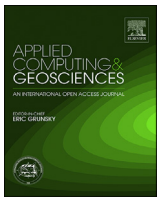
Applied Computing and Geosciences 5 (2020) 100017

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.journals.elsevier.com/applied-computing-and-geosciences



Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation

Michael Greenacre

Department of Economics and Business, Universitat Pompeu Fabra, And Barcelona Graduate School of Economics, Ramon Trias Fargas 25-27, Barcelona 08005, Spain

ARTICLE INFO

Keywords:
Amalgamation
Balance
Clustering
Compositional data
Inverse transformation
Logratio transformation

ABSTRACT

Amalgamations (i.e. summing) of parts can be included as new parts in compositional data analysis, and logratios can then be formed using these amalgamations as well as any of the individual parts themselves. In the first contribution of this paper, a comparison is made of the performance of different logratio transformations in explaining the structure of a geochemical data set – some of the transformations include amalgamations. The second contribution shows how amalgamations suggest a natural way of clustering compositional data, leading to a new clustering algorithm for compositional data. The third contribution deals with the inverse transformation where amalgamations are involved in the logratios. In the case of a linearly independent set of logratios of parts and amalgamated parts, consisting of one less logratio than the number of compositional parts and where each part is present in at least one logratio, it is possible to back-transform this set of logratios to the original parts. The solution is defined by a set of linear equations. A special case is a set of linearly independent pairwise logratios of parts, which are also invertible back to the original parts.

1. Introduction

In the approach to compositional data analysis by Aitchison (1986) 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, abbreviated as LRs, which have been used since the earliest work of Aitchison. For data involving J compositional parts with values denoted by x_1, x_2, …, x_J, there are ½ J (J − 1) unique LRs.

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 ("summated logratio"), an amalgamation logratio is simply defined, for two commonly (but not necessarily) non-overlapping subsets of parts, J_1 and J_2, as:

$$\mathrm{SLR}(J_1, J_2) = \log \frac{\sum_{j \in J_1} x_j}{\sum_{j \in J_2} x_j} \tag{1}$$

Notice that an SLR is a logratio, without any scaling factor, just like any other LR. An SLR can be a simple LR if there are single parts in the numerator and the denominator, so in this sense SLRs represent the general logratio transformation. Amalgamations are often constructed in chemistry based on the understanding of the stoichiometric balances. For example, in geochemistry mafic and felsic associations are known and understood, with mafic materials commonly having dominant elements iron, magnesium and manganese and felsic materials having sodium, potassium, aluminium and silicon. Greenacre, Grunsky and Bacon-Shone (2019) point out the usefulness of SLRs in the practice of compositional data analysis, serving as simple alternatives to the isometric form of "balances". The term "balance" has been associated with these logratios of geometric means, which can have a problematic interpretation, especially when rare parts are included in a geometric mean, as shown by Greenacre et al. (2019). The definition (1) is a more intuitive definition of the balancing of parts, and to avoid confusion will be specifically called a balance of amalgamated parts, or simply an amalgamation balance.

In the book by Pawlowsky-Glahn et al. (2015), the use of amalgamations in logratios is ruled out using a mathematical argument. They specifically state that "amalgamation is incompatible with the techniques presented in this book", preferring to group parts using geometric means which lead to the isometric form of balances. A criticism repeatedly raised by these authors about using amalgamations is that they are not linear in the simplex (see, for example, Egozcue and Pawlowsky-Glahn, 2006, p. 155). However, in terms of geochemistry and mineralogy, amalgamations must be done in the simplex because the stoichiometric formulae are constructed based on crystal structure. The imposition of such a mathematical condition restricts the geochemist from using

E-mail address: michael.greenacre@upf.edu.

https://doi.org/10.1016/j.acags.2019.100017
Received 14 September 2019; Received in revised form 13 December 2019; Accepted 13 December 2019
Available online 19 December 2019

transformations that make perfect substantive sense in practical applications. Greenacre et al. (2019) believe, by contrast, that any transformation that makes substantive sense and respects the basic principle of subcompositional coherence and scale invariance is acceptable for good practice in compositional data analysis. An SLR is scale invariant and is also subcompositionally coherent, with respect to the addition or deletion of parts that are not part of the SLR itself.

This paper makes three contributions to the practice of using amalgamations in compositional data analysis: (a) It demonstrates how amalgamations can successfully define new parts that are treated just like any other parts in defining logratios; (b) A new procedure for clustering parts is defined based on amalgamations; and (c) It is shown how a set of $J-1$ linearly independent logratios of amalgamations (i.e. SLRs), involving all the parts at least once, can be back-transformed to the original J parts, with a special case being a set of $J-1$ linearly independent pairwise logratios of single parts. Section 2 (Material and Methods) describes the data set on which these contributions are illustrated, as well the technical descriptions of the three contributions. Section 3 shows the results and Section 4 concludes with a discussion. Each of the contributions is illustrated with a worked example, showing the computations performed using R functions, some of which are from the easyCODA package (Greenacre, 2018b). The clustering based an amalgamations will become available in a forthcoming version of easyCODA.

