An R Library for Compositional Data Analysis in Archaeometry

C.C. Beardah[1] and M.J. Baxter
School of Biomedical and Natural Sciences, Nottingham Trent University,
Clifton Campus, Nottingham NG11 8NS, United Kingdom.
[1]christian.beardah@ntu.ac.uk

Abstract

Compositional data naturally arises from the scientific analysis of the chemical composition of archaeological material such as ceramic and glass artefacts. Data of this type can be explored using a variety of techniques, from standard multivariate methods such as principal components analysis and cluster analysis, to methods based upon the use of log-ratios. The general aim is to identify groups of chemically similar artefacts that could potentially be used to answer questions of provenance.

This paper will demonstrate work in progress on the development of a documented library of methods, implemented using the statistical package R, for the analysis of compositional data. R is an open source package that makes available very powerful statistical facilities at no cost. We aim to show how, with the aid of statistical software such as R, traditional exploratory multivariate analysis can easily be used alongside, or in combination with, specialist techniques of compositional data analysis.

The library has been developed from a core of basic R functionality, together with purpose-written routines arising from our own research (for example that reported at CoDaWork'03). In addition, we have included other appropriate publicly available techniques and libraries that have been implemented in R by other authors. Available functions range from standard multivariate techniques through to various approaches to log-ratio analysis and zero replacement. We also discuss and demonstrate a small selection of relatively new techniques that have hitherto been little-used in archaeometric applications involving compositional data. The application of the library to the analysis of data arising in archaeometry will be demonstrated; results from different analyses will be compared; and the utility of the various methods discussed.

Key words: archaeometry; compositional data; multivariate analysis; R.

1 Introduction

The development of a library of computational tools for compositional data analysis (CDA) has arisen naturally as a consequence of work reported at current and previous CoDaWork meetings and elsewhere. Recent examples can be found in Baxter and others (2005a, b) and Beardah and others (2003). Our earlier work in this field, for example that reported in Beardah and Baxter (2001), made use of the S-Plus package as the basis for software development. S-Plus features many powerful statistical tools of a kind commonly used in archaeometric study. Its most attractive feature, however, is that it is associated with a high-level programming language, called S (Venables and Ripley, 2000) which allows many non-standard methods to be programmed with relative ease. The major drawback of using S-Plus as a software development platform is its relatively high cost and associated issues of availability.

R (**http://www.r-project.org/**) is an open source package that makes available a powerful array of statistical and graphical facilities at no cost. It is a programming language and environment similar to the S language and environment commercially available as S-Plus. Much code written for S runs under R and the package is available for a wide variety of platforms. Included in an integrated suite of software for data manipulation and graphical display are an effective data handling and storage facility; operators for calculations on matrices and a large collection of tools for data analysis. Since R, like S, is designed around a true computer language, its core functionality can be easily extended via so-called libraries. Several are supplied with the R distribution and many more are available through the Internet. For the reasons outlined above, we now favour the use of R as our software development package of choice.

In this paper we aim to show how, with the aid of statistical software such as R, traditional exploratory multivariate analysis can easily be used alongside, or in combination with, specialist techniques of

compositional data analysis. Our library has been developed from a core of basic R functionality, together with purpose-written routines arising from our own research (for example that reported at CoDaWork '03). In addition, we have included other appropriate publicly available techniques and libraries that have been implemented in R by other authors. In particular, the examples in this paper and in Baxter (submitted) make use of functions made available in the following libraries: **`class`**; **`cluster`**; **`e1071`**; **`fastICA`**; **`MASS`**; **`mda`**; **`nnet`**; **`rpart`**. (Several of these are associated with Venables and Ripley, 2002.) Available functions range from standard multivariate techniques through to various approaches to log-ratio analysis and zero replacement. In addition, we introduce and illustrate some newer methods that might be applied to compositional data. Finally, this paper represents some early work on an assessment of the usefulness of some of these methods for archaeometric applications involving compositional data.

In the next section we give some background detail on the data employed in the first two of our illustrative examples. These and other examples, discussed and illustrated in section 3, are chosen to demonstrate the application of selected components of our CDA library. This is followed in section 4 by a brief discussion of our results.

