Grouping Ceramic Compositional Data: An S-Plus Implementation

Christian C. Beardah and Mike J. Baxter

Department of Mathematics, Statistics and Operational Research, The Nottingham Trent University

Clifton, Nottingham NG11 8NS, United Kingdom

e-mails: christian.beardah@ntu.ac.uk, michael.baxter@ntu.ac.uk

Abstract

The scientific analysis of the chemical composition of ceramics naturally leads to multivariate data that is often explored using techniques such as principal components analysis and cluster analysis. The general aim is to identify groups of chemically similar artefacts. Separate groupings could be assumed to indicate, for example, distinct origins or processing recipes of the raw materials (clay) used to make the artefacts.

In this paper we aim to show how, with the aid of powerful statistical software such as S-Plus, traditional exploratory multivariate analysis can easily be used alongside, or in combination with, a technique designed specifically for grouping ceramics by chemical composition (Beier and Mommsen 1994). This latter technique involves grouping together artefacts whose chemical compositions are "close" with respect to a mathematical measure of dissimilarity. The measure used takes into account uncertainty of measurement and the possibility of constant shifts in the data due, for example, to dilution of the clay or to instrumental variation.

We shall demonstrate the ease of implementation in S-Plus of this methodology and show how graphical exploratory techniques can also be used to create an integrated approach to the grouping of ceramic chemical compositional data. The methodology will be applied to real data of this type.

Key words: ceramics, compositional data, multivariate analysis, cluster analysis, S-Plus

1. Introduction: archaeological and statistical background

This paper presents on-going work by members of the EU funded GEOPRO TMR Network. The aims of the project are to "design, develop and apply a novel approach which integrates fully geochemical and mineralogical analytical techniques, to further the study of the provenance and technology of archaeological ceramics and their raw materials".

From a statistical point of view the early stages of our work has involved an investigation into "state of the art" techniques for the analysis of data resulting from (a) geochemical and (b) mineralogical analysis of ceramics (Baxter et al. in press). Our ultimate aim is to produce a software package, using commercial tools, for the analysis and presentation of mixed-mode data. By *mixed-mode* data we mean that containing *both* geochemical and mineralogical information.

Mineralogical data can consist of variables of a mixture of types, including continuous, categorical and presence/absence. By contrast, geochemical data is typically continuous in nature. In addition, modern analytical methods such as Neutron Activation Analysis (NAA) can provide measurements for the concentration of over 20 elements in a ceramic. For this reason, geochemical data is usually high-dimensional. Such data, represented mathematically as an *n* by *p* data matrix, is often explored using techniques such as principal components analysis and cluster analysis (Baxter 1994). The general aim of such analyses is to identify groups of chemically similar artefacts. Separate groupings could be assumed to indicate, for example, distinct origins or processing recipes of the raw materials (clay) used to make the artefacts.

Here we concentrate solely on the implementation of statistical methods for geochemical data. In particular, we discuss an implementation of a technique designed specifically for grouping ceramics by chemical composition (Beier and Mommsen 1994). This technique involves grouping together artefacts whose chemical compositions are "close" with respect to a mathematical measure of dissimilarity (see section 2.2 for details). The measure used takes into account uncertainty of measurement and the possibility of constant shifts in the data due, for example, to dilution of the clay or instrumental variation.

