Economics Working Paper Series

Working Paper No. 1551

Selection and statistical analysis of compositional ratios

Michael Greenacre

August 2016

Selection and statistical analysis of compositional ratios

submitted 17 August 2016

Michael Greenacre
Department of Economics and Business
Universitat Pompeu Fabra,
Ramon Trias Fargas 25-27
08005 Barcelona
Spain

Email: michael.greenacre@upf.edu

Abstract: Compositional data are nonnegative data with the property of closure: that is, each set of values on their components, or so-called parts, has a fixed sum, usually 1 or 100%. Compositional data cannot be analyzed by conventional statistical methods, since the value of any part depends on the choice of the other parts of the composition of interest. For example, reporting the mean and standard deviation of a specific part makes no sense, neither does the correlation between two parts. I propose that a small set of ratios of parts can be determined, either by expert choice or by automatic selection, which effectively replaces the compositional data set. This set can be determined to explain 100% of the variance in the compositional data, or as close to 100% as required. These part ratios can then be validly summarized and analyzed by conventional univariate methods, as well as multivariate methods, where the ratios are preferably log-transformed.

Keywords: compositional data, logarithmic transformation, log-ratio analysis, multivariate analysis, ratios, univariate statistics.

1. Introduction

Compositional data (Aitchison 1986) are sets of non-negative data that have been expressed relative to a fixed total (usually as proportions summing to 1 or percentages summing to 100%). The original totals are not of interest − rather, the relative values, or *composition*, are relevant for summarizing and statistical analysis. The components of a composition are called its *parts*. If a subset of the parts are considered and the data are re-expressed with respect to the new subtotals, this is called a *subcomposition*. In several situations, where the total of the original data is the same for all samples, considering a subcomposition is usually not an issue. For example, in the case of time budget data where activities such as sleeping, eating, leisure, work, transport, etc., are recorded during a 24-hour day, there is usually no point in dropping an activity and re-expressing the parts relative to the total without that activity. Similarly, concentrations in parts per million (ppm), for example, are analysed as such, without re-expression as proportions. In this technical note I concentrate on compositional data where subcompositions or extended compositions are possible, for example, geochemical data, or fatty acid data in ecology, in other words where the proportions depend on the particular choice of parts made by the researcher.

The act of converting a set of values into its set of relative values by dividing by the total, is called *closure*. The term *normalization* is also used; for example, it is said that "the data are normalized", "the data are closed", "some parts are excluded and the data are renormalized, or reclosed", etc… It is exactly because the compositional values associated with the parts change, after renormalization, that makes compositional data unique, and needing special approaches.

In spite of the compositional values depending on the particular mix of parts chosen by the user, parts of a composition are still often summarized by statistics such as the mean and correlation coefficient. Clearly, these summary statistics make no sense when comparing different studies, unless studies have chosen exactly the same set of parts. In multivariate analysis of compositional data, it has long been recognized that a valid approach to compositional data is to analyse ratios of parts, which are invariant to the choice of the set of parts. Basing the analysis on ratios is an approach with the property of *subcompositional coherence* (see, for example, Greenacre and Lewi, 2005). Log-ratio analysis (Aitchison 1990, Aitchison and Greenacre 2002) is a variant of principal component analysis that displays the reduced dimensional structure of all log-ratios of the parts.

Given the central role of ratios in the subcompositionally coherent approach to compositional data analysis, abbreviated as CoDA, it seems obvious that when it comes to reporting univariate statistics, these should be on ratios of parts rather than the parts themselves. In certain research areas, for example in fatty acid analyses in studies of the marine food web, some ratios are actually proposed as indicators of certain phenomena – see, for example, Kraft et al. (2015). In my opinion, reporting ratios should be the norm rather than the exception, since these are the only quantities that are comparable across studies. The problem is that if there are p parts, then there are $\frac{1}{2}p(p-1)$ ratios to consider. But we do know that the dimensionality of a compositional data set with p parts is $p-1$, and thus only $p-1$ ratios are required to reproduce the variance of the whole data set. In this paper I will show how a relatively small set of ratios can be chosen and evaluated for their ability to replace the original compositional data. The advantage will be that these can be summarized and analysed by regular statistical methods.