2. Material and methods

2.1. Aar Massif data set

The Aar Massif data set is considered, comprising 87 geochemical samples of glacial sediment in Switzerland, with parts the 10 major oxides SiO_2, TiO_2, Al_2O_3, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, Fe_2O_3t. Average percentages are as high as 70.81% (SiO_2) and as low as 0.06% (MnO). These data have been previously analysed by Tolosana-Delgado and von Eynatten (2010), van den Boogaart and Tolosana-Delgado (2013), Martín-Fernández et al. (2018), and Greenacre et al. (2019).

2.2. Use of logratios that involve amalgamations

The effect on the sample geometry using five different ways of transforming the compositional data set are compared, as well as their respective two-dimensional approximations, using principal component analysis (PCA). Some of these involve amalgamated parts. An additional analysis is based on the first two optimal principal balances computed by Martín-Fernández et al. (2018). The various options considered here are:

Option 1: The 10 centered logratios (CLRs), which have a sample geometry equivalent to the full set of 45 LRs (see, for example, Aitchison and Greenacre (2002, Appendix)). The CLRs are the logratios of each one of the 10 parts as numerators with the same denominator equal to the geometric mean of all 10 parts.

Option 2: A set of 9 additive logratios (ALRs), where 9 of the parts are ratioed with respect to the other reference part, in this case Al_2O_3. In a preliminary analysis it was found that using Al_2O_3 as the reference part gave the best agreement with the sample logratio structure, although there is not a large difference if any other part is used as reference. Hence, the ALRs used here are the logratios $\log(SiO_2/Al_2O_3)$, $\log(TiO_2/Al_2O_3)$, $\log(MnO/Al_2O_3)$, etc.

Option 3: A more general set of 9 linearly independent LRs determined by stepwise logratio selection, as described by Greenacre (2018a), i.e. the logarithms of: MgO/Na_2O, K_2O/P_2O_5, SiO_2/K_2O, TiO_2/Na_2O, SiO_2/Na_2O, MgO/CaO, TiO_2/MnO, Al_2O_3/P_2O_5, SiO_2/Fe_2O_3t. The function STEP in the package easyCODA was used to obtain this set of pairwise logratios.

Option 4: A set of logratios that includes three pre-defined amalgamations of parts as additional parts for creating logratios: mafic (=MgO + Fe_2O_3t + MnO), felsic (=Na_2O, SiO_2, Al_2O_3, K_2O) and carbonate-apatite (=CaO + P_2O_5). Using the same stepwise process as in Option 3, implemented by the function STEP in easyCODA, this results in the logarithms of MgO/Na_2O, K_2O/P_2O_5, SiO_2/K_2O, TiO_2/Na_2O, SiO_2/Na_2O, felsic/carbonate-apatite, MnO/carbonate-apatite, Al_2O_3/MgO, TiO_2/Fe_2O_3t (notice that mafic was not chosen in any logratio) – see Greenacre et al. (2019, Table 1). This set of logratios satisfies the condition that each part is mentioned at least once, which is required later for back-transforming the SLRs.

Option 5: The subset of the first five logratios that are common to Options 3 and 4, which together explain 98.7% of the total logratio variance (Greenacre et al., 2019, Table 1).

Options 6–10: The optimal two-dimensional versions of options 1–5 above, as determined by a principal component analysis (PCA) in each case.

Table 1
Matrix of Spearman correlations between the sets of inter-sample distances obtained from the eleven different configurations of the samples, with rows and columns sorted in descending order of the first column (e.g., **dist1** = the distances computed in option 1).

	dist1	dist3	dist4	dist6	dist5	dist8	dist9	dist10	dist2	dist7	dist11
Exact logratio distances											
dist1	1.000										
Distances based on 9 simple LRs											
dist3	0.994	1.000									
Distances based on 9 LRs & SLRs											
dist4	0.993	0.991	1.000								
Distances based on first two PCs of option 1											
dist6	0.987	0.986	0.979	1.000							
Distances based on five best LRs											
dist5	0.983	0.989	0.989	0.970	1.000						
Distances based on first two PCs of option 3											
dist8	0.983	0.990	0.976	0.995	0.976	1.000					
Distances based on first two PCs of option 4											
dist9	0.982	0.982	0.985	0.994	0.975	0.989	1.000				
Distances based on first two PCs of option 5											
dist10	0.970	0.980	0.972	0.979	0.986	0.988	0.983	1.000			
Distances based on ALRs with respect to Al_2O_3											
dist2	0.968	0.966	0.972	0.954	0.980	0.955	0.962	0.969	1.000		
Distances based on first two PCs of option 2											
dist7	0.952	0.956	0.956	0.959	0.967	0.963	0.967	0.979	0.988	1.000	
Distances based on first two principal (isometric) balances											
dist11	0.950	0.950	0.955	0.959	0.953	0.955	0.966	0.960	0.954	0.959	1.000

Option 11: The two-dimensional display using the first two optimal principal balances of Martín-Fernández et al. (2018). The optimal principal balances are a sequence of logratios of geometric means obtained by a recursive partitioning procedure, using an exhaustive search at all the two-way splits of the parts (Martín-Fernández et al., 2018).

