COMPOSITIONAL DATA ANALYSIS OF ARCHAEOLOGICAL GLASS: PROBLEMS AND POSSIBLE SOLUTIONS

C.C. Beardah[1], M.J. Baxter[1], H.E.M. Cool[2]and C.M. Jackson[3]

[1]School of Science, Nottingham Trent University, UK; *michael.baxter@ntu.ac.uk*

[2]Barbican Research Associates, 16 Lady Bay Road, West Bridgford, Nottingham NG2 5BJ, UK

[3]Department of Archaeology, University of Sheffield, Northgate House, West Street, Sheffield S1 4ET, UK

1   Introduction

Table 1, which shows a subset of archaeological glass compositional data for a particular Romano-British glass vessel type, facet-cut beakers, reproduces Table 1 of Baxter and Beardah (2003). The full data set, and data for three other vessel types, is given in Baxter, Cool and Jackson (2003), where the archaeological background is discussed.

Sample	Al	Fe	Mg	Ca	Na	K	Ti	P	Mn	Sb	Pb	Other
10	1.36	0.58	0.49	4.10	17.68	0.67	0.10	0.05	0.06	2.84	0.39	71.68
11	1.77	0.44	0.54	5.51	18.22	0.51	0.07	0.04	0.02	0.76	0.11	72.01
12	2.20	0.36	0.46	7.29	16.71	0.74	0.05	0.06	0.03	1.59	0.01	70.50
13	1.93	0.34	0.45	6.52	18.17	0.49	0.06	0.04	0.02	1.38	0.02	70.58
14	2.38	0.38	0.42	6.63	17.17	1.01	0.06	0.05	0.03	1.47	0.41	69.99
15	1.96	0.33	0.40	6.68	19.46	0.56	0.05	0.06	0.02	2.84	0.14	67.50
24	1.92	0.27	0.25	4.15	17.84	0.53	0.04	0.04	0.03	1.36	0.29	73.28
25	1.84	0.37	0.37	5.66	19.23	0.65	0.06	0.05	0.02	0.98	0.15	70.62
26	1.90	0.43	0.47	5.62	18.94	0.55	0.08	0.04	0.02	0.99	0	70.96
27	1.82	0.33	0.38	6.31	17.80	0.58	0.06	0.05	0.01	0.65	0.08	71.93
28	1.84	0.36	0.40	4.78	18.15	0.46	0.06	0.04	0.02	1.10	0	72.79
29	1.37	0.57	0.45	3.84	18.20	0.65	0.11	0.06	0.07	2.23	0.32	72.13
40	2.10	0.35	0.32	5.21	16.76	0.57	0.05	0.04	0.01	1.23	0.01	73.35
41	2.03	0.33	0.31	5.00	16.73	0.53	0.05	0.03	0.01	1.19	0.01	73.78
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: A subset of compositional data for 63 Romano-British colourless facet-cut beakers. The 'other' category is obtained by difference.

Such data may be analyzed using multivariate methods such as principal components analysis (PCA) to see if there are compositional differences between or within vessel types. For the data shown in Table 1 bivariate analysis reveals two distinct compositional groups in the data. The same is true of other vessel types such as cast bowls and, less confidently, cylindrical cups. The evidence for this is presented in Baxter and Beardah (2003) and Baxter, Cool and Jackson (2003). The distinct groups may be tentatively associated with the use of different sands in the manufacture of the glass.

In the table the final column was obtained by differencing the sum of the other variables from 100% and is labelled here as 'other', though it may be equated approximately with silica. A PCA of the standardized data for both the facet-cut beakers and cast bowls successfully identified the grouping suggested by bivariate methods. A PCA of centered log-ratio data either did not do so, or did not do so very successfully. Since, according to proponents of log-ratio methodology (e.g., Aitchison, Barceló-Vidal and Pawlowsky-Glahn, 2002), the former kind of analysis is incorrect we may ask why such 'incorrect' methodology successfully identifies substantively interpretable structure in the data, whereas the theoretically correct approach does not.

The most important reason appears to be that log-ratio analysis, as it is designed to do, emphasises those variables with a high relative variation. In glass compositional data sets such variables often have a low absolute presence and variation. For the data in Table 1, and other data, the plot of the first two principal components is dominated by Mn and Sb and is similar to a simple bivariate plot of these two variables. Typically, Mn varies in the range 0.01% to 0.04%, and this is consistent with the natural range of variation associated with impurities in the raw materials selected for glass-making. It does not reflect differences in the technology of glass-making, or choice of raw

materials, that grouping associated with other variables present at a higher absolute level can suggest.