Section 2 defines the total variance in a compositional data set, which is important because the chosen ratios will be evaluated according to how well they explain this variance. In Section 3 I will discuss how a "good" set of ratios can be chosen, and how they can be summarized. In Section 4 an application will be described, for an archaeological data set by Baxter, Cool and Heyworth (1990) on the oxide compositions of a set of Roman glass cups. Section 5 concludes with a discussion and conclusion.

2. Total variance of a compositional data set

The total variance in a compositional data set, following Aitchison's approach, is measured by the total log-ratio variance. Suppose that the data are in a samples-by-parts matrix **X** ($n\times p$), where the rows of **X** sum to a constant, which can be set to 1 without loss of generality (hence the data are proportions). Then the (unweighted) log-ratio variance, defined by Aitchison (1983), is

$$\sum\sum_{i<i'}\sum\sum_{j<j'}\left(\log\frac{x_{ij}}{x_{ij'}}-\log\frac{x_{i'j}}{x_{i'j'}}\right)^2=\sum\sum_{i<i'}\sum\sum_{j<j'}\left(\log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j'}}{x_{i'j}}\right)^2 \qquad (1)$$

where, for example, $\sum\sum_{j<j'}$ indicates the double summation on all unique pairs of the index, i.e. ½$p(p-1)$ pairs of parts in this case. The second version on the right hand side of (1) shows that the log-ratio variance is the sum of squares of all the logarithmically transformed odds-ratios based on all unique pairs of rows and columns of the data matrix. Greenacre and Lewi (2009) proposed a weighted version where the rows and columns of the table are weighted proportionally to their marginal totals. These weights sum to 1 in each case, so the row weights are $r_i = 1/n$, constant across samples, and the column weights are $c_j = j^{\text{th}}$ part mean. This is the same weighting used by "spectral mapping" (Lewi 1976, 1980; Wouters et al. 2003). The (weighted) log-ratio variance incorporates the weights as follows, using the second version of (1) in terms of odds-ratios

$$\sum\sum_{i<i'}r_i r_{i'}\sum\sum_{j<j'}c_j c_{j'}\left(\log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j'}}{x_{i'j}}\right)^2=\frac{1}{n^2}\sum\sum_{i<i'}\sum\sum_{j<j'}c_j c_{j'}\left(\log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j'}}{x_{i'j}}\right)^2 \qquad (2)$$

The advantage of the weighted version over the unweighted one has been shown by Greenacre and Lewi (2009) and Greenacre (2015), hence I maintain the weighted version in Eq. (2) as the definition of the log-ratio variance, and qualify it with the adjective "unweighted" when referring to Aitchison's original definition in Eq. (1). The log-ratio variance can be shown to be identical to the weighted average of the variances of the *p* columns of the matrix **Y** of so-called *centred log-ratios*

$$\sum_{j=1}^{p} c_j \frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 \text{, where } y_{ij} = \log(x_{ij}) - \sum_{j=1}^{p} c_j \log(x_{ij}) \text{ and } \bar{y}_j = \sum_{i=1}^{n} r_i y_{ij} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \qquad (3)$$

Notice that the rows of the log-transformed data matrix **X** are first centred by their respective weighted average row means to obtain the matrix **Y**, which is then centred by the arithmetic column means. Hence $y_{ij} - \bar{y}_j$ is a double-centring of the matrix log(**X**), with elements log(x_{ij}), using the column and row weights c_j and 1/*n* respectively, then squared and summed using the column and row weights again. In matrix notation, where the column and row weights are gathered in vectors **c** and **r** respectively, the centred log-ratio matrix is

$$\mathbf{Y} = \log(\mathbf{X})(\mathbf{I} - \mathbf{1c}^{\mathsf{T}})^{\mathsf{T}} \qquad (4)$$

and the double-centred matrix is

$$\mathbf{Z} = \mathbf{Y} - \mathbf{1r}^{\mathsf{T}}\mathbf{Y} = (\mathbf{I} - \mathbf{1r}^{\mathsf{T}})\log(\mathbf{X})(\mathbf{I} - \mathbf{1c}^{\mathsf{T}})^{\mathsf{T}} \qquad (5)$$

Then the total log-ratio variance in (2) and (3) is the weighted sum of squares of **Z**

$$\text{variance} = \text{trace}(\mathbf{D}_r\mathbf{Z}\mathbf{D}_c\mathbf{Z}^{\mathsf{T}}). \qquad (6)$$

where $\mathbf{D}_r$ and $\mathbf{D}_c$ are the respective diagonal matrices of the weights.