To compare all 11 options, the inter-sample distances were computed in their respective full 9-dimensional spaces in the case of options 1–5, and similarly in the two-dimensional displays in the case of options 6–11. This results in a set of ½ × 87 × 86 = 3741 distances for each option. The pairwise similarities between these 11 sample structures were measured using the Spearman correlations between the sets of distances.

2.3. Clustering the parts using amalgamations

When considering any subset of logratios, of whatever type, their ability to reconstruct the complete set of pairwise logratios can be assessed by the percentage of logratio variance explained (Greenacre, 2018a). The same idea can be used to define a clustering of parts based on their amalgamations.

Starting with all J parts, it is obvious that their complete set of log-ratios explains the total logratio variance, by definition. When two parts are amalgamated, then the logratios based on this reduced subset of parts can no longer explain the totality of the logratio variance. But one of the possible ½$J(J-1)$ amalgamations will lower the explained variance the least, which is the pair that is chosen in this first step. In the present example, it will be seen that this first pair is the amalgamation of Si_20 with Fe_2O_3t, leading to a tiny reduction of 0.07% of explained variance. Now there are $J-1$ amalgamations, consisting of $J-2$ single parts and the amalgamation $Si_20 + Fe_2O_3t$ of two parts, and the next amalgamation is sought that leads to a set whose logratios reduce the explained variance the least again: it will turn out that further amalgamating $Si_20 + Fe_2O_3t$ with Al_2O_3 gives a minimum reduction of 0.19%, bringing the variance explained at this stage reduced by 0.26%. Notice that the pairwise log-ratios during this process are all SLRs. There are now $J-2$ amalgam-ations, the amalgamation $Si_20 + Fe_2O_3t + Al_2O_3$ and the remaining $J-3$ single-part amalgamations. And so the iterative process continues, where the agglomerative hierarchical clustering procedure continues and at each step the node height is defined by the cumulative loss of explained variance. In the last step, when all parts have been merged into a single amalgamation, with all values 1 and all logratios 0, none of the logratio variance is explained, i.e. 100% reduction in explained variance.

2.4. The inverse transformation of sets of SLRs and LRs

The method of passing from a set of SLRs back to the original parts is explained by first showing a simple three-part example ($X_1+X_2+X_3 = 1$) and two amalgamation logratios (SLRs):

$$Y_1 = \log\left(\frac{X_1}{X_2 + X_3}\right) \qquad Y_2 = \log\left(\frac{X_2}{X_3}\right)$$

By exponentiating:

$$e^{Y_1} = \frac{X_1}{X_2 + X_3} \qquad e^{Y_2} = \frac{X_2}{X_3}$$

Multiplying out gives two linear equations, and the third equation is the condition that the parts sum to a constant, 1 in this case:

$$\begin{aligned} X_1 - e^{Y_1}X_2 - e^{Y_1}X_3 &= 0 \\ X_2 - e^{Y_2}X_3 &= 0 \\ X_1 + X_2 + X_3 &= 1 \end{aligned}$$

Solving the following system thus gives the original parts X_1, X_2 and X_3:

$$\begin{bmatrix} 1 & -e^{Y_1} & -e^{Y_1} \\ 0 & 1 & -e^{Y_2} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\text{i.e.} \quad \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 & -e^{Y_1} & -e^{Y_1} \\ 0 & 1 & -e^{Y_2} \\ 1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

In this simple case, the solution is easy to find in closed form:

$$X_1 = \frac{e^{Y_1}}{1 + e^{Y_1}} \qquad X_2 = \frac{e^{Y_2}}{(1 + e^{Y_1})(1 + e^{Y_2})} \qquad X_3 = \frac{1}{(1 + e^{Y_1})(1 + e^{Y_2})}$$

In general, for a set of $J-1$ linearly independent SLRs, which should involve all J parts, a $J \times J$ matrix $\mathbf{A}$ should be set up where each of the first $J-1$ rows describes the pattern of the respective SLR, and the last row is a vector of 1s. For the i-th row ($i = 1, \ldots, J-1$) 1s are placed in the columns corresponding to the numerator parts and $-e^{Y_i}$ s are placed in the columns corresponding to the denominator parts, where Y_i is the value of the i-th SLR, otherwise all other elements are 0. The vector $\mathbf{b}$ ($J \times 1$) consists of $J-1$ 0s and the last element 1. The logratios back-transformed to the original parts is the solution for $\mathbf{x}$ of the system of linear equations $\mathbf{Ax} = \mathbf{b}$, i.e. $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

The same strategy applies to a set of $J-1$ pairwise logratios that are independent (e.g., see option 3 of Section 2): each row that corresponds to a logratio would contain a single 1 for the numerator and a single $-e^{Y_i}$ for the denominator in the columns corresponding to the respective parts.

