Approaches to petrographic data analysis using S-Plus

C.C. Beardah, M.J. Baxter and I. Papageorgiou
Department of Mathematics
Statistics and Operational Research
The Nottingham Trent University
Clifton, Nottingham NG11 8NS, UK

M.A. Cau
Department of Archaeology and Prehistory
The University of Sheffield
Northgate House, West Street
Sheffield S1 4ET, UK

Abstract: In this paper we shall show how S-Plus can be used to address issues involved in clustering artefacts on the basis of petrographic thin-section data. If such data are coded as a set of categorical variables or as presence/absence data a variety of analytical options are open. These involve choosing how to measure similarity between cases and how, subsequently, to group these using clustering or scaling techniques. Using S-Plus functions that come with the package, or which are freely available in libraries developed by S-Plus users, options include multiple correspondence analysis and metric and non-metric scaling methods including Sammon mapping and isotonic multidimensional scaling. Some of these will be illustrated, and it will be shown how different methods can produce results that reveal different aspects of a data set. Our interest in pursuing this is that we would like to investigate the possibility of including both petrographic and geochemical data in a statistical analysis of artefact compositional data, and on an equal footing. We conclude by indicating how this might be achieved within the methodological framework discussed in the paper.

Key words: ceramics; petrographic thin-sections; multivariate analysis; cluster analysis; S-Plus.

Introduction

This paper presents the development of previously published work (Baxter et al. in press; Beardah and Baxter 2001) by members of the EU funded GEOPRO TMR Network. The general aims of the project are to integrate geochemical and mineralogical techniques in the study of raw materials and archaeological ceramic provenance. From a statistical point of view, a specific interest is the possibility of incorporating both geochemical and mineralogical data into statistical analyses. Using these two types of data in combination (so-called *mixed-mode* data), or by independently applying appropriate techniques to each, we hope to identify groups within a set of data. Such groups could be assumed to indicate, for example, distinct origins of the artefacts or raw materials used in their manufacture. Our ultimate aim is to produce user-friendly and accessible software, developed using the S-Plus package (Mathsoft 1999; Venables and Ripley 1999, 2000) and allowing "state of the art" analysis and presentation of geochemical, mineralogical and mixed-mode data. The reasons for choosing S-Plus as the basis for the development and distribution of our software are reported in Beardah and Baxter (2001).

In the same paper we showed how a clustering methodology for grouping chemical compositional data from artefacts, developed by Beier and Mommsen (1994), could be implemented in the S-Plus package. Geochemical data is typically continuous in nature. By contrast, as discussed in section 2 below, mineralogical data in its raw form can consist of variables of a mixture of types, including continuous, categorical and presence/absence. In this paper we concentrate solely on the implementation of statistical methods for mineralogical data.

More background on data arising from the analysis of petrographic thin-sections is given in the next section. Section 3 of the paper presents a brief discussion on techniques for grouping petrographic data. Such methods include multiple correspondence analysis and metric and non-metric scaling methods including Sammon mapping and isotonic multidimensional scaling. In section 4 we present an archaeometric case study. This is followed in section 5 by our conclusions.

Analysis and coding of ceramic thin-sections

Examples of ceramic thin-sections are shown in Figure 1. Usually a single analyst, on the basis of examination and comparison, carries out the classification of a collection of thin-sections into groups based upon provenance. This time-consuming approach is well known to rely upon the experience of the observer and the availability of reference groups and collections and furthermore it can be "heavily biased by the personal approach of the analyst involved" (Whitbread 1991) and lack reproducibility.

In attempting to develop an alternative, less "subjective" method of classification, the first challenge we face is how to describe a

collection of thin-sections in a manner that is amenable to statistical analysis. In order to do so, one approach is to represent each thin-section in terms of qualitative or categorical variables. Such variables can then be coded in a manner that can be input into a quantitative statistical analysis. Cau et al. (in preparation) review some reasons for, and attempts at, quantifying thin-section data, and describe the coding system used in our examples.