2 Data

Compositional data naturally arises from the scientific analysis of the chemical composition of archaeological material such as ceramic and glass artefacts. Data of this type can be expressed as an n by p matrix representing n cases and p variables and can be explored using a variety of techniques, from standard multivariate methods such as principal components analysis (PCA) and cluster analysis (Baxter, 1994), to methods based upon the use of log-ratios (Aitchison, 1986). The general aim is to identify groups of chemically similar artefacts that could potentially be used to answer questions of provenance.

Most of our illustrative examples make use of a data set consisting of $n = 241$ specimens of glass, found at sites in Israel and collated by Professor Ian Freestone of Cardiff University. This data set, and sub-sets of it, is the focus of more detailed analysis in Baxter (submitted) and in the CoDaWork'05 companion paper, Baxter and others (2005b). The latter paper illustrates some familiar methodology, partially revisited in section 3.1 below, applied to both standardised and log-ratio transformed data, while the former paper showcases newer methods. Some of these are discussed in section 3.2 below and will be subject to further investigation prior to full implementation in our CDA library.

The $n = 241$ specimens of glass were measured with respect to the chemical composition of SiO_2, CaO, Al_2O_3, FeO, MgO, Na_2O and K_2O. The first five of these $p = 7$ variables are assumed to enter with the sand used in glass-making. The specimens have been classified into five groups, assumed to be of different provenance and compositionally distinct, although we note the possibility that some cases may have been misclassified. As reported in the CoDaWork'05 companion paper, Baxter and others (2005b), two of these five compositional sub-groups, termed *Levantine I* and *Levantine II*, (consisting of $n = 155$ specimens in total) are, with some possible exceptions, known to be separated by geography and chronology.

3 Methodology supported by the CDA library

3.1 Traditional exploratory techniques

Methods such as PCA and cluster analysis (CA) are easy to carry out using standard, built-in R functions called from the command line. However, in an attempt to provide a unified "umbrella" function as the core of our CDA library, we have taken the approach of subsuming as many "standard" methods as possible within a single command.

3.1.1 Example 1

Figure 2 (essentially the same as Figure 2 of the CoDaWork'05 companion paper, Baxter and others, 2005b), can be generated using the R commands shown in Figure 1. This example makes use of the Levantine sand data ($n = 155$, $p = 5$). In Figure 1, line 1 identifies the row numbers of seven specimens considered to be outliers as a result of prior analysis. Line 2 creates the data matrix formed by omitting the outlying rows and considering only columns 1 to 5 (the chemical composition of SiO_2, CaO, Al_2O_3, FeO and MgO; the variables assumed to enter with the sand used in glass-making). Column 9 of the original data matrix, here stored as **`groups`** in line 3, contains the group classification (1-5) of each case;

Levantine I is labelled "4" and *Levantine II* "5". These values are used to label the biplot. (Note that Figure 2 of the CoDaWork'05 companion paper, Baxter and others, 2005b, is labelled by the groups suggested in Figure 1 of that paper and *not*, as here, by the given classification.)

```
1.    outliers <- c(27, 37, 44, 70, 78, 87, 150)
2.    data <- Levantine[-outliers, 1:5]
3.    groups <- Levantine[-outliers, 9]
4.    comp.check(data)
5.    data <- sub2fully(data)
6.    cda(data, analysis="Biplot", labels=as.character(groups))
```

Figure 1: R commands used to generate Figure 2. Subject to a reflection in the horizontal axis and some differences in labelling (explained in the text), this is the same as Figure 2 of the CoDaWork'05 companion paper, Baxter and others (2005b).
Notes: highlighted lines feature CDA library commands; line numbers are given for reference only.