3. Results

3.1. Comparison of 11 options for defining sample structure using logratios

Fig. 1 shows the five PCAs of the analyses of options 6–10, and the scatterplot of the first two "principal balances" (option 11). These all show a high level of similarity in the configurations of the samples, but at this high level there are some differences. Correlating the 11 sets of distances generated by each of these options, using the Spearman correlation, the correlation matrix in Table 1 is obtained. The rows and columns of this table have been reordered in terms of decreasing correlation of the various options with option 1, i.e. with the full-space exact logratio distances. Notice that the logratio distances are the so-called Aitchison distances divided by the square root of the number of parts (square root of 10 in this case) – Greenacre (2018b) prefers the former definition of logratio distance, which does not augment with the number of parts.

Table 1 shows that the two sample configurations generated by the simple LRs and the SLRs (options 3, $r = 0.994$, and 4, $r = 0.993$) are the most similar to the exact logratio distances, followed by the two-dimensional approximation of the logratio distances (option 6, $r = 0.987$) and then the configuration based on just the best five LRs (option 5, $r = 0.983$). Then follow the two-dimensional approximations of the LRs and SLRs (options 8, $r = 0.983$, and 9, $r = 0.982$) and the two-dimensional approximation using the best five LRs (option 10, $r = 0.970$). Finally come the ALRs, first full-dimensional and then two-dimensional (options 2, $r = 0.968$, and 7, $r = 0.952$) and last the scat-terplot of the first two optimal principal balances (option 11, $r = 0.950$).