To avoid this problem without resorting to incorrect methodology some form of weighted analysis, that downweights the influence of variables such as Mn, seems needed. The ratio-map methodology developed by Greenacre (2002) is one possibility, but it has not been much used with archaeometric data. The same can be said of the use of standardized log-ratios. We explore their use, and other analytical possibilities, in this paper.

A secondary, but by no means trivial, problem for any analysis based on logarithms (as the ratio-map approach is) is the presence of zeroes in the data for Pb. These are really censored data below the level of 0.01%. In order to apply methods based on logarithms these zero values need to be dealt with in some way, and we take the opportunity to explore a variety of strategies for dealing with zeroes, some of which have been proposed only recently.

2   Dealing with zeroes

The following possible strategies have been investigated, where we have specialised to our situation where, at most, one element of a composition is zero and the constant row sum is 100%. If $\mathbf{X}$ is the $n \times p$ data matrix then x_{ij} is a typical value.

1. Merge Pb into the 'other' category.

2. Replace zero values of Pb with some value, α, less than of 0.01%, and recalculate the variable 'other' by differencing.

3. Replace zeroes with α and subtract $\alpha/(p-1)$ from all other elements of the composition, for some small α. This is the additive replacement strategy of Aitchison (1986).

4. Replace zeroes with some small value, α, and other elements by $(x_{ij} - \alpha x_{ij}/100)$. This is the multiplicative replacement strategy proposed by Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2002). It is a particular case of what they call the 'simple replacement strategy', in which zeroes are replaced with a small constant and then all elements rescaled so that the sum is 100%.

The first of these options is not ideal since Pb is a variable that can potentially discriminate between glass vessel types, but it does avoid the potential problem of the other options, that results can be sensitive to the value of α used. In particular, spurious clusters associated with the replaced zeroes may occur. Following the suggestion of Sandford, Pierson and Crovelli (1993) we used $\alpha = 0.0055\%$ for strategy (2) and, for consistency, the same value for strategies (3) and (4). A full account of the properties of the additive and multiplicative replacement strategies is provided in Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2002) who express a clear preference for the latter.

In log-ratio analysis, centered log-ratios of the form

$$y_{ij} = \log[x_{ij}/g(\mathbf{x}_i)] = \log(x_{ij}) - p^{-1} \sum_{j=1}^{p} \log(x_{ij}) \tag{1}$$

may be used, where $g(\mathbf{x}_i) = (x_{i1} x_{i2} \ldots x_{ip})^{1/p}$ is the geometric mean of the data for the ith case. In a successful PCA, a plot based on the first two components approximates the Euclidean distances between cases in p dimensions. When y_{ij} as defined in Equation (1) is used this is sometimes called Aitchison distance and is given by

$$d_{ij}^2 = \sum_{k=1}^{p} \left(\log \frac{x_{ik}}{g(\mathbf{x}_i)} - \log \frac{x_{jk}}{g(\mathbf{x}_j)} \right)^2 \tag{2}$$

In Baxter, Cool and Heyworth (1990) it was shown that this distance could be approximated to first order using

$$y_{ij} = (z_{ij} - \bar{z}_i)/\bar{z}_i \qquad (3)$$

where $z_{ij} = x_{ij}/\bar{x}_j$, and $\bar{x}_j$ and $\bar{z}_j$ are the means for variable j. This approximation was originally derived in an attempt to understand why the results from log-ratio analysis, and a correspondence analysis where elements were divided by their variable mean, were so similar. It suggests that, as an alternative to zero replacement in log-ratio analysis, an approximate log-ratio analysis, based on a PCA of Equation (3) which is unaffected by zeroes, could be used, and this is a fifth strategy that we have investigated. This and strategy (1) will be referred to as 'zero avoidance', as opposed to the zero replacement strategies (2)–(4).

3 Analytical strategies

A large number of analyses were undertaken of the data in Table 1, similar data sets given in Baxter, Cool and Jackson (2003), and various combinations of them. We undertook PCAs of the data treated in the following ways.