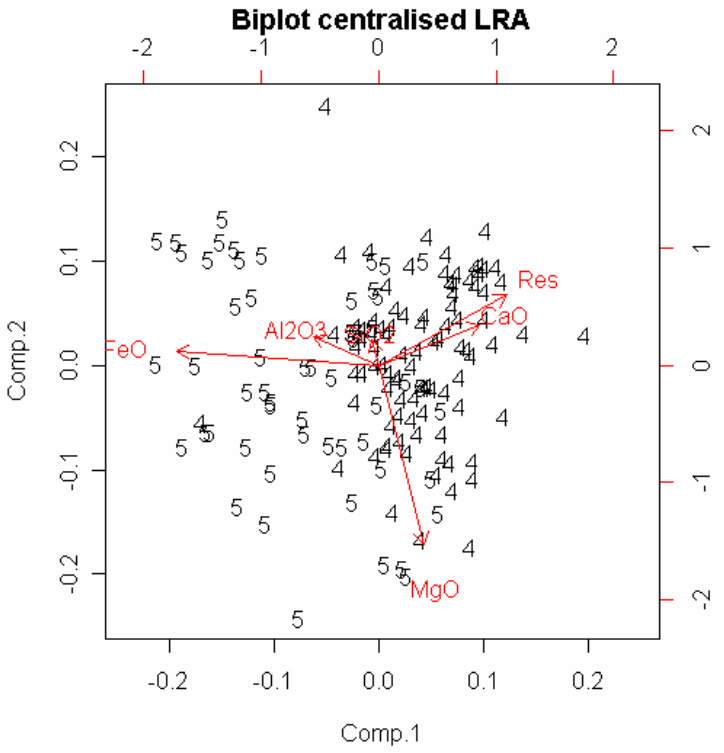

Figure 2: A biplot of the Levantine sand compositional data, using log-ratio analysis after converting to fully compositional data by defining a residual ("Res" in the plot). *Levantine I* is labelled "4" and *Levantine II* "5".
Seven compositional outliers have been omitted.

Line 4 of Figure 1 is a CDA library command used to check whether or not the data matrix is compositional in nature. Here the default row sum of 100 is assumed, but this value can be specified as an additional input argument if desired. In this case the output (not shown) reveals that the data matrix does not contain compositional data, since we have only used the sub-composition consisting of the first five variables; hence rows sum to less than 100 in each case.

In line 5 we process the data matrix to ensure that it is fully compositional; specifying a residual variable by differencing from 100 (Baxter and others, 2005b; Barceló-Vidal, 2003). Finally, line 6 performs the analysis. The simple form of the command in line 6 obscures the fact that extensive options are available. Many other methods are supported, as are various data transformations and zero replacement strategies (section 3.1.3). In this case zero replacement is not necessary and Figure 2 shows the resulting biplot based upon the default centred log-ratio data transformation (Aitchison, 1986).

Supported data transformations include **"none"**, **"scale"** (to standardise), **"clr"** (centred log-ratios; the default) and **"div"**. The latter option implements the approach advocated by Buxeda i Garrigós (1999) whereby the log-ratio transformation is performed with respect to a selected variable used as the divisor. Statistics based on the variation matrix described in Buxeda i Garrigós (1999) are used to return a set of values, one per variable. The maximum of these values determines the variable to use as divisor in the subsequent log-ratio transformation.

3.1.2 Example 2

Our second example makes use of the full data set of n = 241 cases. Figure 1 of Baxter (submitted) is reproduced here as Figure 4 and can be generated using the R commands shown in Figure 3 below. Here we have applied a linear discriminant analysis (LDA) to the raw data. This data is approximately compositional in nature, by which we mean that rows sum to values less than, but generally within 1 or 2 of 100. The **trans** argument to the **cda** function has been used to indicate that no data transformation is to be applied and that the data is not compositional. Alternatively, very similar results can be obtained after first making the data fully compositional by defining a residual variable (Figure 1: line 5) or by making the data completely compositional by normalising using the CDA library command **data <- sub2comp(data)**.

```
1.    data <- IFmaster[, 1:7]
2.    groups <- IFmaster[, 9]
3.    cda(data,   grouping=groups, analysis="LDA",
                  trans=list(method="none", comp=FALSE),
                  labels=as.character(groups))
```

Figure 3: R commands used to generate Figure 4. This is the same as Figure 1 of Baxter (submitted).
Notes: highlighted lines feature CDA library commands; line numbers are given for reference only.