3. Choosing a set of ratios for univariate analysis

The most obvious way of choosing a set of ratios is that experts do it based on their knowledge of the compositional parts. Whatever the choice is, the relationship of the ratios to the original data set can be measured, as I will explain below. As an alternative, the ratios can be chosen automatically based on statistical criteria, selected to represent the original data set in an optimal way. A further option is a combination of expert and automatic selection, illustrated in the application of Section 4.

Given a proposed set of ratios, the proportion of total variance in Eq. (6), equivalently in Eqs. (2) and (3), that is explained by the set of log-ratios can be calculated. This is conveniently evaluated using the `rda` function for redundancy analysis in R (R core team 2015) in the **vegan** package (Oksanen et al. 2015). Redundancy analysis (Rao 1964, van den Wollenberg 1977) is a generalization of multiple regression analysis, where a set of interval-level response variables is modelled as a linear function of observed predictors. Alternatively in the **vegan** package, the `adonis` function can be used, which has the flexibility of being able to take a rectangular matrix or square symmetric distance matrix as the response matrix. In the present application, the response matrix should be the double-centred matrix **Z** scaled by the square roots of the row and column weights, so that the sum of squares of the response matrix is the total variance: i.e., the response matrix is $\mathbf{D}_r^{1/2}\mathbf{Z}\mathbf{D}_c^{1/2}$ (cf. Eq. (6)).

An automatic way of identifying a "good" set of ratios can proceed in a stepwise fashion, trying in the first step every individual log-ratio as an explanatory variable in explaining the variance in Eq. (6), and selecting the one with the highest percentage of variance explained. This ratio is then fixed as the first log-ratio and then the second best log-ratio in combination with the first is sought, then fixed, and so on. Care must be taken to choose ratios that are "independent" of the ones already chosen: for example, if A/B and B/C have already been selected, then A/C is no longer a candidate for selection, since it depends on the others, or on the log-scale, log(A) –log(C) is linearly dependent on log(A) –log(B) and log(B) –log(C). Since the dimensionality of a p-part compositional data set is $p - 1$, and if all the parts have appeared in at least one log-ratio after $p - 1$ steps of the above procedure, the variance explained will be 100%, as is the case in the application of the following section.

4. Application: oxides in ancient Roman glass cups

To illustrate the procedure, the data set of Baxter et al. (1990) on the percentages by weight of 11 different oxides in a sample of 47 Roman glass cups found in archaeological sites in Colchester, where oxygen is combined with silicon (Si), aluminium (Al), iron (Fe), magnesium (Mg), calcium (Ca), sodium (Na), potassium (K), titanium (Ti), phosphorus (P), manganese (Mn) and antimony (Sb). The data are reproduced in Table 2 of Greenacre and Lewi (2009), who also highlight the difference between unweighted and weighted log-ratio analysis of these data. The total log-ratio variance of this data set is 0.002339, a quite low value in the range of log-ratio variances, due to the high similarity between the compositions of the cups.

The process of selecting the log-ratios starts with looking for the one that explains the most of this variance – I used the `adonis` function in `R` mentioned above, although the same results are obtained using the `rda` function. Of the ½×11×10 = 55 possible ratios, the log-ratio of Si/Ca turned out to be the best, explaining 61.5% of the variance. The second best is Si/Sb, explaining an additional 12.6%, bringing the variance explained up to 74.1%, and so on. The sequence of ratios and their accumulated variances are given in Table 1. In addition, Table 1 reports the medians of these ratios, as well as their reference ranges based on the estimated 0.025 and 0.975 quantiles (i.e., 2.5% and 97.5% percentiles). These statistics are perfectly comparable with other comparable archaeological studies, whether the list of oxides is extended or not, since the ratios are invariant to the parts chosen by the researcher.