In order to describe the thin-sections, a system of 19 categorical variables covering the most common features normally recorded in ceramic petrology has been developed (Cau et al., in preparation). Variables 1-5 record aspects of the thin-section that can be used to reveal "technological information", for example firing temperature or forming techniques. The main rock types, organized by general families, and the main rock forming minerals that are found in Mediterranean ceramics are recorded in variables 6-18. Finally, variable 19 records packing, an estimate of the general amount of particles in a given sample.

It is important to note that in the system reported by Cau et al., a combination of categories within a specimen will result in the formation of a new category. As a result, the number of categories for each variable can be as many as 30, or as few as two (presence/absence). Furthermore, the system is open, that is new variables and categories can be added with relative ease.

Statistical methodologies

After description using the system described in section 2, a collection of n thin-sections can be represented by a table with n rows and 19 columns. Each row contains the description of a single thin-section; each column represents a variable; and each cell contains a number indicating the category. An important point is that *any* system similar to that of Cau et al., which results in an n by p table of categorical data, where p is the number of descriptive variables, can be treated using the statistical methods described in this section.

Converting categorical data into 0/1 data

In order to apply statistical methods, one approach is to convert the table of categorical data into a data matrix consisting of 0/1 entries. There are two ways of doing this, illustrated here by means of a simple example. Table 1 shows an example of a 4 by 1 table of categorical data. Here the single variable is divided into four categories, the last of which consists of the presence of both category 1 *and* 3. (Recall that in the system reported by Cau et al., a combination of categories within a specimen will result in the formation of a new category). For this particular data set, category 2 is not represented.

Coding method 1

Here we introduce one dummy variable for each category appearing within the data set, with a case being coded 1 for the variable whose category it belongs to and zero otherwise. Thus applying coding method 1 to the categorical data of Table 1 results in the 4 by 3 data matrix, denoted G, shown in the bottom-right of Table 2. Here the number of categories present in the variable is three (1, 3 and 4); no column appears for category 2 as this does not appear with these data. (Any zero columns would need to be omitted prior to the subsequent statistical analysis.)

Also, note that since our example has only one categorical variable, this coding method results in a single non-zero entry for each row of the matrix G. More generally, for an n by p table of categorical data, where variable i has L_i levels present in the table, we introduce L_i dummy variables corresponding to the levels of variable i. This gives rise to an n by L data matrix, G, where

$$L = \sum_{i=1}^{p} L_i$$

and the sum of the entries in each row is p.

Coding method 2

Using the descriptive system of Cau et al. we note that many categories are in fact combinations of those previously defined. For example, category 4 of our example consists of the combination of categories 1 and 3. An alternative method of coding the data is to only record the presence or absence of the "core categories" (1, 2 and 3 for our example). Thus applying coding method 2 to the categorical data of Table 1 results in the 4 by 2 data matrix shown in the bottom-right of Table 3. Here the number of core categories present in the variable is two (categories 1 and 3). Again, zero columns would need to be omitted prior to the subsequent statistical analysis and are therefore not included.

Although coding method 2 is conceptually simpler than method 1, it results in a more unwieldy data matrix, G. In particular, for our example the sum of the entries in each row is no longer 1, as was the case for coding method 1. More generally, for an n by p table of categorical data coding method 2 leads to an n by L data matrix, G, where the sum of the entries in each row varies and in each row is $>=p$.

Statistical methodologies

A variety of statistical methodologies can be used to investigate the data resulting, in each case, in a 'map' showing how similar cases are to each other.

- Multiple Correspondence Analysis (MCA) can be used if the data are coded using method one, and is simply correspondence analysis applied to the matrix G. This will have the effect of giving greater weight to rare categories.
- Principal component analysis (PCA) can be applied directly to G and, using coding method one, is equivalent to classical multidimensional scaling (MDS) applied to a dissimilarity matrix whose elements are the square-roots of the elements of the matrix $(P - GG^{T}/p)$ where P is an $(n \times n)$ matrix of 1s and ^{T} indicates a matrix transpose.
- Non-metric MDS methods, including Sammon's

mapping and Kruskal's isotonic MDS (Venables and Ripley 1999: 333-335) can be applied to the dissimilarity matrix described above.