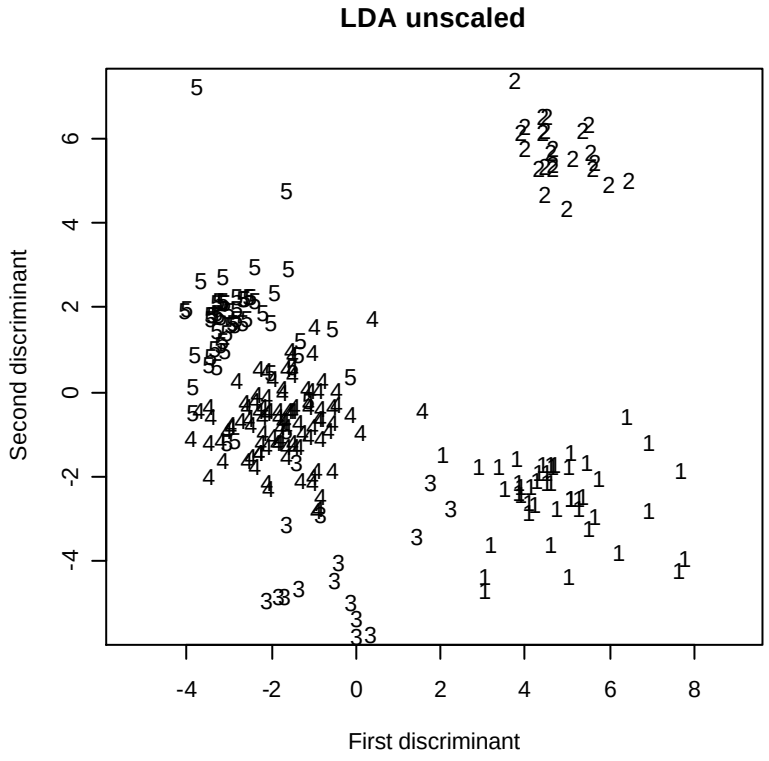

Figure 4: A linear discriminant analysis of the full data set of n = 241 cases labelled by assumed type.
Levantine I is labelled "4" and *Levantine II* "5".

Since the focus of this paper is on methodology and implementation rather than on substantive archaeological interpretation, we note in passing that Figure 4 shows reasonably good separation of the five types; a few fairly clear outliers and some potential misclassifications.

3.1.3 Example 3

At CoDaWork'03, Beardah and others (2003) presented analyses based upon the chemical composition of $n = 63$ colourless Romano-British glass facet-cut beakers, determined by inductively coupled plasma spectroscopy and expressed as percentages. Analyses based upon two other glass vessel types, cast bowls and cylindrical cups were also given. Since the final column of these data (termed the "residual") was obtained by differencing the sum of the other variables from 100%, these data are fully compositional. However, we note that in order to apply methods based upon logarithms, zero values (which occur for PbO) need to be dealt with in some way.

Various methods of "zero avoidance" or zero replacement are discussed and illustrated in Beardah and others (2003). Two of these are illustrated here. Firstly, zero values can be avoided altogether by merging PbO with the "residual". In Figure 5, line 1 defines the data matrix under consideration and line 2 is used to check that these data are compositional. This being the case, we then use the **`residual.add`** function to merge PbO (column 11) with the residual (assumed to be the final column). Line 4 then carries out the ratio-map methodology of Greenacre (2002).

```
1.    data <- bacallsi.Type2[,-1]
2.    comp.check(data)
3.    data <- residual.add(data, 11)
4.    cda(data, analysis="RatioMap")
```

Figure 5: R commands used to generate Figure 6, a ratio-map using the zero avoidance strategy described in the text. Notes: highlighted lines feature CDA library commands; line numbers are given for reference only.

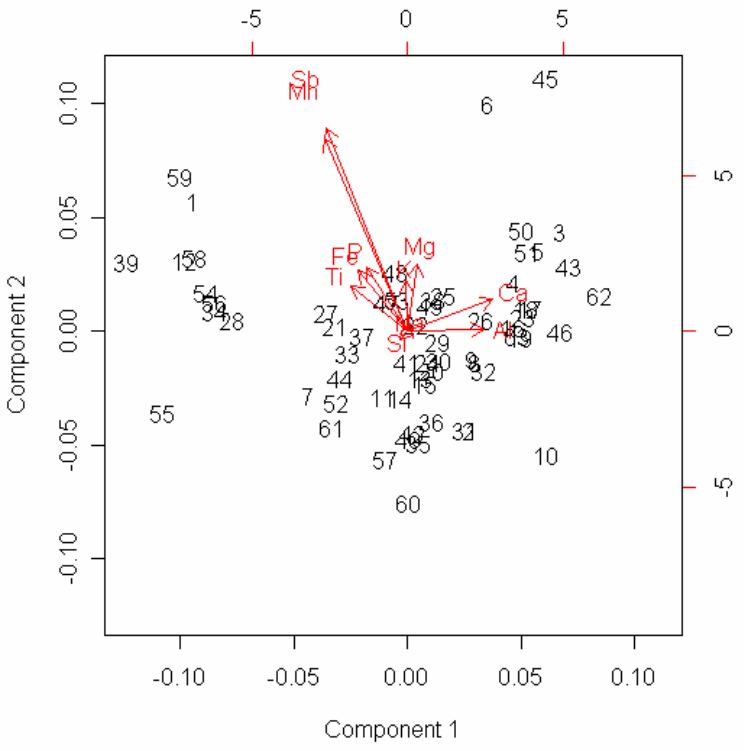

Figure 6: A ratio-map for data representing the chemical composition of $n = 63$ colourless Romano-British glass facet-cut beakers. Zero values are avoided by merging PbO with the "residual" variable.