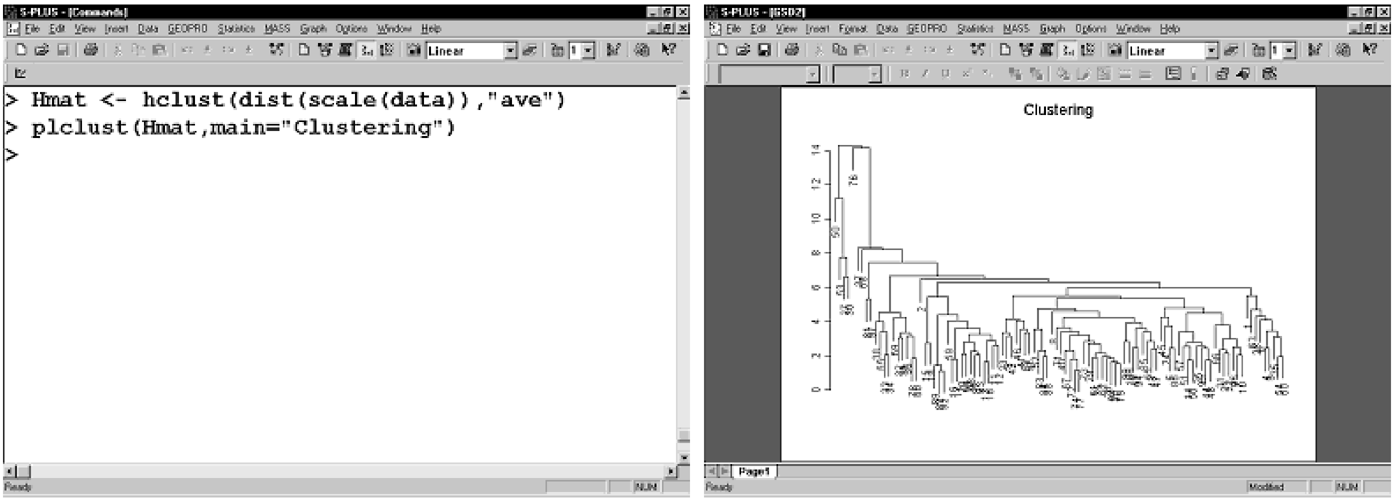
Although this work is at a relatively early stage, a number of decisions with long-term implications have been taken. One of these is to use the S-Plus package as the basis for the development and distribution of our software. This package has many powerful "off-the-shelf" statistical tools of a kind commonly used in archaeometric study. Its most attractive feature, however, is that it is associated with a powerful programming language, called "S" (Venables and Ripley 2000) which allows many non-standard methods to be programmed with relative ease. The intention is to program non-standard methods of choice and to provide the end user with a user-friendly and accessible "front end", through the S-Plus user interface, which can also be manipulated via the S programming language.

More background detail on statistical issues is given in the next section. This includes a brief description of Beier and Mommsen's method of grouping, based upon Mahalanobis distance. Section 3 of the paper presents an archaeometric case study and is followed in section 4 by a discussion of some software-related issues. Our conclusions are contained in section 5.

2. Statistical issues

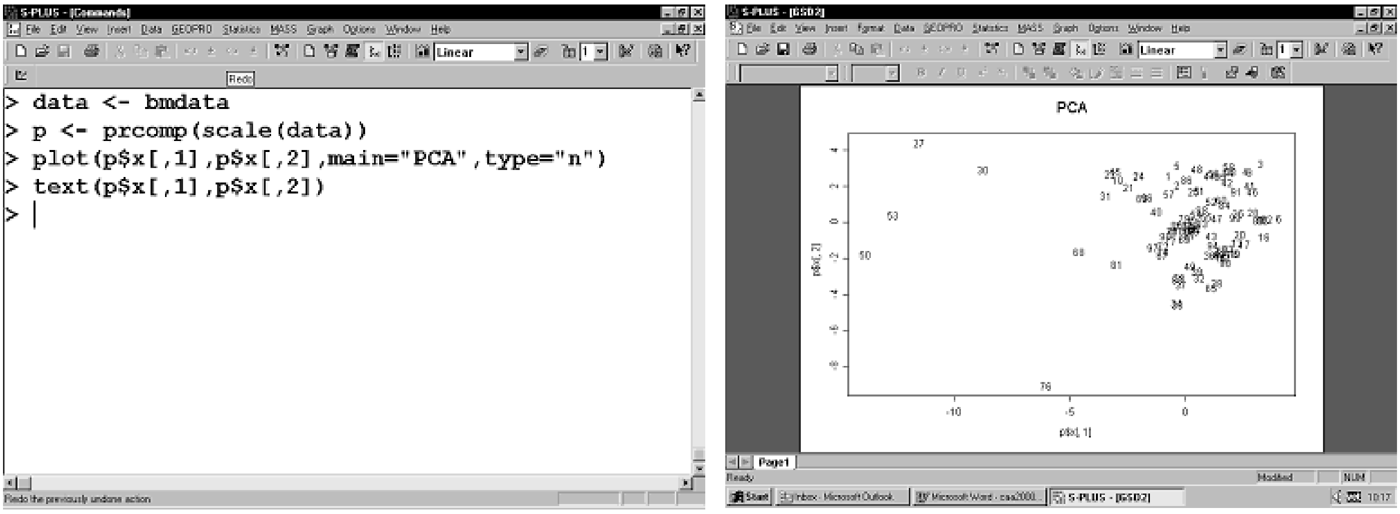
2.1. Traditional exploratory techniques: principal components analysis and cluster analysis

When analysing multivariate statistical data of the kind that often arises in archaeometry, it is almost always useful to apply a bat-



```
> Hmat <- hclust(dist(scale(data)),"ave")
> plclust(Hmat,main="Clustering")
>
```

Figure 1: (a) S-Plus commands and (b) resulting output when carrying out a PCA. These data consist of 27 elemental concentrations for each of 100 ceramic samples from Nichoria in the Peloponnese and are part of the Perlman-Asaro databank of Mycenaean samples (Mommsen et al. in press).



```
> data <- bmdata
> p <- prcomp(scale(data))
> plot(p$x[,1],p$x[,2],main="PCA",type="n")
> text(p$x[,1],p$x[,2])
> |
```

Figure 2: For the same data set as in figure 1, (a) S-Plus commands and (b) resulting output when carrying out an average linkage CA.

tery 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.