3.2. Cluster analysis based on amalgamating parts

Fig. 2(a) shows the hierarchical clustering of the parts using the new

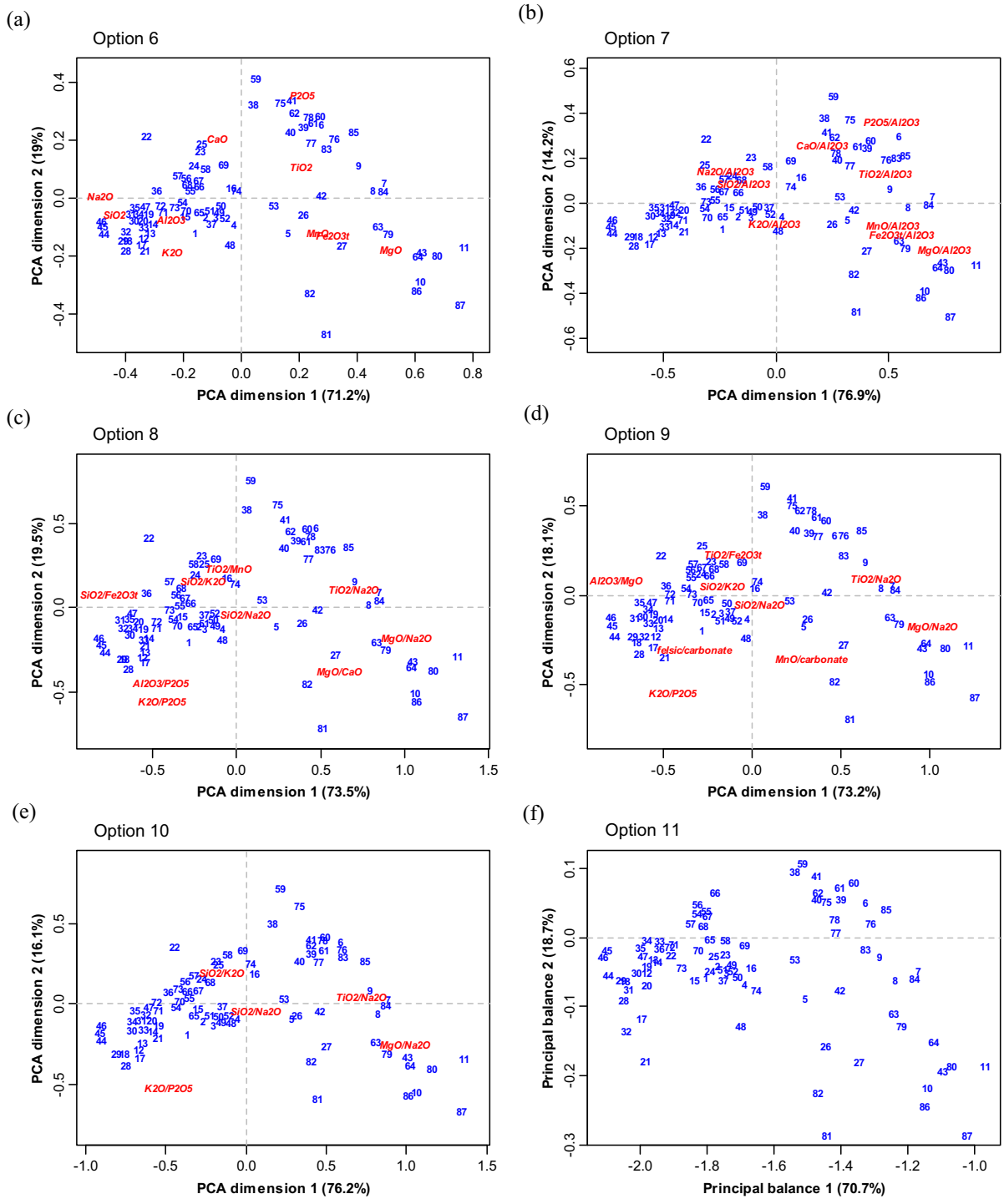


Fig. 1. Two-dimensional configurations of options 6–11. (a) PCA of the CLRs, which is equivalent, as far as the sample configuration is concerned, to the PCA of all the LRs. (b) PCA of the ALRs with respect to Al_2O_3. (c) PCA of the set of nine LRs chosen stepwise. (d) PCA of the 9 LRs that include the amalgamations of felsic and carbonate-apatite, chosen stepwise. (e) PCA of the five best LRs. (f) Scatterplot of the first two optimal "principal balances". Explained variances are indicated in each case.

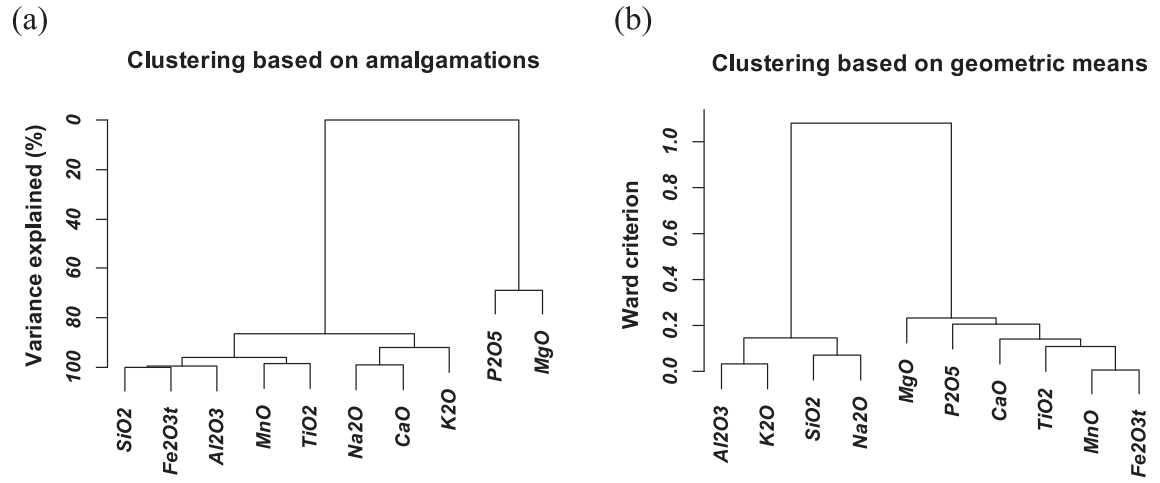


Fig. 2. (a) Hierarchical clustering based on the new algorithm that depends on amalgamating parts, with cumulative percentages of explained logratio variance defining the node heights; (b) Hierarchical Ward clustering, using function clustCoDa_qmode() in R package 'robCompositions', based on geometric means.

clustering algorithm based on amalgamations, where the vertical axis is in cumulative amounts of explained logratio variance by the pairwise logratios at each level of the tree. At the lowest level, with all single parts, their logratios explain all the variance. As parts are amalgamated, their logratios explain less and less variance, and at the top when all parts are amalgamated into one cluster, zero logratio variance is explained. In Fig. 2(a) the scale is annotated in percentages of the total logratio variance (equal to 0.1523), from 100% at the bottom to 0% at the top. Notice that the first three amalgamations can be made, after four steps of the algorithm, namely SiO_2 + Fe_2O_3t + Al_2O_3, MnO + TiO_2, and Na_2O + CaO, with very little loss of explained variance – the loss is only 1.38%.

As a comparison, the result in Fig. 2(b) is obtained using the function clustCoDa_qmode() in the R package 'robCompositions'. This clusters the Aitchison distances between the parts, using Ward clustering, where the Aitchison distances differ from the logratio distances of Greenacre (2011, 2018a, b) by a simple scalar multiplier, as explained earlier. Clusters of parts are represented as geometric means, as suggested by Martín-Fernández et al. (2018), rather than amalgamations.