What is not clear from the stepwise selection is that at some steps there are more than one ratio competing for entry, giving the same additional benefit of variance explained. For example, in Table 1, the third ratio chosen was Na/Sb explaining an additional 12.3% of the variance (increasing from 74.1% to 86.4%), but exactly the same increase would have been obtained if

Si/Na or Ca/Na had entered. The important aspect of this third step is the entry of Na, which can be in a ratio with either Si, Ca or Sb. in my algorithm the ratio was chosen randomly from the three possibilities and turned out to be Na/Sb. It is at this point that an expert could intervene to choose one of the "competing" ratios that has some relevant substantive meaning and interpretation in the context of the data.

Figure 1 sheds light on the choice of the ratios. This is the (weighted) log-ratio biplot (see Greenacre and Lewi 2005, Greenacre 2009, 2010, 2011), where the contribution biplot scaling (Greenacre 2013) is used, where the parts most contributing to the variance are shown more distant from the origin. Clearly the Si vs. Ca opposition is the most important along the first axis, hence the choice of the first ratio as Si/Ca. It is no surprise either that Si/Sb is then chosen, to include Sb which is the most important contributor on the second axis.

Figure 2 shows the principal component analysis (PCA) biplot based on the 10 log-ratios of Table 1. This analysis requires the ratios to be weighted proportional to the products of the weights of the pair of parts, as in (2). The resemblance with Fig. 1 is clear, and there is a large increase in variance explained (remember that both Figure 1, explaining 79.7% of the variance, and Fig. 2, explaining 88.5%, are explaining the same total log-ratio variance).

Furthermore, the three most important ratios, Si/Ca, Si/Sb and Na/Sb are clearly dominating the two-dimensional solution, so Fig. 3 is the same analysis using just these three ratios, and hardly differs from Fig. 2. Notice that the high variance explained of 99.9% in Fig. 3 is relative to the total variance of just these three ratios.

Based on expert knowledge, a selection of ratios can be made that have a substantive interpretation, or a combination of expert knowledge and automatic selection can be made. For example, Tanimoto and Rehren (2008) consider the composition of glasses from the late bronze age and point out some elements that are "rather heterogeneous in their composition, particularly in their ratios of soda (Na_2O) to potash (K_2O) and lime (CaO) to magnesium (MgO)". These ratios can be "forced" into the first two steps of the present algorithm, after which the same stepwise procedure can be performed. In the present data set it turns out that those two ratios explain only 16.6% of the variance. The automatic selection that follows immediately brings in Si/Ca (or equivalently Si/Na), which increases the variance explained dramatically to 74.1% . A sequence of ratios then follows, bringing in a similar sequence of elements as in Table 1, and reaching 100% with 10 ratios, as before.

5. Discussion and conclusion

The main point of this article is to show that a simple choice of ratios can account for all the variance in a compositional data set, and can be used for univariate or multivariate analysis as a substitute for the original data. Univariate analysis of the ratios is particularly relevant since these are subcompositionally coherent and comparable across studies, whereas univariate statistics of the original parts are not. The definitive book on CoDa, edited by Pawlowsky-Glahn and Buccianti (2011), contains almost no mention of univariate analysis of compositional data, except a passing reference by Lovell et al. (2011) to a paper by Filtzmoser, Hron and Reimann (2009), who use the *isometric log-ratio* transformation to arrive at a set of $p-1$ variables that replace the original data set. These new variables are defined as proportional to ratios of parts to geometric means of parts as follows (the constant of proportionality is not relevant here)

$$\log\left(\frac{x_{ij}}{(\prod_{k=j+1}^{p} x_{ij})^{1/(p-j)}}\right) = \log(x_{ij}) - \frac{1}{p-j}\sum_{k=j+1}^{p}\log(x_{ij}) \quad \text{for } j = 1,\ldots, p-1 \tag{7}$$

The problem with these log-ratios is that they have no easy interpretative meaning and also depend on a trivial property of the parts, namely their ordering in the data set. However, the parts could be re-ordered so that the first ratio is the highest variance-explaining one and so on, in a stepwise manner again.