As previously mentioned, analysis of the chemical composition of ceramics naturally leads to multivariate data that is often explored using techniques such as principal components analysis (PCA) and cluster analysis (CA). These methods are easy to carry out using standard, built-in S-Plus functions and can be called from menus and associated dialogue boxes, or from the command line. For example, figures 1 and 2 show (a) S-Plus commands and (b) the resulting output when carrying out respectively a PCA and an average linkage CA on the same data set. The S programming language may be thought of as a collection of commands, like those shown in figure 1(a) and figure 2(a), that implement powerful statistical and mathematical operations, together with traditional programming constructs such as looping, conditional (if … else) branching and facilities for graphical and textural output.

In the next section we outline a grouping method (Beier and Mommsen 1994) based upon the Mahalanobis distance between a point, and a group of points, in *p*-dimensional space. Here groups are iteratively “grown” from an initial group that can be as small as just one object. Although the idea behind this methodology is simple, the existing implementation is less so. Our aim is to make the procedure more accessible by implementing it within a commercially accessible package, namely S-Plus.

One way in which this can be achieved is by utilising the “traditional” exploratory techniques discussed here to provide some of the many methods that could be used to obtain an initial grouping. Other possible methods include scanning the data set for pairs of objects that are close together in the Euclidean sense, or using less well known exploratory methods, such as kernel density estimation in combination with contouring (Beardah et al. 2000), to identify potential groups. Several of these possible starting procedures have been provided as options in our S-Plus implementation of Beier and Mommsen’s grouping procedure.

2.2. Statistical concepts: grouping using Mahalanobis distance

The grouping method introduced by Beier and Mommsen (1994) is based largely upon the Mahalanobis distance between a point, $\underline{x}$, in *p*-dimensional space, and a group with mean $\underline{y}$ and *p* by *p* covariance matrix S. This is given by

$$d_{xy}^2 = \left(\underline{x} - \underline{y}\right)^T \Sigma^{-1} \left(\underline{x} - \underline{y}\right) \qquad (1)$$

where $\underline{x}$ and $\underline{y}$ are represented by *p* by 1 vectors.

In simple terms, new objects are added to an existing group if they are "close" to the centre of the group. Groups are iteratively "grown" from an initial group that can be as small as just one object. As discussed in section 2.1, "traditional" exploratory methods could be used to provide this initial grouping. Once a group has ceased to grow, the procedure is re-started from a different initial group. This process is continued until the whole data set has been classified as belonging to a group or not. The original presentation of this methodology (Beier and Mommsen 1994) extends the basic idea discussed above to take into account (a) uncertainty of measurement and (b), the possibility of constant shifts in the data due, for example, to dilution of the clay or instrumental variation. These additional properties are achieved by using modifications to the definition of Mahalanobis distance given in equation (1). The most general modified Mahalanobis "distance" used by Beier and Mommsen is given by

$$d_{xy}^2 = \left(f_0 \underline{x} - \underline{y}\right)^T \left(f_0^2 S_x + S_y\right)^{-1} \left(f_0 \underline{x} - \underline{y}\right) \cdot$$

Here S_y denotes the covariance matrix of the group (previously denoted by S), S_x denotes a diagonal matrix, the *k*th diagonal of which is the squared "uncertainty" associated with the *k*th element of $\underline{x}$ and f_0 is the "best-relative fit" factor, calculated as the solution of

$$\frac{\partial}{\partial f}\left[\left(f\underline{x} - \underline{y}\right)^T \left(f^2 S_x + S_y\right)^{-1} \left(f\underline{x} - \underline{y}\right)\right] = 0 \cdot$$

In other words f_0 is chosen to minimise d^2_{xy}. The calculation of f_0 can be achieved using standard numerical minimisation methods.

It should be noted that, in order for the Mahalanobis distance to be used, the size of the current group, n_g, needs to be greater than the number of variables, *p*, indeed $n_g >> p$ is preferable. If the size of the current group is too small for the Mahalanobis distance to be utilised, weighted Euclidean distance is used instead. As the analysis of chemical composition of artefacts in archaeology typically involves "large" numbers of variables and "small" sample sizes, this restriction can be problematic for our main application. Fortunately, this potential drawback can be alleviated in several ways. For example, variable selection methods can be used to reduce the number of variables considered, without adversely affecting the outcome of the statistical analysis (Baxter and Jackson in press). Alternatively, we could apply a dimension reducing technique such as PCA to the data before subjecting the transformed data to the grouping procedure, using a subset of components.