Correspondence Analysis of the 0/1 matrix *G* can be applied to data derived from coding method two, and any of the MDS methods can be applied to a suitable dissimilarity matrix derived from *G*.

All of these methods result in graphical output similar in nature to that of a PCA. The number of components used to display the results can be varied.

Case study

The methodologies introduced in the previous section are now illustrated using mineralogical data arising from a collection of Late Roman Cooking Ware (LRCW) from the Balearic Islands and the eastern Iberian peninsula. In examining these data our aims included the determination of origin, in other words, were the wares local or imported? The full data-set contains 115 samples of LRCW from different sites in the region. (For these data we have both chemical and petrographic information.) A subset of these data contains 25 samples of LRCW from Can Sora (Eivissa). These samples come from a cistern, used as a rubbish dump during Late Antiquity. Samples came from two different layers (from the 5th and 6th centuries A.D.)

Data format and conversion

Figure 2(a) shows a part of the 25 by 19 table of categorical data for these data as it appears in S-Plus. The second column, whose entry in row number 1 is "CS-2 (pl)", contains labels for each case. Such labels can be extremely useful when analysing (in S-Plus) graphical output of the type generated by the methods of section 3. Symbols in brackets denote the result of a preliminary classification carried out by one of us (MAC) and based solely upon a petrographic analysis of the kind discussed in section 2. This analysis led to the identification of five types, labelled *pl*, *v*, *m*, *p* and *f*, and five outliers labelled *o* (Cau et al., in preparation).

Conversion between categorical representation and 0/1 representation can be achieved via our **catdist.mat** routine. This routine can be called from the GEOPRO menu (figure 3(a)) and results in the dialog box shown in figure 3(b).

This returns a list, **cansora.GX**, containing the matrices *G* and *X* (see section 3). To extract *G* and store it in a separate, appropriately named variable we can use the S-Plus command

cansora.G <- cansora.GX[["G"]]

Figure 2(b) shows part of the matrix (*G*) of 0/1 entries so obtained, again as it appears in S-Plus. Coding method 1 has been used (see section 3.1.1). For illustration, note that variable 1 (optical activity) has three categories (1. active, 2. inactive and 3. intermediate). Only the final case in the data set, labelled CS-27, falls into category 2 for this variable. As a result, the matrix *G* has three dummy variables, labelled **opt.act.1**, **opt.act.2** and **opt.act.3**, corresponding to the categorical variable describing optical activity. Those cases, for example CS-2 and CS-3, falling into category 1 for optical activity thus have a 1 in the column of G labelled **opt.act.1** and zero entries in those columns of *G* labelled **opt.act.2** and **opt.act.3**.

Of course, the matrix *G* may have already been coded (using method 1 or method 2) and entered directly. This possibility is allowed for in our implementation of the various statistical methods for the analysis of petrographic data. In either case we must be wary of columns whose sum is zero. Such columns are automatically detected and removed upon submission to the routines for performing MCA, MDS etc.

Analysis of petrographic data

Using a combination of existing S-Plus library routines (Venables and Ripley 1999) and new implementations of existing techniques, adapted to make the most of the graphical capabilities of the package, we have made the following methods available.

1. Multiple Correspondence Analysis (MCA).
2. Classical Multi-Dimensional Scaling (MDS).
3. Non-metric MDS methods: Sammon's mapping and Kruskal's isotonic MDS.
4. Correspondence Analysis of the 0/1 matrix *G*.

We initially use the Can Sora data set for illustration (recall that this is a subset of the full LRCW data set).

In order to make the most of the interactivity available within S-plus, the application of these methods is a two-stage process:

1. Application of the method (**MCA/MDS Scores** menu item, figure 4(a));
2. Plotting the results graphically (**MCA/MDS Plots** menu item, figure 4(b)).

The calculation stage results in the creation of a variable named according to the method used, for example, if we use Multiple Correspondence Analysis, then the variable is called **last.MCA**. It is important to note that the output from MCA *must* be stored in this variable.

Different methods result in different numbers of component scores that can be plotted. Table 4 shows the maximum number of these currently available for each method.