1. Standardized as $(x_{ij} - \bar{x}_j)/s_j$ where s_j is the estimated standard deviation of variable j.

2. Log-transformed data, to base 10.

3. Standardized log-transformed data.

4. Centered log-ratio data, as in Equation (1).

5. Standardized centered log-ratio data.

6. Weighted log-ratio data, as in Greenacre's (2002) ratio-map methodology.

7. Approximate log-ratio analysis, using y_{ij} as defined in Equation (3).

For (2)–(6) the five zero replacement or avoidance strategies outlined in the previous section were used.

Data treatments (1), (2) and, to a lesser extent, (3) are common in the archaeometric literature, but would, with the exception to be noted, be regarded as incorrect by proponents of log-ratio analysis. The exception arises when only trace element data are used, when it can be shown that analyses based on the log- and log-ratio transformation are equivalent (Aitchison, Barceló-Vidal and Pawlowsky-Glahn, 2002). For our data the log-ratio approach, (4), suffers from the problems identified in Baxter and Beardah (2003) and at the outset of this paper, and data treatments (5) and (6) represent different ways of attempting to avoid these problems, either by giving all variables equal weight, or downweighting the contributions of those variables with high relative variation but low absolute presence. Treatment (7) should, if the approximation to (4) is reasonable, produce similar results but without any need for a zero replacement strategy.

It is well-known that PCA can be dominated by high-variance variables, and standardization to give variables equal weight is a common response to this problem. Standardization after log-ratio transformation does not, however, seem to have been much used, presumably because it means that the emphasis on high relative variation that is a feature of the log-ratio approach is lost.

If $\mathbf{L}$ is the matrix of logarithmically transformed data, a PCA (or biplot) of log-ratios can be obtained by double-centering $\mathbf{L}$ with respect to the row and column means and using the singular value decomposition (SVD) to obtain a least-squares matrix approximation to the matrix so transformed. In the ratio-map method the double-centering is with respect to weighted averages based on the row and column totals, followed by a weighted least-squares matrix approximation

also based on these totals. Algebraically, the method may be described in the form of a generalized SVD of the weighted doubly-centered data. Geometrically, and with respect to the rows, the aim is to approximate weighted distances between cases defined by

$$d_{ij}^2 = \sum_{k=1}^{p} c_j \left(\log \frac{x_{ik}}{\tilde{g}(\mathbf{x}_i)} - \log \frac{x_{jk}}{\tilde{g}(\mathbf{x}_j)} \right)^2$$

where c_j is the column mass or total and

$$\tilde{g}(\mathbf{x}_i) = (x_{i1}^{c_1} x_{i2}^{c_2} \ldots x_{ip}^{c_p})^{1/p}$$

is the weighted geometric mean of row i (Greenacre, 2002).

4 Results

Analyses are described below for three different glass vessel types. For one type, cylindrical cups, only 1/97 cases had a non-zero value for Pb. It, two outliers, and Pb were omitted from statistical analysis. For the other types, cast bowls and facet-cut beakers there were, respectively, 22/34 and 14/63 zero values. With such a high proportion of zeroes any replacement strategy is likely to be fraught with difficulties, but we will persevere nevertheless.

4.1 Facet-cut beakers

In Baxter and Beardah (2003) bivariate analyses suggested that there were at least two distinct compositional groups for two of the types, cast bowls and facet-cut beakers. These are taken as given for present purposes, and plots will be labelled so that these groups, and cases with zero values of Pb, can be identified. Figure 1 shows a variety of plots based on PCAs of centered log-ratios, for the facet-cut beaker data and Figure 2 shows plots based on PCAs of standardized data and the approximation to log-ratio data of Equation (3), and two ratio-maps obtained using zero avoidance strategy (1) and replacement strategy (2). These and other figures are presented as 'score' plots since the usual emphasis in archaeometric practice is on identifying groups of cases in the data. Their interpretation, and in particular claims to the effect that a particular variable such as Mn or Pb dominates an analysis, has, however, been informed by the inspection of biplots for the analyses.

Only one of the zero replacement strategies (2)–(4) is illustrated as we found that results were very similar, whatever approach was used. Analyses of unstandardized log-transformed data are not shown, since they were very similar to those obtained from the comparable log-ratio analyses. We have observed this phenomenon in numerous previous analyses of ceramics and glass; essentially both methods are dominated by the same small number of variables with high variance on the transformed scale. Analyses of standardized log-transformed data are not shown as the results are very similar to those for standardized data (Fig. 2A). This phenomenon has also previously been observed (Baxter, 1995) and can be expected if there are no clear outliers in the data (on either scale).