Of course the outcome of this method depends upon what we mean by new objects being "close" to the centre of the current group. If we desire the grouping procedure to act in a completely automatic manner, without user-intervention, we may quantify "closeness" by the definition of two threshold values for the weighted Euclidean and Mahalanobis distance, D_E and D_M respectively. If the size of the current group is big enough to measure distances in the Mahalanobis sense then we add a new object $\underline{x}$ to the current group if $d^2_{xy} < D_M$. On the other hand, if the size of the current group is small then we add a new object to the current group if $d^2_{xy} < D_E$. Based on properties of d^2_{xy} under the assumption that groups have a multivariate normal distribution Beier and Mommsen (1994) suggest using D_M=1.6.

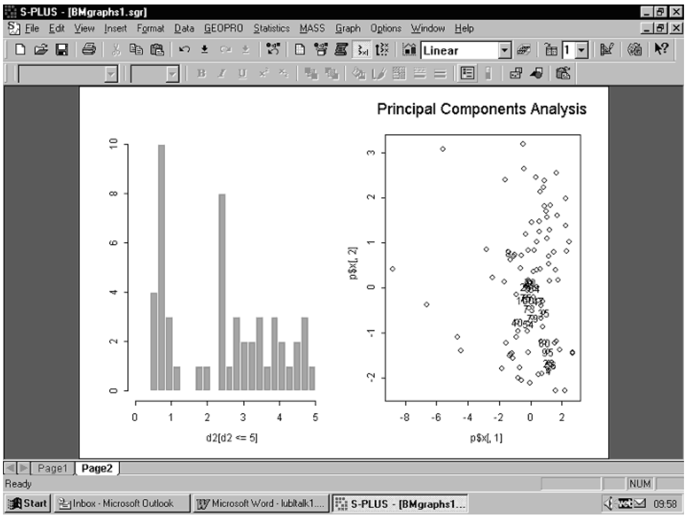

S-PLUS - [BMgraphs1.sgr]
File Edit View Insert Format Data GEOPRO Statistics MASS Graph Options Window Help
Linear
1
Principal Components Analysis
0
2
4
6
8
10
0 1 2 3 4 5
d2[d2 <= 5]
p$x[, 2]
-2
-1
0
1
2
3
-8 -6 -4 -2 0 2
p$x[, 1]
Page1
Page2
Ready
NUM
Start
Inbox - Microsoft Outlook
Microsoft Word - lubltalk1....
S-PLUS - [BMgraphs1...
09:58

Figure 3: The left-hand plot shows a histogram of squared distance values. A plot of the first two PC scores is shown on the right. Labels have been used to highlight the objects in the current group. These data consist of a subset of 11 elemental concentrations for the same 100 ceramic samples from Nichoria in the Peloponnese considered throughout the paper (Mommsen et al. in press).

Applying the methodology as a "black box" as described above may be appropriate in some circumstances, for example if the data set is very large (Beardah et al. in press). In other situations more routinely encountered in the analysis of ceramics, Beier and Mommsen's grouping procedure is best used interactively, with the user making decisions regarding what is meant by an object being "close" to the centre of a group. To illustrate, suppose a group is in the process of being "grown". The group will have a mean $\underline{y}$ and will consist of n_g objects that are "close" to $\underline{y}$ as measured by either weighted Euclidean or Mahalanobis distance depending upon the current size of the group. To assess whether the group can be grown further, we measure the distance of all objects in the data set from $\underline{y}$. This results in *n* squared distance values $d_1^2, d_2^2, \ldots, d_n^2$. As a visual aid to group growth, a histogram of the values d_i^2 for *i=1, …, n* can be plotted. For a well-defined group of objects separated from the remainder of the data set a "valley" or "edge" should occur at relatively small values of squared distance. The range of values of squared distance where this occurs will depend upon the number of variables under consideration (Beier and Mommsen 1994). For example, in the case of the *p=27* elements considered in our example in section 3, we would expect to see the "valley" or "edge" at values of squared distance in the range of 1.5 to 2.0.