Several zero replacement strategies (Aitchison, 1986; Martín-Fernández and others, 2003) are supported directly via an optional argument of the **cda** command. For example, specifying

```
zero=list(method="aitch", delta=0.0055, tol=1e-8)
```

enables us to carry out the additive replacement strategy proposed by Aitchison (1986). Here zeroes are replaced by some small value d, here taken to be 0.0055, and we subtract d/(*p*-1) from all other elements of the composition, where *p* is the number of columns in the data matrix. Values less than **tol** are considered to be zero.

In addition to those illustrated earlier, methods supported by our CDA library also include the "weighted correspondence analysis" (WCA) technique of Baxter, Cool and Heyworth (1990). This method is also illustrated in Beardah and others (2003) where it is applied to the data under discussion here. This method provides an approximate log-ratio analysis which is unaffected by the presence of zeroes. Figure 7 shows WCA output that, subject to a reflection in the horizontal axis, is the same as that shown in Figure 2B of Beardah and others (2003). This output can be generated by defining a suitable **data** matrix (Figure 5: line 1) and applying the command **cda(data, analysis="WCA")**.

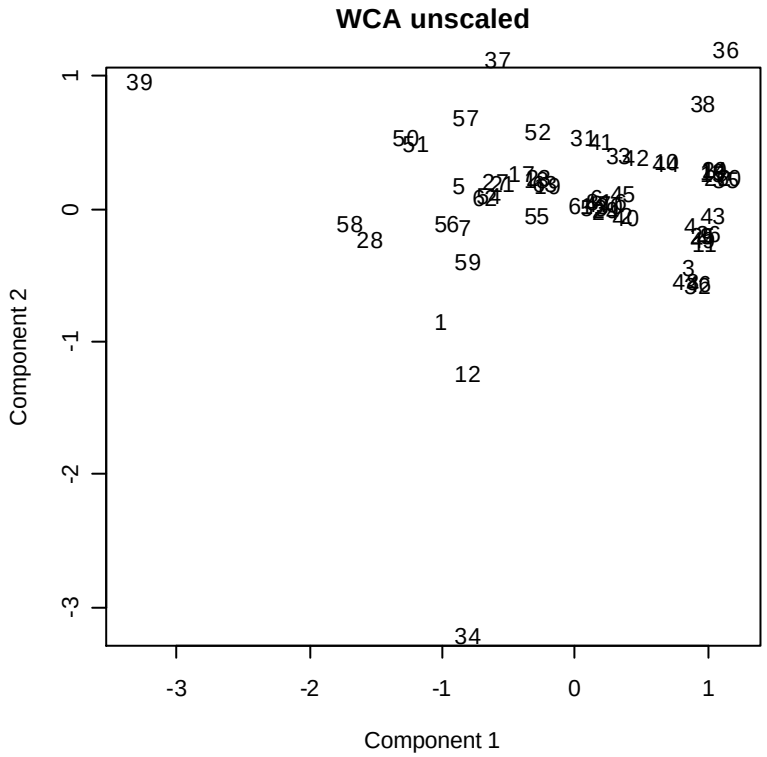

Figure 7: A weighted correspondence analysis for data representing the chemical composition of *n* = 63 colourless Romano-British glass facet-cut beakers. Labels correspond to case numbers.