In all the analyses where a zero replacement strategy has been used (Fig. 1B,D and Fig. 2D) the zeroes are identified as a separate group. This is also the case for the approximate log-ratio analysis (Fig. 2B). All the analyses separate out the distinctive compositional group, labelled 'a', most clearly in those analyses based on standardized data, whatever the transformation, and least clearly in the unstandardized log-ratio analysis and the approximation to it. The log-ratio analyis using zero replacement (Fig. 1B) and the approximate log-ratio analysis (Fig. 2B) do not seem that similar at first sight, but close examination shows that the patterns revealed are quite similar except that in the latter analysis there is some 'distortion' caused by the two apparent outliers on the plot.
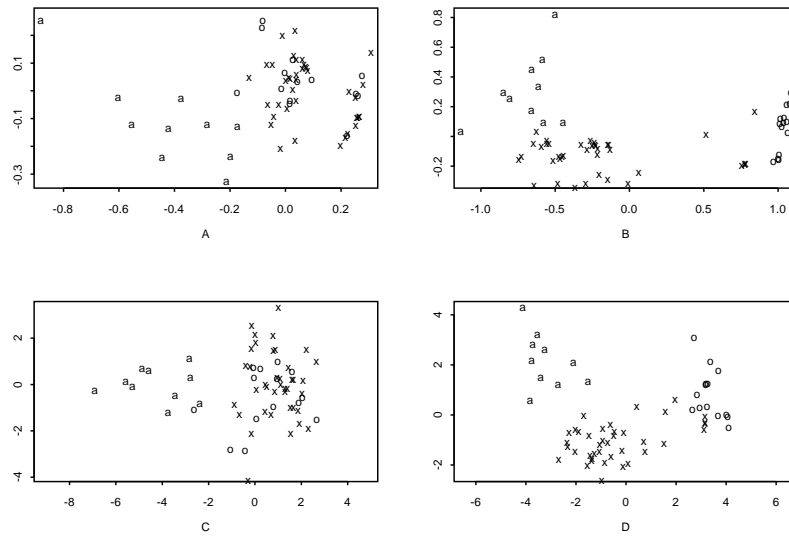
Figure 1: Plots of the first two principal components from a PCA of centered log-ratios of the data in Table 1. Plots A and C use zero avoidance strategy (1); plots B and D use zero replacement strategy (2). Plots A and B are based on unstandardized log-ratios and C and D on standardized ratios. The labels 'a' identify a distinctive compositional group revealed by bivariate analysis (Baxter and Beardah, 2003), and 'o' identifies cases with zero values for Pb.
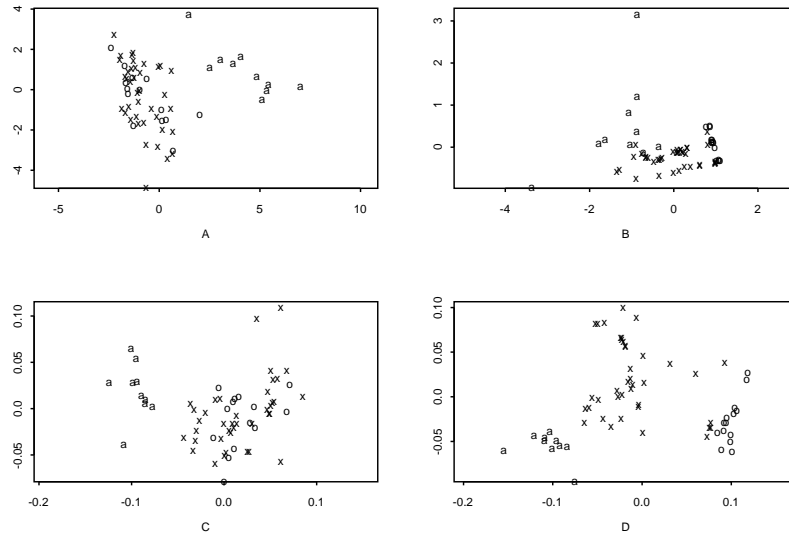

Figure 2: Plots of the first two principal components from a PCA of the standardized data of Table 1 (A); the approximate log-ratio data obtained from Equation (3) (B); and ratio-maps using use zero avoidance strategy (1) and zero replacement strategy (2) (C and D). The labels 'a' identify a distinctive compositional group revealed by bivariate analysis (Baxter and Beardah, 2003), and 'o' identifies cases with zero values for Pb.