Another approach is to use so-called *balances* (Egozcue and Pawlowsky-Glahn 2011), which are log-ratios of geometric means of groups of parts. If these are defined as a sequential binary partition, involving $p-1$ balances, then there is an ordering in terms of highest-to-lowest variance explanation in the sequence of balances. This is reminiscent of fatty acid compositional studies where researchers might express the sum of saturated fatty acids, for example, in a ratio with the sum of the unsaturated ones. This is not the ratio of two geometric means, but the ratio of two sums, and the ratio of two sums of subsets of parts is not subcompositionally coherent, but it is the same general idea as a balance. Basing balances on expert knowledge could once more increase the interpretative value of these new log-ratios, combined with other ratios to increase the log-ratio variance explained.

A specifically chosen set of simple part ratios, as I propose here, serves the same purpose as the above approaches, and has an easier interpretation, especially if guided by experts who are familiar with the data context. These also provide simple univariate statistics that can be validly summarized by regular statistical measures of centrality such as the mean and median, and dispersion measures such as standard deviation and quantiles. These ratios can even be correlated or combined in multivariate analyses such as regression and principal component analysis, with the assurance that they are subcompositionally coherent. In a particular field, for example the archaeology of ancient glass where the set of parts is fairly similar across studies, one can imagine a set of ratios becoming a benchmark for easier comparison of data sets.

References

Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70:57−65

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

Aitchison J (1990) Relative variation diagrams for describing patterns of compositional variability. Math Geol 22(4):487–511

Aitchison J, Greenacre MJ (2002) Biplots for compositional data. J R Stat Soc Ser C (Appl Stat) 51(4):375–392

Baxter MJ, Cool HEM, Heyworth MP (1990) Principal component and correspondence analysis of compositional data: some similarities. J Appl Stat 17:229–235

Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment 407: 6100−6108.

Greenacre MJ (2009) Log-ratio analysis is a limiting case of correspondence analysis. Math Geosci 42: 129–134

Greenacre MJ (2010) Biplots in Practice. BBVA Foundation, Bilbao. Free download from www.multivariatestatistics.org

Greenacre MJ (2011) Compositional data and correspondence analysis. In: Pawlowski-Glahn V, Buccianti A (eds) Compositional Data Analysis. Wiley, Chichester UK, pp.104−113

Greenacre,MJ (2013) Contribution biplots. J Comp Graph Stat 22: 107–122

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

Kraft A, Graeve M, Janssen D, Greenacre MJ, Falk-Petersen S (2015) Arctic pelagic amphipods: lipid dynamics and life strategy. J Plank Res, XXXX.

Lewi PJ (1976) Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim Forsch (Drug Res) 26:1295–1300

Lewi PJ (1980) Multivariate data analysis in APL. In: van der Linden GA (ed) Proceedings of APL-80 conference. North-Holland, Amsterdam, pp 267–271

Lovell D, Müller W, Taylor J, Zwart A, Helliwell C (2011) Proportions, percentges, ppm: do the molecular biosciences treat compositional data right? In: Pawlowski-Glahn V, Buccianti A (eds) Compositional Data Analysis. Wiley, Chichester UK, pp.193–207

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

Pawlowski-Glahn V, Buccianti A (eds) Compositional Data Analysis. Wiley, Chichester UK

R core team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rao CR (1964) The use and interpretation of principal component analysis in applied research. Sankhya A 26: 329–358

Tanimoto S, Rehren T (2008) Interactions between silicate and salt melts in LBA glassmaking. J Archaeol Sci 35: 2566–2573

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

Wouters L, Göhlmann HW, Bijnens L, Kass SU, Molenberghs G, Lewi PJ (2003) Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics 59: 1131–1139.

Table caption

Table 1: List of 10 ratios, selected sequentially, which on a log-scale add the most variance explained stepwise to the total log-ratio variance of the 47×11 compositional data set of Baxter et al. (1990). Median ratios and 95% reference range, based on 0.025 and 0.975 quantiles, are also given.