Highlighting subgroups

Figure 5 shows the graphical output upon application of MCA to the Can Sora data set. Graphical output based upon three components has been displayed. The window on the left shows a display based upon the first three components, while that on the right shows all possible plots based upon two of the first three components. The subgroup classified petrographically and labelled *v* has been highlighted. This is achieved by clicking rows in the **last.MCA** data window. (The centre of the three windows shown in Figure 5; the **ctrl** and **shift** keys can be used to highlight individual rows or sets of consecutive rows.) Corresponding points in the component plots are automatically

highlighted. It is clear from the left window of Figure 5 that this subgroup also separates on the basis of a plot of the first three components of the MCA. If this were not immediately apparent, the three-dimensional plot could be rotated and examined from different viewpoints. Furthermore, plots based upon components 1 and 2, and 2 and 3 also exhibit clear separation of this subgroup. Based upon the plot of components 1 and 3, separation is less clear. Despite this, there is overwhelming evidence that this subgroup would be identified by this statistical technique (MCA) independently of any knowledge of the prior petrographic analysis.

Of course, we may not have a petrographic classification available. In these circumstances we will be attempting to group the data by using purely statistical techniques, possibly including methods (such as PCA) for the analysis of geochemical data, if available. S-plus can be used to identify and highlight subgroups in many different ways. Figure 4(b) shows a simple example. Here we have used the "Highlight rows" box to identify objects where the second component of the MCA output is negative. However, more complicated expressions can be used to highlight rows. In fact we can enter any valid S-plus expression in the "Highlight rows" box. For example, the subgroup shown in figure 5 could be defined by noting that these are the cases where the first and second MCA components are negative. Hence we could use the S-plus expression

last.MCA[,1]<0 & last.MCA[,2]<0

in the "Highlight rows" box to get the same effect as that shown in figure 5. Furthermore, the expression used could depend upon the analysis using a different method entirely. So we might choose to highlight (in the MCA output) the rows where the second *MDS* component is positive. This might be useful way of comparing the subgroups obtained using different methods, especially if the dataset is large. Simple exploratory techniques, for example histograms, can be used to look at the distribution of individual components in an effort to identify criteria for separating groups in this way. A slight twist on the traditional histogram, a labelled histogram, is provided via the GEOPRO menu. An example is shown in figure 6. This shows that the Can Sora data set separates clearly into two groups on the basis of whether the second MCA component is positive or negative.

More usually, and perhaps especially in analyses involving smaller datasets, subgroups may be identified visually, directly from the graphical output. By positioning the mouse pointer over a plotted point, the row number (or label) of the point and its component values are shown (see the left window in figure 5; here the mouse was positioned over the rightmost point of the subgroup, representing the case labelled CS-17). For example, as previously discussed, it is likely that the subgroup highlighted in figure 5 would have been identified in the absence of a prior petrographical classification, by visual observation of the MCA output alone. (Since it separates nicely in plots of both the first two, and first three, components.) By using the mouse, we can easily identify that this subgroup consists of rows 8, 9, 10, 13, 14, and 15 of the dataset (see the centre window of figure 5).

Finally, and perhaps most conveniently, subgroups can be highlighted directly in windows displaying graphical output by means of the **Select Data Points** feature of the **Graph Tools** menu within S-Plus. Clicking and dragging the mouse can be used to highlight sets of points in the graphics window. (The **ctrl** key can be used to highlight subgroups consisting of several individual clusters.) Selected points are automatically highlighted in the data window also (as illustrated in figure 5).

It would be useful to investigate whether the subgroup previously identified separates just as easily when we use other methods. Returning to the **MCA/MDS Plots** dialog we can now generate the graphical output resulting from, for example, Kruskal's non-metric MDS (**IsoMDS**). The dialog box is shown in figure 7. Note that the entry in the "Highlight rows" box is generated automatically, since these rows were previously highlighted in a data window. The graphical output so generated reveals that the aforementioned subgroup separates just as easily with this method. In fact, all five methods discussed here identify this subgroup with little difficulty.