4.2 Cast bowls

Figures 3 and 4 present similar analyses for a different type of vessel, cast bowls. The same generalizations made about the facet-cut beaker analysis, concerning the similarities of different methods and replacement strategies apply here. Bivariate analyses suggest two distinct compositional groups in the data, and this forms the basis of the labelling on the plots. As zeroes occur in both groups no attempt has been made to identify them separately, although we note that the zeroes concentrate into groups to the right-hand side of Figures 3B,D and 4B.
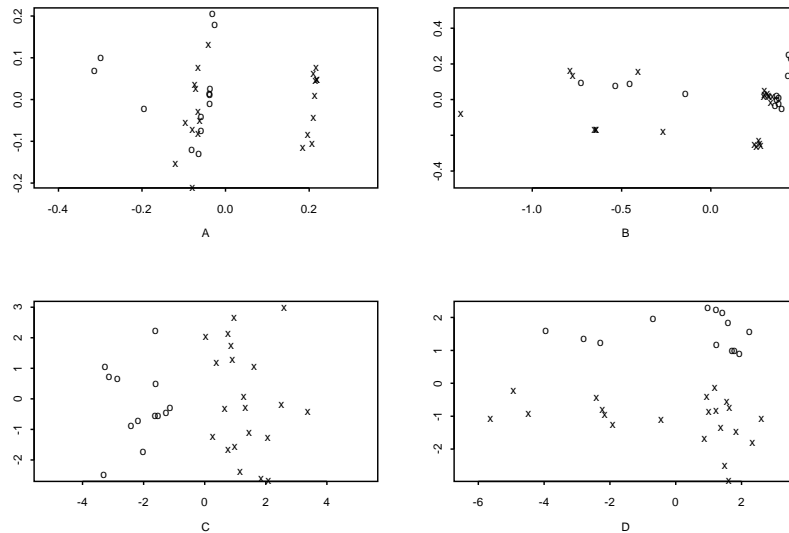
Figure 3: Plots of the first two principal components from a PCA of centered log-ratios of compositional data for Romano-British cast bowls. Plots A and C use zero avoidance strategy (1); plots B and D use zero replacement strategy (2). Plots A and B are based on unstandardized log-ratios and C and D on standardized ratios. The labels 'o' and 'x' identify two distinct compositional groups suggested by bivariate analysis (Baxter and Beardah, 2003).

The unstandardized log-ratio analyses do not successfully suggest the structure in the data. For the zero avoidance strategy (1) (Fig. 3A) the first axis, and the patterns suggested, are almost entirely dependent on Mn; for the zero replacement strategies (Fig. 3B) the first axis, and the patterns suggested, are almost entirely dependent on Pb. We comment on this further in our concluding section. The appearance of the plot based on the approximate log-ratios is dominated by an outlier (Fig. 4B) but, making allowance for this, gives the same message (uninformative) as Figure 3B.

Analyses of standardized data (Fig. 4A) and standardized log-ratio data with zero avoidance strategy (1) (Fig. 3C) identify the groups in the data well. The groups are also evident in plots using a zero replacement strategy (Fig. 3D), but on the second axis, as the first is dominated by Pb. The ratio-map analyses largely, but not completely, separate out the two groups and the zero treatment strategy seems to have little influence on the result (Figs. 4C,D).

4.3 Cylindrical cups

Of 97 cylindrical cups, only one had a non-zero value for Pb. It and Pb have been excluded from the following analyses, along with two outliers.