The explained logratio variances can be studied for any set of contrasts on the respective trees. As an example, when the two trees in Fig. 2 are each cut to give two groups of parts, this implies a single logratio. The percentage of variance explained is 68.7% for the SLR balance in Fig. 2(a) formed by sums of {SiO_2, Fe_2O_3t,Al_2O_3,MnO,TiO_2,Na_2O,CaO,K_2O} in the numerator and of {P_2O_5, MgO} in the denominator, and 69.0% for the "balance" in Fig. 2(b) formed by the geometric means of {Al_2O_3,K_2O,-SiO_2,Na_2O} in the numerator and of {CaO,P_2O_5,MgO,TiO_2,MnO,Fe_2O_3t} in the denominator. As in many cases reported before, by Greenacre (2018a), Greenacre et al. (2019) as well as in Fig. 1, the approach using sums of parts, functions well compared to the approach using geometric means, explaining only slightly less variance and having a straightforward interpretation.

Even though the Ward clustering in Fig. 2(b) is not based on explained variance, as is that of Fig. 2(a), the successive loss of variance explained can be computed at each node. The cumulative percentage losses at each node of the two dendrograms of Fig. 2 are given in Table 2. In the case of Fig. 2(a), the values in the first column of Table 2 are 100 minus the heights of the nodes, whereas for Fig. 2(b) these have been computed externally, using geometric means to combine the parts at each node. The values are quite comparable, even though different sequences of parts are clustered.

Table 2
Cumulative percentage losses in explained logratio variance as parts are amalgamated in Fig. 2(a) or combined using geometric means in Fig. 2(b). The values 31.33% and 30.96% for unexplained variance at Node 8, subtracted from 100%, correspond to the explained variances of 68.7% and 69.0% mentioned earlier, when there are only two subsets of combined parts available as a single logratio balance in the respective cases.

Node	Clustering based on amalgamations	Clustering based on geometric means
1	0.07	0.06
2	0.26	0.19
3	0.64	0.58
4	1.38	1.93
5	4.06	4.66
6	7.93	8.48
7	13.59	13.23
8	31.33	30.96
9	100	100

3.3. *Back-transformation of SLRs to the original parts*

For the Aar Massif data set, using option 4 given previously, the nine selected ratios, in which every part appears at least once, lead to the following matrix defining the system of equations:

$$\begin{bmatrix} 0&0&0&0&1&0&-e^{Y_1}&0&0&0\\ 0&0&0&0&0&0&0&1&-e^{Y_2}&0\\ 1&0&0&0&0&0&0&-e^{Y_3}&0&0\\ 0&1&0&0&0&0&-e^{Y_4}&0&0&0\\ 1&0&0&0&0&0&-e^{Y_5}&0&0&0\\ 1&0&1&0&0&-e^{Y_6}&1&1&-e^{Y_6}&0\\ 0&0&0&1&0&-e^{Y_7}&0&0&-e^{Y_7}&0\\ 0&0&1&0&-e^{Y_8}&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&-e^{Y_9}\\ 1&1&1&1&1&1&1&1&1&1 \end{bmatrix} \begin{bmatrix} \mathbf{SiO_2}\\ \mathbf{TiO_2}\\ \mathbf{Al_2O_3}\\ \mathbf{MnO}\\ \mathbf{MgO}\\ \mathbf{CaO}\\ \mathbf{Na_2O}\\ \mathbf{K_2O}\\ \mathbf{P_2O_5}\\ \mathbf{Fe_2O_3t} \end{bmatrix} = \begin{bmatrix} 0\\0\\0\\0\\0\\0\\0\\0\\0\\1 \end{bmatrix} \quad (2)$$

The function **invSLR** has been written to set up these equations and perform the inversion, given the SLRs and their definitions. The function has been added to the latest version 0.32 of the **easyCODA** package (Greenacre, 2018b) in R (R core team, 2019). The latest version of this package can always be installed in R from R-Forge as follows.

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

The code to compute the inverse using Eqn (2) is illustrated in the

Appendix, where the ratios are coded in the following format: **"num/den"**, where **num** is the numerator part or amalgamation and **den** the denominator part or amalgamation. An amalgamation is coded as **"p1&p2&..."** where **p1** is part 1, **p2** is part 2, etc. Thus, for example, the 7th SLR in (2), which contrasts MnO against carbonate-apatite (which is an amalgamation of CaO and P2O5), is coded as **"MnO/CaO&P2O5"**. The Appendix gives the R code to perform the stepwise selection of the logratios, including those that involve amalgamations, and then back-transforming the resultant SLRs to the original parts.