3.1.4 Example 4

Our final example in this section illustrates the relative variation biplot of Aitchison and Greenacre (2002). Figure 9, subject to a reflection in the vertical axis, reproduces Figure 2 of that paper and can be obtained using the commands shown in Figure 8 below. The data represents the *p* = 6-part colour compositions of *n* = 22 paintings created for teaching purposes. Here a so-called form biplot (the default) has been produced. This preserves distances between rows (Aitchison and Greenacre, 2002). By specifying additional input options we can just as easily produce a covariance biplot that preserves the covariance structure between log-ratios (as in Figure 3 of Aitchison and Greenacre, 2002).

```
1.    data <- colours
2.    comp.check(data)
3.    data <- sub2fully(data)
4.    cda(data, analysis="CompBiplot")
```

Figure 8: R commands used to generate Figure 9, a relative variation (form) biplot for data representing the six-part colour compositions of $n = 22$ paintings created for teaching purposes.
Notes: highlighted lines feature CDA library commands; line numbers are given for reference only.

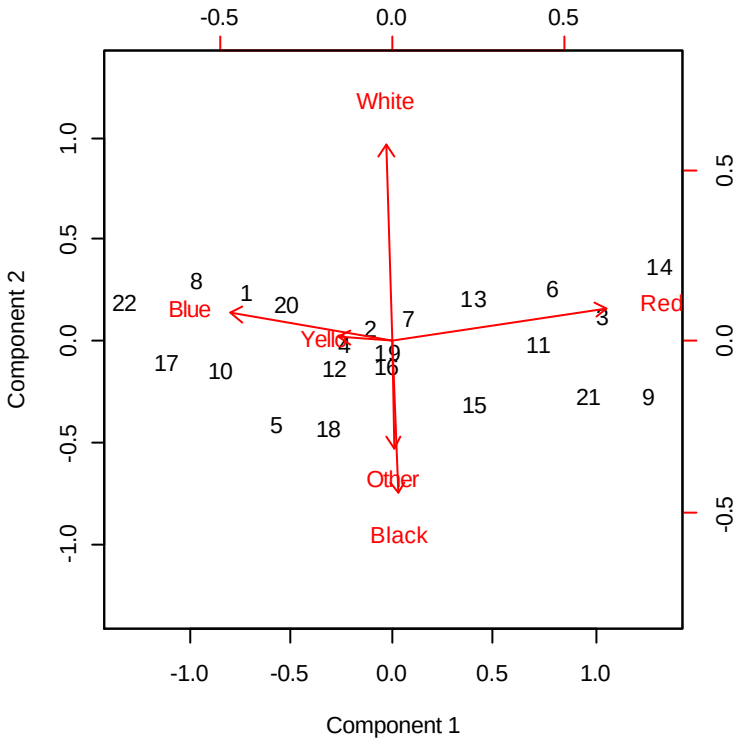

Figure 9: A form biplot for data representing the six-part colour compositions of $n = 22$ paintings created for teaching purposes.

3.2 Self-Organising Maps

The Self-Organising Map (SOM) is an example of the class of so-called unsupervised learning methods. Other examples of unsupervised learning methods include independent components analysis and well-known techniques such as PCA and various forms of cluster analysis. Linear discriminant analysis is an example of a supervised learning method, as are quadratic, flexible and mixture discriminant analysis. Other examples are techniques such as logistic regression analysis, support vector machines, learning vector quantisation, classification trees and feed-forward neural networks. The application of all of these methods (and others) in an archaeometric context, not solely restricted to compositional data analysis, is discussed in Baxter (submitted).

The utility of the methods discussed above to the kind of compositional data encountered in archaeometric applications will the focus of on-going work, as will the incorporation of such methods into our CDA library. Here we restrict ourselves to just one example of the use of SOMs.

Figure 10 shows the output obtained when applying the **`batchSOM`** function (from the **`class`** library in R) to the Levantine sand compositional data, omitting outliers. Figure 9 of Baxter (submitted) is similar. As Figure 10 shows, the outcome of the SOM technique is a graphical display or "map"; cases that are "close" in the p-dimensional space from which they originate hopefully appear close in the map. Here the two groups have separated out reasonably well, although we note that (a), this would not be obvious

without the labelling and (b), as explained in section 2, some cases are potentially misclassified. Since adjacent cells on the map may not, in reality, be particularly close to each other, there may be some advantage in using additional visualisation methods (to be explored) to help identify separate clusters.