Figure 5 shows the outcome of a PCA of centered log-ratio data. This is similar to a plot of Sb against Mn, and a group of 'high' Mn specimens with values between 0.05% and 0.25% is identified.
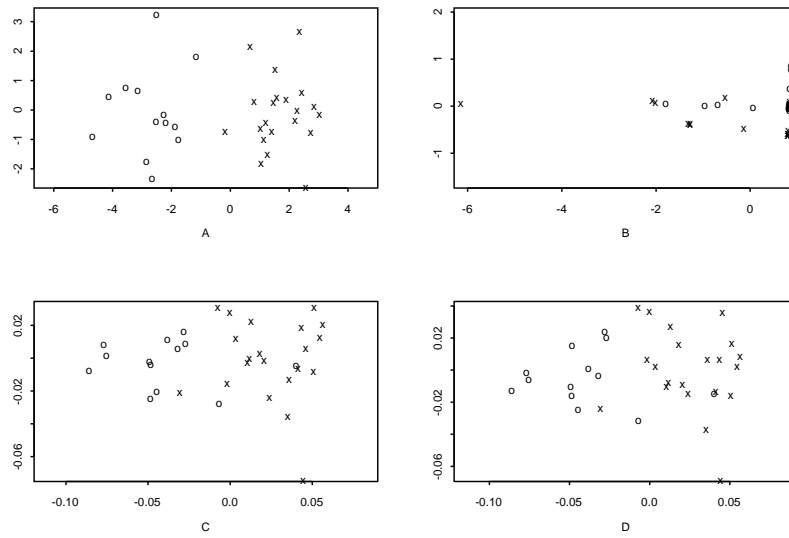
Figure 4: Plots of the first two principal components from a PCA of standardized compositional data for Romano-British cast bowls (A); the approximate log-ratio data obtained from Equation (3) (B); and ratio-maps using use zero avoidance strategy (1) and zero replacement strategy (2) (C and D). The labels 'o' and 'x' identify two distinct compositional groups suggested by bivariate analysis (Baxter and Beardah, 2003).
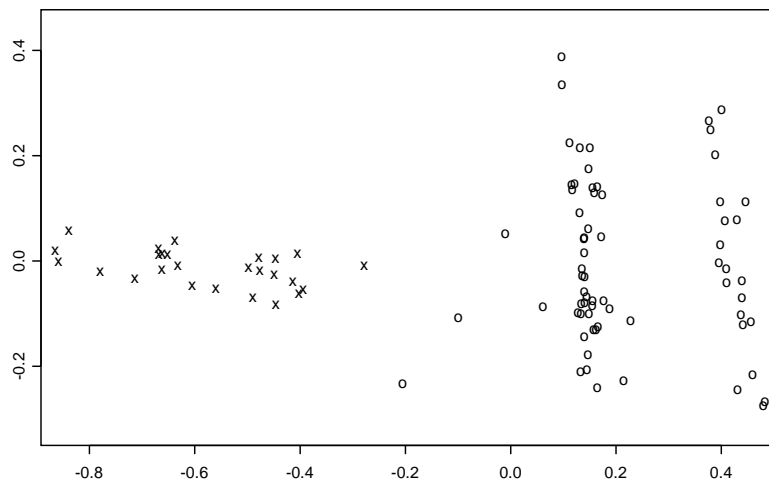

Figure 5: Plots of the first two principal components from a PCA of centered log-ratios of compositional data for 94 Romano-British cylindrical cups.

It is not clear whether this could represent deliberate addition (Mn can be deliberately introduced as a decolourizing agent), but it is noteworthy that there are few similar cases in the other glass vessel types, and for illustration we will treat this as a distinct compositional group.
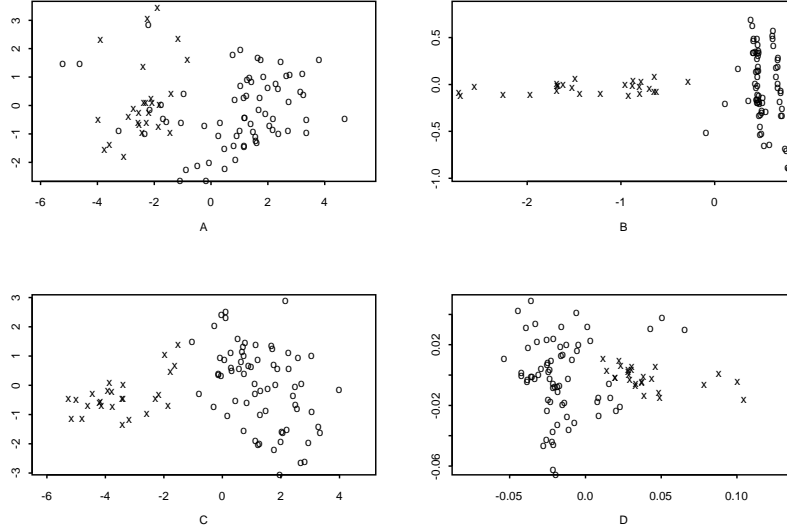

Figure 6: Plots of the first two principal components from a PCA of the standardized compositional data for Romano-British cylindrical cups (A); the approximate log-ratio data obtained from Equation (3) (B); standardized log-ratio data (C); and a ratio-map (D). The labels 'o' and 'x' identify two distinct compositional groups suggested by Figure 5.