Outlier removal

It will often be the case that examination of the graphical output from one or more methods will reveal the presence of outliers. Removal of outliers can help to further separate other genuine subgroups within the data set. For example, figure 8 reveals that, based upon the output from MCA, it is fairly clear that rows 6 and 11 (cases CS-7 and CS-13) could be considered to be outliers. This observation is supported, to a greater or lesser extent, by the other methods. At this point it is worth recalling that, for the Can Sora dataset, five cases were identified petrographically as outliers. Rows 6 and 11 were two of these. The others were rows 7, 22 and 23 (cases CS-8, CS-24 and CS-25). There is fairly strong evidence, from methods other than MCA, that row 23 is an outlier, however the situation is less clear, for rows 7 and 22, which seem to associate quite well with the subgroup labelled *pl* on the basis of petrography. This is an example of where the application of statistical techniques possibly results in a different conclusion to the application of traditional petrographic analysis.

Outliers can be removed at the first (calculation) stage of the application of our methods. Figure 9 shows the application of MCA to the Can Sora dataset minus rows 6, 11 and 23.

Combining petrographic and geochemical analyses

For the full LRCW data set (115 cases) we have both geochemical and petrographic information. The former consists of 25 concentration values and the latter has already been coded as a matrix (G) of 0/1 entries using coding method 1. We can calculate and plot PCA scores in much the same way as previously discussed. Figure 10 shows the dialog box called from the GEOPRO menu entry **PCA Scores**.

The initial plots of the component scores are quite cluttered and would possibly benefit from the removal of some outliers. The most obvious outliers are cases MC-16 and MC-19 (rows 33 and 34 respectively; these cases were also identified as outliers on the basis of the petrographic analysis). Re-calculating the PCA scores with these cases omitted gives a slightly clearer

plot. To do this we would use the same dialog as in figure 8, but with **c(33,34)** entered in the **Omit these cases** box. There is now evidence of an outlying group consisting of cases U-1 and U-10 (rows 60 and 69 of this reduced dataset). Care needs to be taken here, as cases U-1 and U-10 correspond to rows 62 and 71 of the original dataset (with no outliers removed). This shows the value of assigning labels to cases! Re-calculating the PCA scores with rows 33, 34, 62 and 71 omitted gives a plot that is rather more spread out and that could be used to identify more subgroups on the basis of chemistry.

Of course we can also apply our various methods for grouping petrographic data to the LRCW dataset with rows 33, 34, 62 and 71 omitted. Analysing the two-dimensional output from MCA reveals a potential subgroup of four cases (CS-18, 19, 20 and 25) in the bottom-centre of the plot of components 2 and 3. Interestingly, on the basis of the three-dimensional output, we could probably discard CS-25 from this potential subgroup. CS-18, 19 and 20 has also been identified as a subgroup on the basis of petrography. (CS-25 was identified as an outlier on the basis of the petrographic analysis.) However, highlighting these cases in the plot of the PCA scores (with rows 33, 34, 62 and 71 omitted) reveals that based upon the analysis so far, these cases do not separate on the basis of chemistry.

Mixed-mode data

In the previous section we have seen that one way of incorporating both petrographic and geochemical information in our statistical analysis is to analyse the chemistry and petrography separately, but possibly concurrently, using methods appropriate for each. Using this approach, we can investigate whether subgroups identified with the various methods for analyzing petrographic data are also identified using exploratory methods for analyzing geochemical data. Such methods available within S-Plus include, for example, PCA, cluster analysis, and other, less traditional methods such as that proposed by Beier and Mommsen (1994) and discussed in Beardah and Baxter (2001).

However, within the framework outlined so far, we can incorporate both chemical and petrographic data. This could be achieved, for example, by first converting the elemental concentrations of the chemical data into categorical data before applying the techniques outlined here. Alternatively, some methods can deal directly with data of mixed type (e.g. continuous chemical data and 0/1 petrographic data). To do so, we need to measure dissimilarity between objects of mixed type (Gower and Hand 1996; Kaufman and Rousseeuw 1990). Such approaches will be the focus of future work.