The histogram on the left of figure 3 shows an example of the idea discussed above, as featured in our S-Plus implementation of Beier and Mommsen's grouping procedure. As the next step in this particular grouping process we would have to decide whether to use

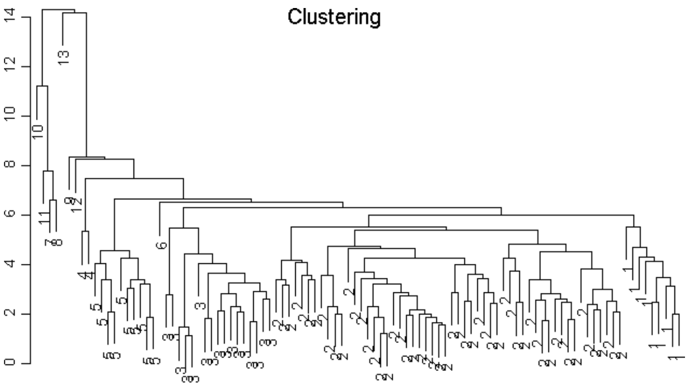


Figure 4: A grouping of the data provided by cutting the dendrogram resulting from an average linkage CA in S-Plus. (In this case the dendrogram was cut at a value of just less than 6.) The user is prompted to enter a group number that will define the initial group for Beier and Mommsen's grouping procedure.

a cut-off value for squared distance of approximately 1.5, or approximately 2.1. As can be seen from the histogram, the latter value would result in the addition to the group of two objects that the former value would leave out. In fact, in this case a cut-off value of approximately 1.5 would result in the group ceasing to grow at this point whereas the higher cut-off value would result in the group growing further.

As an additional aid to making a decision regarding an appropriate cut-off value, the plot of the first two PC scores on the right of figure 3 labels the objects currently in the group. By clicking the mouse on the histogram, the user can experiment with the effect of choosing different cut-off values. After each choice, the PC scores plot on the right is updated and user is asked to accept the current group or to choose an alternative cut-off value. Once a group is accepted, the next iteration of the grouping procedure can take place. This consists of the calculation of (a) the current group mean and (b), the squared distances from each object in the data set to the current group mean. At this stage a new histogram is plotted and the group may or may not grow further, subject to the choice of cut-off value. (Note that as the choice of cut-off value is unrestricted, the user also has the option of *shrinking* the current group.)

Interaction with the graphical display of the type reported above is easily implemented within S-Plus and can also be used in other ways. For example, figure 2(b) of section 2.1 shows the graphical output resulting from carrying out an average linkage CA in S-Plus. As mentioned in section 2.1, such traditional exploratory techniques could be used as a method of choosing initial groups for Beier and Mommsen's procedure. In the case of output from a cluster analysis, the user may achieve this by "cutting" the tree with the mouse at an appropriate point on the dendrogram. The result is a grouping of the data set defined by the point at which the tree was cut. Figure 4 shows such a grouping, labelled by group number. The user is then prompted to enter a group number that will define the initial group for Beier and Mommsen's procedure. PCA output can also be used to identify an initial grouping by clicking on points in a plot of the first two PC scores.

3. Case study

Our implementation of Beier and Mommsen's grouping procedure is now illustrated using a data set consisting of 27 elemental concentrations (measured by Neutron Activation Analysis (NAA) in Berkeley) for each of 100 ceramic samples from Nichoria in the Peloponnese. These data form part of the Perlman-Asaro databank of Mycenaean samples. More information can be found in Mommsen et al. (in press). Using the original implementation of the methodology described here, it is claimed that the majority of these data can be divided into 6 well-separated groups. (Several chemical loners, that is objects not belonging to any group, are also identified.) Rather than presenting a full grouping of the data set, we discuss the growth of three example sub-groups that, for illustration, utilise a variety of methods for providing an initial grouping.

Before describing the grouping process in more detail, it is worth pointing out that presently ours is not a full implementation of Beier and Mommsen's grouping procedure. Their original implementation consists of several thousand lines of FORTRAN code and has evolved over many years to include subtleties that are not yet reflected in our S-Plus implementation. Furthermore, the published grouping of these data (Mommsen et al. in press) is influenced by scientific and statistical decisions not replicated here. For example, we perform our grouping based upon the full data matrix of 100 objects by 27 variables, whereas Mommsen et al. use 25 variables, omitting Ca and Na. It would therefore be surprising if the procedure outlined below gave the same grouping, however we do hope to get a similar grouping.