Figure 6 shows four other analyses of these data. The approximate log-ratio analysis (Fig. 6B) is very similar to the log-ratio analysis and it and the use of standardized log-ratios (Fig. 6C) separate out the 'high' Mn group. The group plots coherently, but not separately, in the analyses of standardized data (Fig. 6A) and the ratio-map approach (Fig. 6D).

5 Discussion

The investigation reported here was stimulated by our previous observation that, for archaeological glass compositions, log-ratio analysis frequently produced unsatisfactory results. The reason for this is that variables with a low absolute presence but high relative variance (and particularly Mn and Sb in our examples that ignore Pb) often dominate the analysis. This is not 'wrong' but can lead to results that have no substantive interpretation in archaeological terms, even when archaeologically interpretable structure exists. Our analyses of facet-cut beakers and cast bowls illustrates this. The analysis of cylindrical cups is more equivocal. The log-ratio analysis successfully suggests possible structure, but only because it is associated with a variable, Mn, that dominates the analysis.

It may reasonably be noted that our demonstration depends on the assumption that the structure we suggest genuinely exists within the facet-cut beaker and cast bowl data. Such structure was originally identified using bivariate analysis, and its archaeological meaning is discussed in Baxter, Cool and Jackson (2003). We remark here that the compositional groups defined are distinct with respect to several variables (e.g., Fe, Mg, Ca for the cast bowls, and Al, Fe, Sb for the facet-cut beakers) and we have no reason to doubt that they are genuine. For the cylindrical cups the identification of distinct compositions is largely dependent on a single variable, Mn. If this is removed from analysis the 'high' Mn group continues to plot coherently, but not separately.

Given the unsatisfactory nature of log-ratio analysis with data of the kind to hand we sought alternatives that could recapture the structure believed to exist in the data. In doing so we had to deal with the problem of zeroes in one potentially important variable. We distinguished between zero avoidance and zero replacement strategies, investigating three of the latter. There was very little difference between any of them with our data, and they were all unsatisfactory in the sense that analyses of unstandardized data, including the ratio-map approach, were dominated by Pb and isolated those cases with zero Pb as a distinct group. It is possible that this is a function of the replacement value we chose to use, 0.0055%, but there is a more fundamental problem. Even if Pb is increased to 0.01%, it will continue to dominate analyses based on logarithms or log-ratios because of its high relative variance (for the same reason that Mn dominates analyses that ignore Pb). One way of viewing this is that zero replacement simply replaces one problem with another, at least as far as log-ratio analysis is concerned.

The approximate log-ratio analysis used, that avoids the need for zero replacement, seems viable but reproduces the unsatisfactory features of log-ratio analysis. It also had a greater tendency to highlight outliers in some of our examples.

The following observations may be made of the alternatives to unstandardized log-ratio analysis that we explored.

1. The use of PCA of standardized data may be incorrect but did recover interpretable structure that the use of unstandardized log-ratios (and logarithms) failed to detect (without some experimentation).

2. The use of standardized log-ratios (and logarithms) generally worked well, but produced results that were, for the most part, similar to the use of standardized PCA.

3. The ratio-map analyses were adversely affected by the zeroes. For the analysis of facet-cut beakers it nevertheless succeeded in suggesting the groups in the data known to be there. Analysis of the cast bowls was less successful.

In summary, for the two data sets where we are sure there are distinct compositional groups, standardization worked well whatever the prior data treatment. The ratio-map approach, with the caveats entered above, gave reasonable results that were less successful than those for standardized data, but better than those for unstandardized log-ratios.