Table 1

Ratio	Cumulative Explained Variance	Median	95% Reference Range
Si/Ca	61.5%	13.3	10.1-15.0
Si/Sb	74.1%	206.5	120.4-403.5
Na/Sb	86.4%	53.3	32.1-93.6
Fe/Sb	93.6%	0.871	0.437-1.422
Ca/K	96.6%	11.4	9.3-14.7
Mg/Na	98.4%	0.0255	0.0179-0.0310
Al/Ca	99.2%	0.347	0.291-0.389
Si/Ti	99.5%	1043	726-1485
Ti/Mn	99.8%	6.00	3.08-8.00
Al/P	100.0%	38.4	25.9-48.1

Figure captions

Figure 1: Log-ratio contribution biplot of the glass cups data set.

Figure 2: PCA biplot of the log-ratios of Table 1, using the contribution biplot scaling for the ratios.

Figure 3: PCA contribution biplot of the three top log-ratios of Table 1 (also the most outlying in the PCA contribution biplot of Fig. 2).

Figure 1

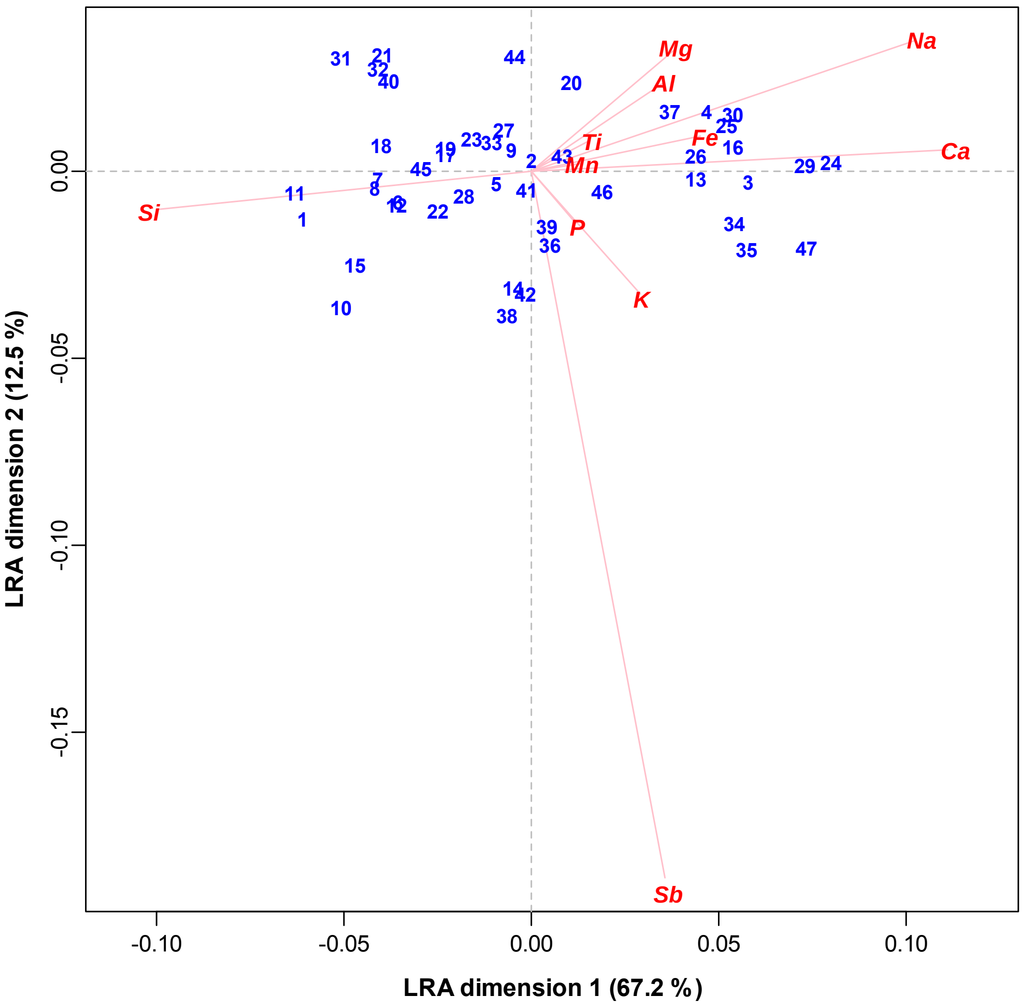
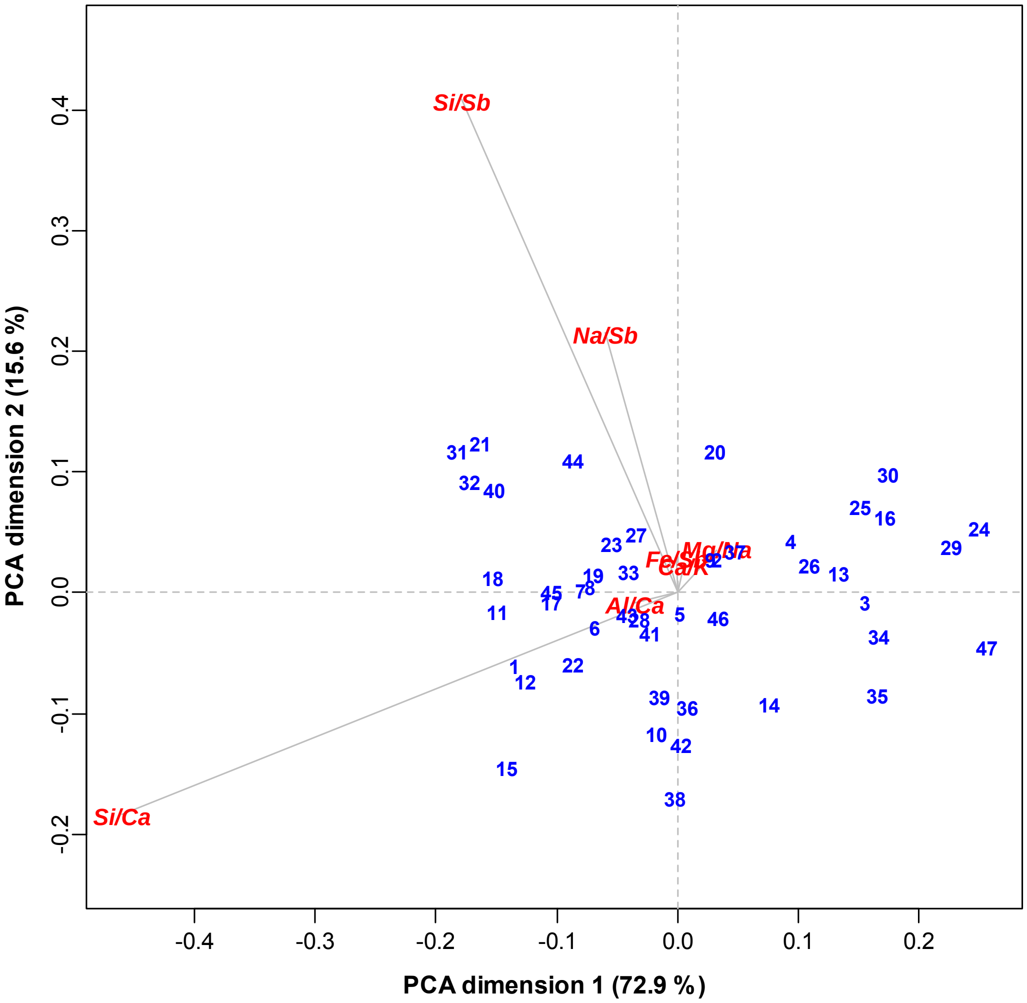
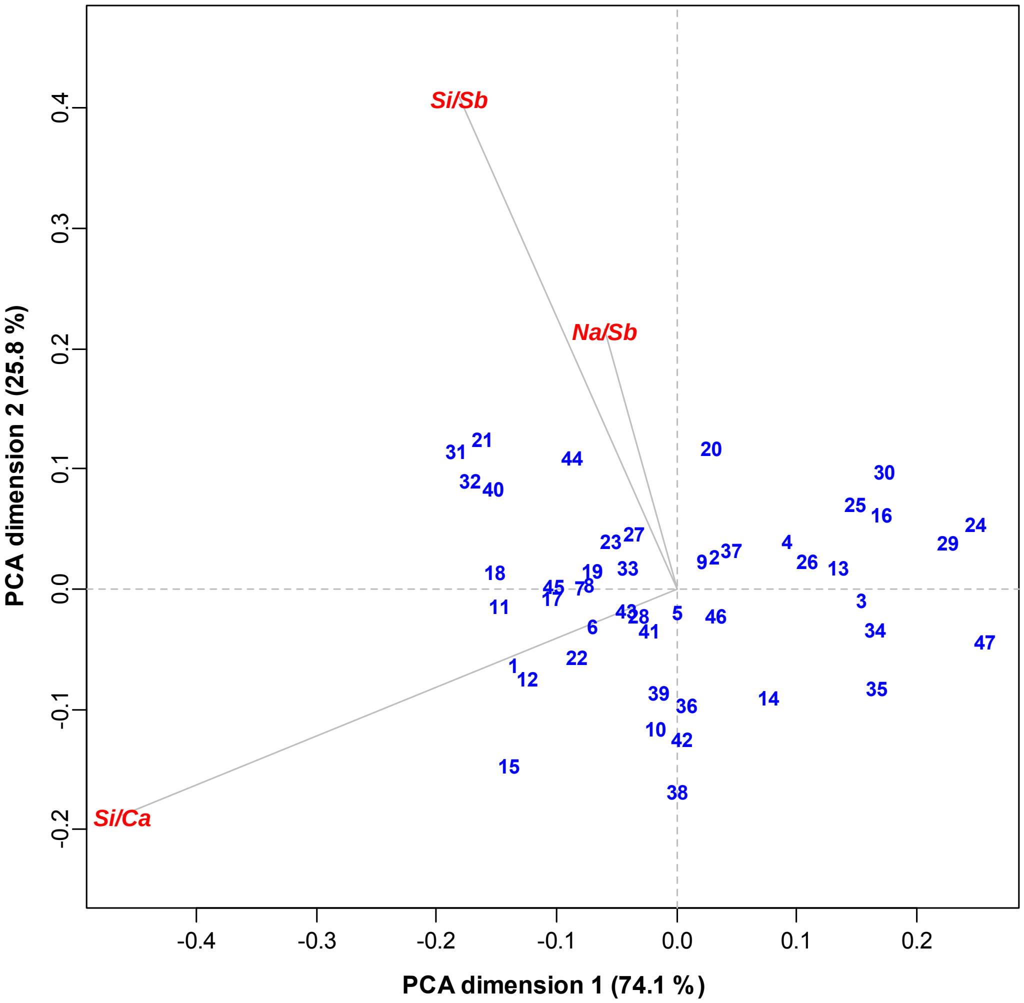