3.1. Group 1

A preliminary scan through the data set, searching for pairs of objects that are close with respect to weighted Euclidean distance reveals that, for example, objects 74 and 77 are very close together. Using these two objects as our initial grouping the group mean is calculated and, with the aid of a histogram of (squared Euclidean) distance values, we look for objects that are close to our current group. Using a cut-off value of approximately 3.5 adds the four nearest objects to the group, which now consists of objects 26, 61, 64, 70, 74 and 77. The group mean is re-calculated and we again look for objects that are close to our current group. Using a cut-off value of approximately 1.6 adds four more objects to the group, which now consists of objects 26, 61, 62, 64, 70, 73, 74, 75, 77 and 79. The group mean is again re-calculated and the resulting histogram of squared distance values is shown in figure 5. This reveals (a) that the ten members of our current group are all close to the group mean and (b), shows a pronounced "edge" at a distance value of less than 1, meaning that the group as a whole is distant from other objects. The PCA plot on the right also illustrates that the final grouping is quite compact (compare with the full PCA plot in figure 1(b)). Our current group of ten objects is therefore classified as a genuine sub-group within these data. All the members of this sub-group are contained within a sub-group of 13 objects identified by Mommsen et al. (in press). The objects missing from our sub-group are 47, 67 and 72. It is worth noting that the closest object to our sub-group of ten is object 67, but at a distance that makes inclusion difficult to justify. Object 72 is also close to the group, being the fourth nearest object not included.

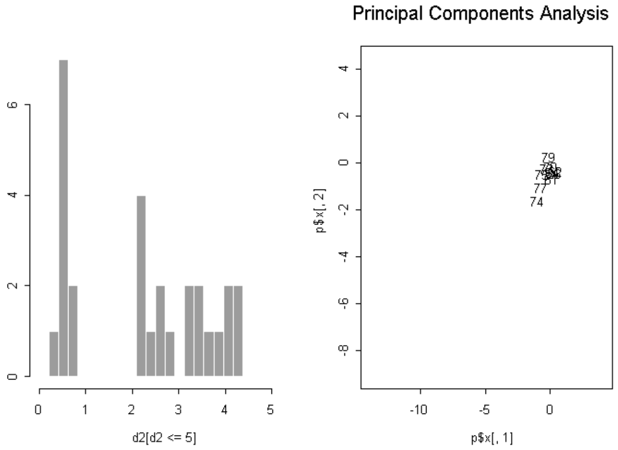


Figure 5: The result of growing the initial grouping consisting of objects 74 and 77. The histogram of squared distances (left) reveals a pronounced "edge" at a value of less than 1 and the labelled PCA plot on the right reveals that the final grouping is quite compact (compare with the full PCA plot in figure 1(b)).

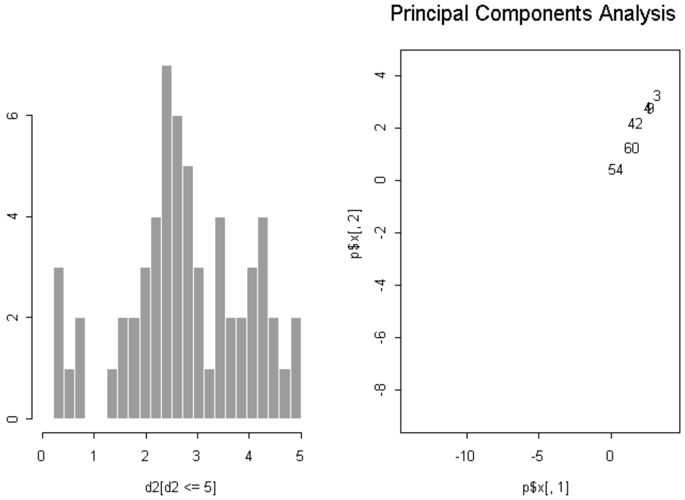


Figure 6: The result of growing the initial grouping consisting of objects 3, 4 and 9.

3.2. Group 2

Having found one sub-group, we now continue, trying to find others. An average linkage CA reveals that objects 33, 34, 38 and 65 are close together. Using these objects as our initial grouping the stages of Beier and Mommsen's grouping procedure give the results presented in table 1. The process stabilises after stage three yielding a sub-group of nine objects. This sub-group is almost identical to a sub-group of ten objects identified in Mommsen et al. (in press). The only difference is that object 37 is missing from our sub-group. Indeed this object lies some distance from the centre of our sub-group.

3.3. Group 3

As a final example, we note that a PCA plot of the first two PC scores reveals that objects 3, 4 and 9 are close together in the extreme top-right of the plot. Selecting these objects directly from the PCA plot (by clicking on them with the mouse) and growing the group yields the results presented in table 2. The process stabilises after stage two yielding a sub-group of six objects. Note that this method of selecting an initial grouping is presented mainly to illustrate the kind of interaction that is possible within S-Plus. Though it works well in this case, using the first two PC scores in this way is less preferable than the methods of initial group selection discussed in sections 3.1 and 3.2.

Stage	Group	Group size	Approximate cut-off value
0	33 34 38 65	4	
1	32 33 34 36 38 65 78 80	8	1.8
2	32 33 34 36 38 39 65 78 80	9	1.5
3	32 33 34 36 38 39 65 78 80	9	1.5

Table 1: Growth of the initial grouping of objects 33, 34, 38 and 65 (identified using an average linkage CA).