A special case is when there are no amalgamations but a complete set of independent pairwise logratios (LRs) that explain all the logratio variance. Such a set of LRs defines an acyclic connected graph of the parts (Greenacre, 2018a, b). The optimal set, given previously as option 3, leads to the following matrix defining the system of equations, where there is only a single 1 and a single $-e^{Y_i}$ in each row, indicating the numerator and denominator parts respectively:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -e^{Y_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -e^{Y_2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -e^{Y_3} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -e^{Y_4} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -e^{Y_5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -e^{Y_6} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -e^{Y_7} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -e^{Y_8} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -e^{Y_9} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{SiO_2} \\ \mathbf{TiO_2} \\ \mathbf{Al_2O_3} \\ \mathbf{MnO} \\ \mathbf{MgO} \\ \mathbf{CaO} \\ \mathbf{Na_2O} \\ \mathbf{K_2O} \\ \mathbf{P_2O_5} \\ \mathbf{Fe_2O_3t} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (3)$$

In this case, without the need to define the amalgamations, the code is simpler – see the R script in the Appendix.

5. Discussion and conclusion

From the results given above, it can be concluded – at least, for this data set – that SLRs are valid transformations of compositional data. In this particular example, amongst all alternatives considered, they approximated the logratio data structure the best, both when only LRs were used and also when amalgamations were included, whereas the two optimal principal balances fared the worst, worse even than the set of ALRs. The best five logratios, which involve six out of the ten parts, also fare better than the principal balances, which involve all ten parts. Admittedly, this ranking amongst the 11 options is in terms of relatively small differences between the correlations, where all the options can be considered to give good representations of the structure. Nevertheless, this example goes to show that SLRs should never be ruled out from possible consideration, coupled with their simple interpretation and the fact their definition is usually based on substantive geochemical grounds. If certain SLRs are preferred based on domain knowledge, these can be forced into the stepwise search, with a small sacrifice in explained variance. For example, it might be preferable in the present example to ensure that the ratio mafic/felsic be specified as an initial logratio, based on the geochemistry of igneous and metamorphic rocks, and then let the algorithm reveal other logratios – Greenacre et al. (2019, Fig. 6) show exactly this possibility, where the effect is to reduce the explained variance by only a tiny 0.3 percentage points.

Although the balances using geometric means have attractive theoretical properties, imposing the strict use of these balances on the practitioner presents difficulties in interpretation, as detailed by Greenacre et al. (2019). Balances based on summations of parts can rather be used as the basis of an alternative and more intuitive clustering of the parts, which has a straightforward interpretation now that clustering means amalgamating, and which enables the creating and testing of geochemical models. Moreover, the logratios of amalgamations (SLRs) are more correctly called balances: for example, when an SLR is zero, then this means that the sums of the parts in the numerator and denominator are the same. This is not true for a balance of geometric means, although many authors misinterpret this balance as a balancing of the sums – see, for example, Pawlowsky-Glahn et al. (2015, page 41), Calle (2019, page 7).

Finally, like the different logratio transformation options presently available in compositional data analysis (ALR, CLR and ILR), it has been shown that a linearly independent set of SLRs, one less than the number of parts, and including all parts, can be back-transformed to the original part values by solving a set of linear equations. This includes the special case of a set of linearly independent LRs that includes all the parts. Thus, both SLRs and simple pairwise LRs satisfy all the requirements for being useful transformations in the practice of compositional data analysis.

Author contributions section

Michael Greenacre is solely responsible for the theoretical part of the article, as well as the applications, the R script and the easyCODA package in R.

Conflict of interest

Michael Greenacre declares that he has no conflict of interest.

Acknowledgments

The co-operation of Raimon Tolosana-Delgado is gratefully acknowledged for supplying the Aar Massif data set and for recently making this data set available in the R package 'compositions' (van den Boogaart et al. 2019). The encouraging comments and constructive criticisms of two referees significantly improved this paper.

