
require(Ternary)

#   full plot (Fig.2a)
par(mfrow=c(1, 1), mar=rep(0.3, 4))
TernaryPlot(alab="Spirits \u2192", blab="Beer \u2192", clab="\u2190 Wine",
            point='up', lab.cex=1.5, grid.minor.lines = 0,
            grid.lty='solid', col=rgb(0.9, 0.9, 0.9), grid.col='white',
            axis.col=rgb(0.6, 0.6, 0.6), ticks.col=rgb(0.6, 0.6, 0.6),
            padding=0.08)
AddToTernary(points, wbs[,c(3,2,1)], pch=21, cex=0.9, col="blue", bg="lightblue")

#   partial plot (Fig. 2b) with regression model in Fig. 1a back-transformed
for(i in 1:100) {
  beerwine.seq[i] <- beer.seq[i]/wine.seq[i]
  spirits.seq[i]  <- -1.855 + 0.1892 * log(beerwine.seq[i])
  spirits.seq[i]  <- exp(spirits.seq[i])/(1+exp(spirits.seq[i]))
}
wbs.add <- cbind(wine.seq, beer.seq, spirits.seq)
TernaryPlot(xlim=c(-0.3,-0.08), ylim=c(0,0.17), alab="Spirits \u2192", blab="Beer \u2192", clab="\u2190 Wine",
            point='up', lab.cex=1.5, grid.minor.lines = 0, grid.lty='solid', col=rgb(0.9, 0.9, 0.9),
            grid.col='white', axis.col=rgb(0.6, 0.6, 0.6), ticks.col=rgb(0.6, 0.6, 0.6), padding=0.22)
AddToTernary(points, wbs[,c(3,2,1)], pch=21, cex=0.9, col="blue", bg="lightblue")
AddToTernary(lines, wbs.add[,c(3,2,1)], lwd=2, col="gray30")


### For a fixed value of beer+wine=0.88, how does the geometric mean vary? (Figure 3)

beer.sim <- seq(range(beer)[1], range(beer)[2], length=100)
wine.sim <- 0.88-beer.sim
beerwine.gm.sim <- sqrt(beer.sim*wine.sim)

par(mar=c(4.2,4,3,1), font.lab=2, cex.lab=1.2, mgp=c(2.7,0.7,0), mfrow=c(1,1), las=1)
plot(beer.sim/wine.sim, beerwine.gm.sim, type="n", xlab="beer/wine ratio", ylim=c(0.36,0.44),
     ylab="geometric mean of beer and wine", main="For fixed sum beer+wine=0.88")
lines(beer.sim/wine.sim, beerwine.gm.sim, lwd=2, col="blue")

### Aar Massif data assumed in data.frame 'aar'

### Comparison of ILRs and SLRs (Figure 4) for aar and aar.sub (without MnO)

aar.sub <- aar[,-4]
aar.sub <- aar.sub / apply(aar.sub, 1, sum)

# SiO2, Na2O, MnO, P2O5 are numbers 1, 7, 4, 9 in aar
# SiO2, Na2O, P2O5 are numbers 1, 6, 8 in aar.sub
```

```
# remember that ILR in easyCODA divides the classic definition by the square root
# of the number of parts in the (sub)composition

ilr1 <- ILR(aar, numer=c(1,7,4), denom=9, weight=FALSE)$LR
ilr1.sub <- ILR(aar.sub, numer=c(1,6), denom=8, weight=FALSE)$LR
par(mar=c(4.2,4,1,2), font.lab=2, cex.lab=1.3, mgp=c(3,0.7,0), mfrow=c(1,3), las=1)
plot(ilr1.sub, ilr1, xlab="ILR of {SiO2, Na2O}:P2O5", ylab="ILR of {SiO2, Na2O, MnO}:P2O5")

gm1 <- (aar[,1]*aar[,7]*aar[,4])^(1/3)
gm1.sub <- (aar.sub[,1]*aar.sub[,6])^(1/2)
plot(gm1.sub, gm1, xlab="GM of {SiO2, Na2O}", ylab="GM of {SiO2, Na2O, MnO}")

slr1 <- SLR(aar, numer=c(1,7,4), denom=9, weight=FALSE)$LR
slr1.sub <- SLR(aar.sub, numer=c(1,6), denom=8, weight=FALSE)$LR
plot(slr1.sub, slr1, xlab="SLR of {SiO2, Na2O}:P2O5", ylab="SLR of {SiO2, Na2O, MnO}:P2O5")

### Stepwise selection of ratios, including the three amalgamations

#   Define the amalgamations and add them to the set of 10 parts
mafic      <- apply(aar[,c(5,10,4)], 1, sum)
felsic     <- apply(aar[,c(7,1,3,8)], 1, sum)
carbonate  <- apply(aar[,c(6,9)], 1, sum)
aar.amalg  <- cbind(aar, mafic, felsic, carbonate)

#   Perform the stepwise analysis
aar.step   <- STEP(aar.amalg, aar, weight=FALSE)

# Table 1
cbind(aar.step$ratios,
      round(100*aar.step$R2max,1),
      round(aar.step$pro.cor,3))

### LRA of original data and PCA of the 9 selected logratios

rownames(aar) <- 1:nrow(aar)
par(mar=c(4.2,4,3,3), font.lab=2, cex.lab=1.2, mgp=c(2.7,0.7,0), las=1, mfrow=c(1,2))

### LRA of original data and PCA of the 9 selected logratios

rownames(aar) <- 1:nrow(aar)
par(mar=c(4.2,4,3,3), font.lab=2, cex.lab=1.2, mgp=c(2.7,0.7,0), las=1, mfrow=c(1,2), cex.axis=0.8)

#   LRA (logratio analysis, = PCA of the CLRs)
aar.lra <- LRA(aar, weight=FALSE)
PLOT.LRA(aar.lra, map="contribution")

#   PCA of the selected logratios
rownames(aar.step$logratios) <- 1:nrow(aar)
# invert K2O/P2O5
aar.step$logratios[,2] <- -aar.step$logratios[,2]
colnames(aar.step$logratios)[2] <- "P2O5/K2O"
aar.ratios.pca <- PCA(aar.step$logratios, weight=FALSE)
PLOT.PCA(aar.ratios.pca, map="contribution", axes.inv=c(1,-1), rescale=2)

#   Procrustes correlation of two configurations of samples in two dimensions
protest(aar.ratios.pca$rowpcoord[,1:2], aar.lra$rowpcoord[,1:2])$t0
# [1] 0.9971076

#   Procrustes correlation of full-space geometry of samples
protest(aar.ratios.pca$rowpcoord, aar.lra$rowpcoord)$t0
# [1] 0.993197
```