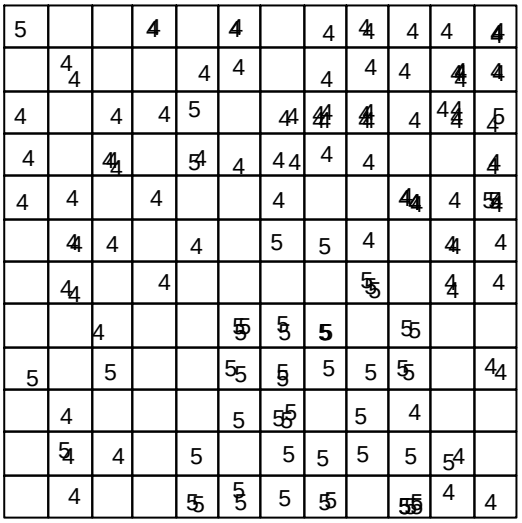

Figure 10: A SOM of the Levantine sand compositional data. Levantine I is labelled "4" and Levantine II "5". Seven compositional outliers have been omitted.

Another possible difficulty with the method is that the final output depends upon the random initialisation used and therefore a different map results each time the method is used, even for identical input data and parameters. Furthermore, as Figure 9 of Baxter (submitted) shows, members of the same group can appear in opposite corners of the map.

4 Discussion

As we have previously stated, in Beardah and Baxter (2001), "when analysing multivariate statistical data of the kind that often arises in archaeometry, it is almost always useful to apply a battery of techniques, rather than one method in isolation. Indeed the long-term aim of our research is not to recommend a particular method of statistical analysis, but to make a variety of methods available in widely accessible and user-friendly form. We hope that this will enable informed users to develop their own views about the advantages and drawbacks of different approaches."

Since it is freely-available, easily extensible and well supported, we have chosen to use the R package as the platform for our software development. In this paper we have illustrated some of the work in progress towards the development of an R library for CDA. However, our consideration of newer techniques such as SOMs does not necessarily mean that we believe that such methods will prove to be of much use for archaeometric data analysis. Some of the newer methods, including many discussed in more detail in Baxter (submitted), were originally motivated by applications very different from those typically encountered in archaeometry. It is possible that "typical" archaeometric data is not sufficiently complex or large, as measured by the number of cases or variables, to warrant the application of some of these methods. Additionally, some techniques require the careful use of various "tuning" parameters. In cases where default values are available the methods can be applied in a "black-box" fashion; however this can make them less accessible to the non-specialist user. For the reasons outlined above, our on-going aim is

to provide a selection of methods, from which the suitably informed user can select those that are of particular use to them.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics* 51, 375-392.

Barceló-Vidal, C. (2003). When a data set can be considered compositional? CoDaWork'03: Compositional Data Analysis Workshop, Girona, Spain. (http://ima.udg.es/Activitats/CoDaWork03/)

Baxter, M.J. (1994). *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh University Press.

Baxter, M.J. (submitted). Supervised and unsupervised pattern recognition in archaeometry. Submitted to *Archaeometry*.

Baxter, M.J., Beardah, C.C., Cool, H.E.M. and Jackson, C.M. (2005a). Compositional data analysis of some alkaline glasses. *Mathematical Geology* 37(2), 183-196.

Baxter, M.J., Beardah, C.C. and Freestone, I.C. (2005b). Compositional analysis of archaeological glasses. Presented at CoDaWork'05: Compositional Data Analysis Workshop, Girona, Spain.

Baxter, M.J., Cool, H.E.M. and Heyworth, M.P. (1990). Principal component and correspondence analysis of compositional data: some similarities. *Applied Statistics* 17, 229-235.

Beardah, C.C. and Baxter, M.J. (2001). Grouping ceramic compositional data: an S-plus implementation. In Z. Stancic and T. Veljanovski (Eds.), *Computer Applications and Quantitative Methods in Archaeology 2000*, BAR International Series 931, pp 53-9. Oxford: Archaeopress.

Beardah, C.C., Baxter, M.J., Cool, H.E.M. and Jackson, C.M. (2003). Compositional data analysis of archaeological glass: problems and possible solutions, CoDaWork'03: Compositional Data Analysis Workshop, Girona, Spain. (http://ima.udg.es/Activitats/CoDaWork03/)

Buxeda i Garrigós, J. (1999). Alteration and contamination of archaeological ceramics: the perturbation problem. *Journal of Archaeological Science* 26, 295-313.

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)

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeroes and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* 35(3).

Venables, W.N. and Ripley, B.D. (2000). *S Programming*. New York: Springer-Verlag.

Venables, W.N. and Ripley, B.D. (2002). *Modern applied statistics with S (4th edition)*. New York: Springer-Verlag.