Appendix

R script for some of the computations (notice that the implementation of the clustering based on amalgamations is not yet implemented in the easyCODA package, but will appear soon in a new version of the package)

```
### Aar Massif data set is in data object 'Aar'
### in the package 'compositions'
library(compositions)
data(Aar)
### Columns 3 to 12 are the oxide data
aar <- Aar[,3:12]
### Close the rows
aar <- aar / rowSums(aar)
round(head(aar),5)
#      SiO2    TiO2   Al2O3     MnO     MgO     CaO    Na2O     K2O    P2O5  Fe2O3t
# 1 0.73638 0.00221 0.14236 0.00037 0.00482 0.01043 0.04344 0.04194 0.00060 0.01746
# 2 0.73958 0.00260 0.14019 0.00038 0.00511 0.01142 0.04326 0.03955 0.00070 0.01722
# 3 0.74027 0.00301 0.13883 0.00041 0.00551 0.01153 0.04320 0.03829 0.00070 0.01824
# 4 0.78219 0.00320 0.11374 0.00043 0.00571 0.01021 0.03414 0.03164 0.00060 0.01812
# 5 0.78943 0.00391 0.10313 0.00061 0.00862 0.01072 0.02816 0.02926 0.00070 0.02546
# 6 0.74817 0.00543 0.11942 0.00073 0.00945 0.02232 0.02965 0.03217 0.00292 0.02975

### Compute the three amalgamations
mafic      <- rowSums(aar[,c(5,10,4)])
felsic     <- rowSums(aar[,c(7,1,3,8)])
carb_apat  <- rowSums(aar[,c(6,9)])

### Add the three amalgamations to the existing parts
aar.amalg  <- cbind(aar, mafic, felsic, carb_apat)

### Perform the stepwise logratio selection using function
### STEP in the easyCODA package: logratios are formed from the
### columns in aar.amalg, with the target aar to be explained
### (see Option 4 in Section 2)
require(easyCODA)
aar.step   <- STEP(aar.amalg, aar, weight=FALSE)

### Names of the logratios in $names of the STEP object
aar.step$names
# [1] "MgO/Na2O"  "K2O/P2O5"          "SiO2/K2O"        "TiO2/Na2O"
# [5] "SiO2/Na2O" "felsic/carb_apat"  "MnO/carb_apat"   "Al2O3/MgO"
# [9] "TiO2/Fe2O3t"
ratio.names <- aar.step$names

### The ones involving amalgamations have to be renamed
### with the definitions of the amalgamations, as required by invSLR
ratio.names[6] <- "Na2O&SiO2&Al2O3&K2O/CaO&P2O5"
ratio.names[7] <- "MnO/CaO&P2O5"

### The names of the parts (for matching with logratio names)
part.names <- colnames(aar)

### Transform the logratios back to the original parts using invSLR
### The logratios are in $logratios of the STEP object
SLRinverses <- invSLR(aar.step$logratios, part.names, ratio.names)
round(head(SLRinverses), 5)
```

```
#         SiO2    TiO2   Al2O3     MnO     MgO     CaO    Na2O     K2O    P2O5  Fe2O3t
# [1,] 0.73638 0.00221 0.14236 0.00037 0.00482 0.01043 0.04344 0.04194 0.00060 0.01746
# [2,] 0.73958 0.00260 0.14019 0.00038 0.00511 0.01142 0.04326 0.03955 0.00070 0.01722
# [3,] 0.74027 0.00301 0.13883 0.00041 0.00551 0.01153 0.04320 0.03829 0.00070 0.01824
# [4,] 0.78219 0.00320 0.11374 0.00043 0.00571 0.01021 0.03414 0.03164 0.00060 0.01812
# [5,] 0.78943 0.00391 0.10313 0.00061 0.00862 0.01072 0.02816 0.02926 0.00070 0.02546
# [6,] 0.74817 0.00543 0.11942 0.00073 0.00945 0.02232 0.02965 0.03217 0.00292 0.02975
### The values check with the part values shown at the start

### The stepwise search for the best pairwise logratios
### (see Option 3 in Section 2)
require(easyCODA)
aar.step2 <- STEP(aar, weight=FALSE)
ratio.names2 <- aar.step2$names
part.names2  <- colnames(aar)
### The inverse transformation
SLR.inverses2 <- invSLR(aar.step2$logratios, part.names2,
                        ratio.names2)
round(head(SLR.inverses2), 5)
```

```
#         SiO2    TiO2   Al2O3     MnO     MgO     CaO    Na2O     K2O    P2O5  Fe2O3t
# [1,] 0.73638 0.00221 0.14236 0.00037 0.00482 0.01043 0.04344 0.04194 0.00060 0.01746
# [2,] 0.73958 0.00260 0.14019 0.00038 0.00511 0.01142 0.04326 0.03955 0.00070 0.01722
# [3,] 0.74027 0.00301 0.13883 0.00041 0.00551 0.01153 0.04320 0.03829 0.00070 0.01824
# [4,] 0.78219 0.00320 0.11374 0.00043 0.00571 0.01021 0.03414 0.03164 0.00060 0.01812
# [5,] 0.78943 0.00391 0.10313 0.00061 0.00862 0.01072 0.02816 0.02926 0.00070 0.02546
# [6,] 0.74817 0.00543 0.11942 0.00073 0.00945 0.02232 0.02965 0.03217 0.00292 0.02975
```

References

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

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Appl. Stat. 51, 375–392.

Calle, M.L., 2019. Statistical analysis of metagenomics data. Genom. Inform. 17 (1), e6. https://doi.org/10.5808/GI.2019.17.1.e6.

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

Greenacre, M., 2018a. Variable selection in compositional data analysis, using pairwise logratios. Math. Geosci. 51, 649–682. https://doi.org/10.1007/s11004-018-9754-x.

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

Greenacre, M., Grunsky, E., Bacon-Shone, J., 2019. A comparison of amalgamation logratio balances and isometric logratio balances in compositional data analysis. In revision at Computers & Geosciences.

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

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

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

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

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

van den Boogaart, K.G., Tolosana-Delgado, R., Bren, M., 2019. Compositions: Compositional Data Analysis. R Package Version 1, pp. 40–43. https://CRAN.R-project.org/package=compositions.