Mg
Na
Al
Ti
Fe
Mn
Ca
Si
P
K
Sb
31
21
32
40
44
20
37
4
30
25
16
18
19
17
23
33
27
9
2
43
26
29
24
45
7
8
11
1
6
12
22
28
5
41
46
13
3
39
36
34
35
47
15
10
14
42
38
0.00
-0.05
-0.10
-0.15
-0.10
-0.05
0.00
0.05
0.10
LRA dimension 1 (67.2 %)
LRA dimension 2 (12.5 %)

Figure 2



Si/Sb
Na/Sb
Fe/Si
Mg/Na
Ca/K
Al/Ca
Si/Ca
31
21
44
20
32
40
30
25
16
24
29
4
26
13
23
27
9
2
37
18
19
33
45
17
7
8
3
11
43
28
5
46
6
41
34
47
1
12
22
39
36
14
35
10
42
15
38
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
PCA dimension 1 (72.9 %)
PCA dimension 2 (15.6 %)

Figure 3


Si/Sb
Na/Sb
Si/Ca
31
21
44
20
32
40
30
25
16
24
23
27
29
4
9
2
37
26
13
18
19
33
45
7
8
17
11
3
43
28
5
46
6
41
34
47
22
1
12
39
36
14
35
10
42
15
38
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
PCA dimension 2 (25.8 %)
PCA dimension 1 (74.1 %)