Summary and conclusions

In this paper we have seen that

- Otherwise complex methodology for the analysis of petrographic data can be implemented in S-Plus with relative ease. In addition, many techniques (e.g. the non-metric methods) have been implemented in S-Plus and made freely available, for example as part of the MASS library (Venables and Ripley 1999).
- The ability to manipulate the user interface makes it possible to design user-friendly routines.
- The nature of the S-plus interface makes it possible to (a) easily identify sub-groups within the data and (b) compare the results when using different methods independently, including those appropriate for the analysis of geochemical data.
- Some of the methods reported here can be used to analyse data of mixed type.

Our implementation has been illustrated using data arising from the petrographic and chemical analysis of 115 specimens of Late Roman Cooking Ware (LRCW) from the Balearic Islands and the eastern Iberian peninsula. Whilst not fully illustrated in this paper, the general approach to the data analysis is an iterative process consisting of identifying obvious groups and outliers, "peeling" these away from the data set, and proceeding with an examination of what remains. See Cau et al. (in preparation) and Papageorgiou et al. (2001) for more detail on this approach.

Finally, as stated at the outset, our involvement in the GEOPRO project has the aim of developing an S-plus library of user-friendly routines for grouping ceramics using chemical and/or mineralogical data. In support of the software, materials such as documentation and tutorials, both paper and web based, will be provided. The final collection of routines will be made freely available, via the Internet, to the archaeometric community.

Acknowledgements

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., I. Papageorgiou, M.A. Cau, P.M. Day and C.M. Jackson. In press. Integrating geochemical and mineralogical data in studies of ceramic provenance: some statistical issues (the GEOPRO project). Proceedings of CAA99, Dublin. Oxford, BAR.

Beardah, C.C. and M.J. Baxter. 2001. Grouping ceramic compositional data: an S-plus implementation. Proceedings of CAA2000, Ljubljana. Oxford, BAR.

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

Cau, M.A., I. Papageorgiou and M.J. Baxter. In preparation. Exploring automatic grouping procedures in ceramic petrology.

Gower, J.C. and D.J. Hand. 1996. *Biplots*. London: Chapman and Hall.

Kaufman, L. and P.J. Rousseeuw. 1990. *Finding Groups in Data*. New York: John Wiley.

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

Papageorgiou, I., M.J. Baxter and M.A. Cau. 2001. Model-based cluster analysis of artefact compositional data. (Submitted for publication.)

Venables, W.N. and B.D. Ripley. 1999. *Modern Applied Statistics using S-Plus.* New York: Springer-Verlag.

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

Whitbread, I.K. 1991. Image and data processing in ceramic petrology, in A. Middleton & I. Freestone (Eds.), *Recent Developments in Ceramic Petrology*, British Museum, Occasional Paper no. 81, pp. 369-388.