The examples caution against the unthinking use of log-ratio analysis and show that less correct approaches can provide practically useful results. We have, for the most part, restricted ourselves to analyses based on the full composition, using as many variables as possible. In practice, analysis based on subcompositions of the data (whatever the methodology used) would be explored, and for these data bivariate analysis is adequate for identifying groups in the data. Log-ratio analysis undertaken after merging Pb, Mn, Sb and P into the 'other' category will successfully recover the groups in the facet-cut beaker and cast bowl data. The general problem for multivariate analysis is that a certain amount of experimentation is needed to achieve this. In our examples we know what the groups in the data are and what we are looking for. In general this may not be the case and it cannot be assumed that the variables omitted above are unimportant.

As a final example Figure 7 is reproduced from Baxter, Cool and Jackson (2003). One of our interests in undertaking the analyses described here was to see if we could demonstrate a compositional difference between different glass vessel types without presuming that such differences exist. Figure 7 was obtained, after some experimentation, using unstandardized log-ratios after merging Pb (because of the large number of zeroes) and Mn into the 'other' category. It demonstrates successfully that facet-cut beakers are reasonably distinct from other types and also separates out, to the bottom-left, the distinctive compositional sub-group within the facet-cut beakers. It does so, however, by effectively ignoring Pb which we know to be a variable that helps discriminate between facet-cut beakers and other types, so the analysis is not ideal. The first axis is dominated by Sb,
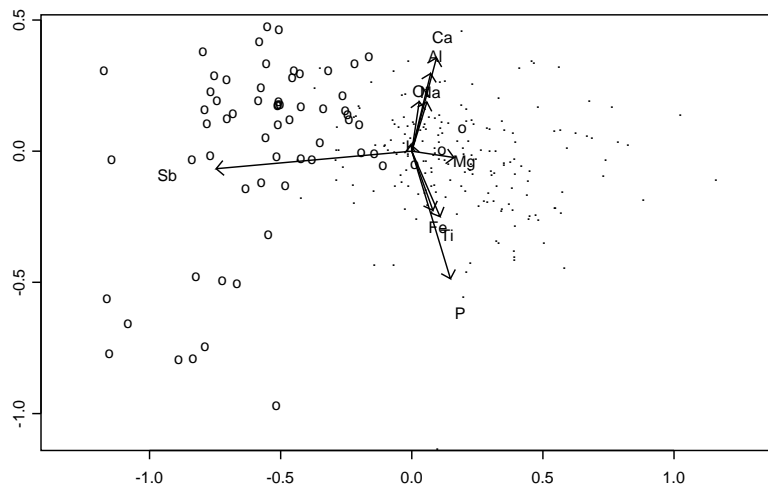
Figure 7: In this form biplot for compositional data the 'other' category includes Mn and Pb. Facet-cut beakers are labelled "o", with other types labelled ".".

but this is not a problem here as the facet-cut beakers tend to have high levels of Sb compared to other types.

Faced with these kind of problems and others identified earlier we prefer to retain a pragmatic approach to practical data analysis rather than letting such analysis be dictated by theoretical concerns divorced from practical problems. A suite of R functions for undertaking the kinds of analysis described here is available from the first author.

Acknowledgments

References

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, London, 416 p.

Aitchison, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2002, Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis: Archaeometry v. 44, p. 295-304.

Baxter, M.J., 1995, Standardization and transformation in principal component analysis, with applications to archaeometry: Applied Statistics v. 44, p. 513-527.

Baxter, M.J., Cool, H.E.M., and Heyworth, M.P., 1990, Principal component and correspondence analysis of compositional data: some similarities: Journal of Applied Statistics v. 17, p. 229-35.

Baxter, M.J., Cool, H.E.M., and Jackson, C.J. 2003, Further studies in the compositional variability

of colourless Romano-British glass: (in preparation).

Baxter, M.J., and Beardah, C.C., 2003, Compositional data analysis in archaeometry: (in preparation).

Greenacre, M.J., 2002, Ratio maps and correspondence analysis: Departament d'Economia i Empresa, Universitat Pompeu Fabra, Working Paper 598. (http://www.econ.upf.es/cgi-bin/onepaper?598)

Martiín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2002, Dealing with zeroes and missing values in compositional data sets using nonparametric imputation: Mathematical Geology (to appear).

Sandford, R.R., Pierson, C.T., and Crovelli, R.A., 1993, An objective replacement method for censored geochemical data: Mathematical Geology v. 25, p. 59-80.