Stage	Group	Group size	Approximate cut-off value
0	3 4 9	3	
1	3 4 9 42 54 60	6	2.1
2	3 4 9 42 54 60	6	1.3

Table 2: Growth of the initial grouping of objects 3, 4 and 9 (identified using a plot of the first two PC scores).

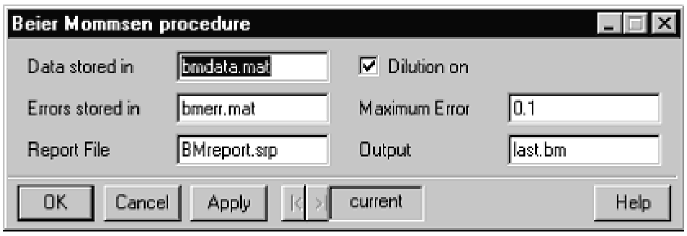
The final histogram of squared distance values is shown in figure 6. Again note the pronounced "valley" around distance values of approximately 1. It is possible to further grow this group as six objects occur at distances of between 1.4 and 2.0 from our subgroup of six objects. However, if this is attempted the group quickly spreads out and the resulting histograms of squared distance fail to reveal the "valley" or "edge" that signifies a coherent sub-group. We therefore classify our six objects as a genuine sub-group within these data. This sub-group is identical to one identified by Mommsen et al. (in press).

4. Software issues

4.1. Extending the graphical user interface

Using S-Plus enables user-friendly calling of user-developed routines. For example our implementation of Beier and Mommsen's grouping procedure can be called from a menu list consisting of other novel grouping methods. Selecting this menu item results in the appearance of a dialogue box of the type shown in figure 7. The specific details of this dialogue box are less important than the ease with which it can be created from within S-Plus. However, the three points outlined below are worth making.

1. The dialogue box controls input to the grouping routine by specifying the names of the S-Plus objects containing the data and the measurement errors if known.
2. The dialogue box also controls output from the grouping routine by specifying the names of an S-Plus report file and an S-Plus output object. The latter contains numerical information that results from the statistical analysis and the former provides a user-friendly format for this information by embedding it in a specially designed written report.



Data stored in: bmdata.mat
☑ Dilution on
Errors stored in: bmerr.mat
Maximum Error: 0.1
Report File: BMreport.srp
Output: last.bm

Figure 7: Dialogue box for calling our implementation of Beier and Mommsen's grouping procedure.

3. This dialogue box consists of only three basic user-interface elements: (a) text input via four editable boxes, (b) numeric input via a single editable box and (c), a single "check box" to define whether the dilution correction should be turned on or off. However, there are several other elements that can be used to construct such dialogue boxes. These include sliders, tabbed pages, push buttons, radio buttons, pop-up menus, scrollable lists etc. (MathSoft 1999).

4.2. The S programming language

As a brief example to show the power of the S programming language, we illustrate in figure 8 a simple S code listing for a routine to calculate the Mahalanobis distance between a point, $\underline{x}$, in p-dimensional space, and a group with mean $\underline{y}$ and p by p covariance matrix S. Recall that this is given by equation (1). Reconsidering equation (1), we can write

$$d_{xy}^{2} = \left(\underline{x} - \underline{y}\right)^{T} \Sigma^{-1} \left(\underline{x} - \underline{y}\right) = \left(\underline{x} - \underline{y}\right)^{T} \underline{v} \qquad (2)$$

where

$$\underline{v} = \Sigma^{-1} \left(\underline{x} - \underline{y}\right).$$

Mathematically, calculating explicitly the inverse of the p by p covariance matrix S is an ill-conditioned process and not to be recommended. However, multiplying both sides of the above equation by S gives

$$\Sigma \underline{v} = \underline{x} - \underline{y}\,.$$

In other words the vector $\underline{v}$ required in the definition of the Mahalanobis distance (2), is the solution of a linear system of p equations. This system is solved in the third line of the S routine illustrated in figure 8, using the *solve.default(Sigma, x-y)* command. The result is stored in the vector *v*. However note that in this simple example we have made no attempt to trap errors caused by the properties of the matrix stored in *Sigma*. In particular, this matrix will be rank deficient if the group size is less than *p*, the number of variables. S-Plus has many other built-in commands for carrying out complex mathematical operations on matrices. This makes it an ideal platform for the development and implementation of algorithms in the field of multivariate statistics where the data is naturally represented as an *n* by *p* data matrix.

The fourth line of the code calculates d_{xy}^{2} as defined by equation (2) and stores the result in the variable *d2*. Here the *%*%* operation represents matrix multiplication. The fifth line of the code signifies that it is the value of *d2* that is to be returned as the output from the routine. Given S-Plus variables *x, y* and *Sigma* storing respectively two *p* by 1 vectors, $\underline{x}$ and $\underline{y}$ and a *p* by *p* covariance matrix S, the command *d2* **<-** *maldist(x, Sigma, y)* calculates the Mahalanobis distance between $\underline{x}$ and $\underline{y}$ and stores the result in the variable *d2*.

```
1. Maldist <- function(x, Sigma, y)
2. {
3. v <- solve.default(Sigma, x-y)
4. d2 <- t(x-y) %*% v
5. return(d2)
6. }
```

Figure 8: S-Plus code listing for a routine to calculate the Mahalanobis distance between a point, $\underline{x}$, in p-dimensional space, and a group with mean $\underline{y}$ and p by p covariance matrix S. Note that line numbers are given only for ease of reference and are not required as part of an S-Plus code.

Finally it is worth noting that this example is chosen simply for illustration, as S-Plus has a built-in command, *mahalanobis*, that is called in a very similar manner to our example. However, as we have previously pointed out in section 2.2, Beier and Mommsen's grouping procedure makes use of modifications to the definition of Mahalanobis distance, given in equation (1), to take into account uncertainty of measurement and dilution effects. These modifications require additional programming effort, especially when we take dilution effects into account.

5. Summary, conclusions and future work

In this paper we have:

1. Shown that otherwise complex methodology can be implemented in S-Plus with relative ease. In particular, we have implemented a powerful but complex approach to the statistical analysis of ceramic compositions developed by Beier and Mommsen.
2. Demonstrated that traditional methods of exploratory multivariate analysis can be used alongside, or in combination with, a technique designed specifically for grouping ceramics by chemical composition.
3. Reported on the development of a user-friendly interface for our S-Plus routines.

Our implementation has been successfully applied to the analysis of chemical compositional data consisting of 27 elemental concentrations for each of 100 ceramic samples from Nichoria in the Peloponnese.

Looking ahead, our involvement in the GEOPRO project has the wider aim of developing an S-plus library of user-friendly routines for grouping ceramics using chemical and/or mineralogical data. In particular, over the duration of the project and beyond, our work will involve the evaluation, development and implementation of many more statistical techniques. A particular difficulty is that other authors have implemented advanced statistical methods for the analysis of mixed-mode data in the programming language FORTRAN. In the interests of widening access to such methodology, some progress has been made in interfacing FORTRAN-based software with S-plus. For similar reasons, we will be giving careful consideration to the provision of support materials (for example documentation and tutorials, both paper and web based). The final collection of routines will be packaged as an S-Plus library and made freely available, via the Internet, to the archaeometric community.

Acknowledgements

We are particularly grateful to Thomas Beier, Hans Mommsen and Anno Hein for fruitful discussions on this work. There is no implication that they necessarily agree with our interpretations and any misunderstandings and errors are entirely our responsibility. Hans Mommsen is thanked, additionally, for providing the data used throughout. This work forms part of the GEOPRO Research Network funded by the DGXII of the European Commission, under the TMR Network Programme (Contract Number ERBFMRX-CT98-0165).

References

BAXTER, M.J., 1994. *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh University Press.

BAXTER, M.J., PAPAGEORGIOU, I., CAU, M.A., DAY, P.M. and JACKSON, C.M., in press. Integrating geochemical and mineralogical data in studies of ceramic provenance: some statistical issues (the GEOPRO project). *Proceedings of CAA99*, Dublin. Oxford, BAR.

BAXTER, M.J. and JACKSON, C.M., in press. Variable selection in artefact compositional studies. *Archaeometry*.

BEARDAH, C.C., PORCINAI, S. and BAXTER, M.J., in press. Clustering with KDEs: art historical and archaeological applications. *Proceedings of CAA99*, Dublin. Oxford, BAR.

BEIER, T. and MOMMSEN, H., 1994. Modified Mahalanobis filters for grouping pottery by chemical composition, *Archaeometry* 36: 287-306.

MATHSOFT 1999. *S-Plus 2000 Programmer's Guide*. Seattle: MathSoft, Inc.

MOMMSEN, H., BEIER, T. and HEIN, A., in press. The Berkeley Neutron Activation Analysis data on Mycenaean pottery - A complete chemical grouping. In Kilikoglou, V. and Maniatis, Y. (eds.), *Proceedings of the 5th European Meeting on Ancient Ceramics*. Athens, 18-20 October 1999. Oxford, BAR.

VENABLES, W.N. and RIPLEY, B.D., 2000. *S Programming*. New York: Springer-Verlag.