Figures

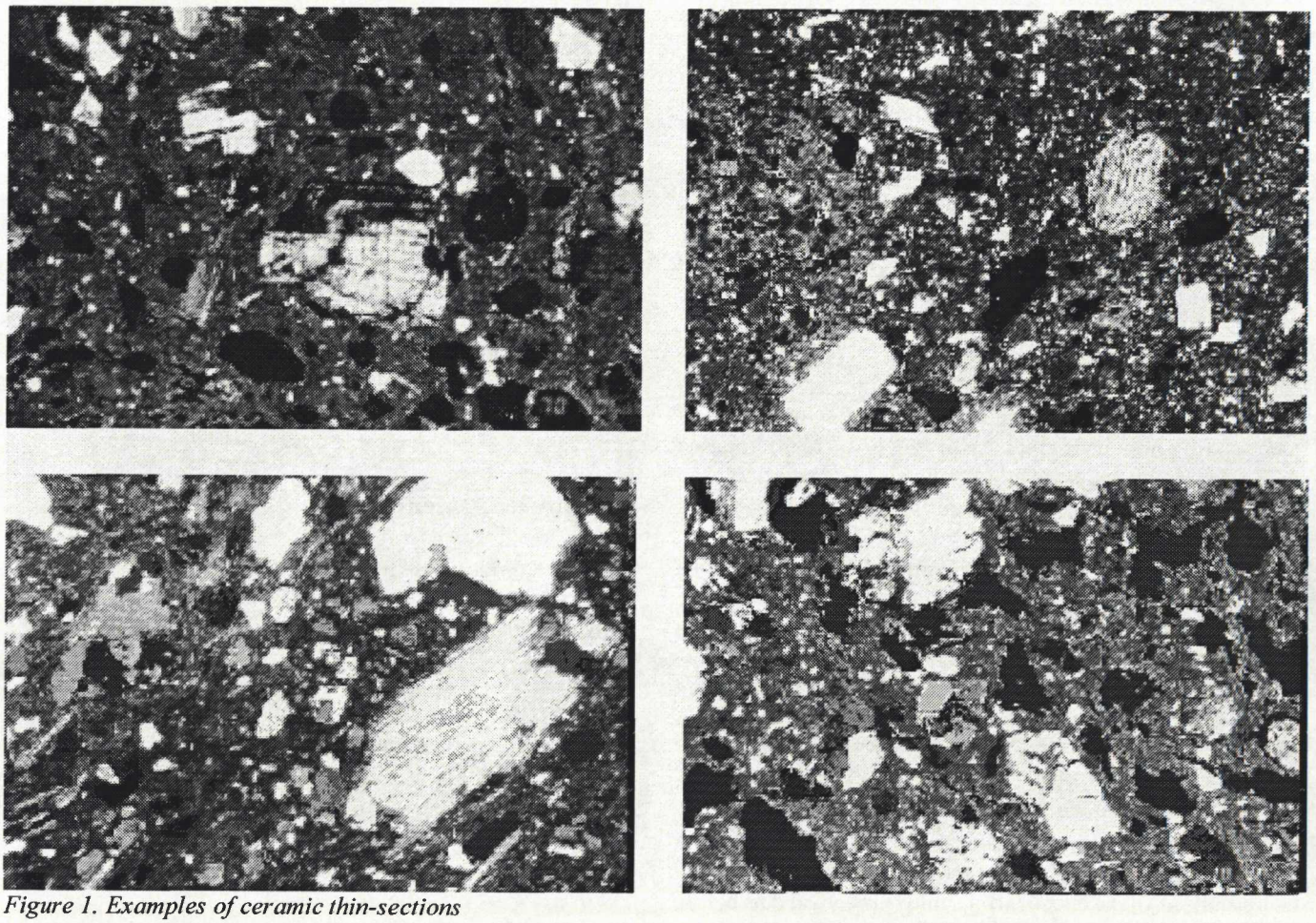

Figure 1. Examples of ceramic thin-sections

		1	2	3	4	5	6	7	8
		opt.act	min.orient	Voids.ori.	texture	special.comp	plut.rocks	volc.rocks	metam.rocks
1	CS-2 (pl)	3	3	1	4	11	3	1	1
2	CS-3 (pl)	1	1	1	4	11	3	1	1
3	CS-4 (pl)	3	1	1	5	9	3	1	1
4	CS-5 (pl)	1	2	1	4	11	3	1	10
5	CS-6 (pl)	3	2	1	4	11	3	1	1
6	CS-7 (o)	3	1	1	3	9	2	1	4
7	CS-8 (o)	3	2	1	5	9	3	1	1
8	CS-9 (y)	3	1	1	4	11	1	2	1
9	CS-10 (y)	1	1	1	4	11	1	2	1
10	CS-11 (y)	1	1	1	4	11	1	2	1

		1	2	3	4	5	6	7	8
		opt.act.1	opt.act.2	opt.act.3	min.orient.1	min.orient.2	1	texture.3	texture.4
1	CS-2 (pl)	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
2	CS-3 (pl)	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
3	CS-4 (pl)	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00
4	CS-5 (pl)	1.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00
5	CS-6 (pl)	0.00	0.00	1.00	0.00	1.00	1.00	0.00	1.00
6	CS-7 (o)	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
7	CS-8 (o)	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00
8	CS-9 (y)	0.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00
9	CS-10 (y)	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
10	CS-11 (y)	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00

Figure 2. The Can Sora data-set expressed as (a) a table of categorical data, (b) a matrix (G) of 0/1 entries obtained using coding method 1. Symbols in brackets denote the result of a preliminary petrographic classification

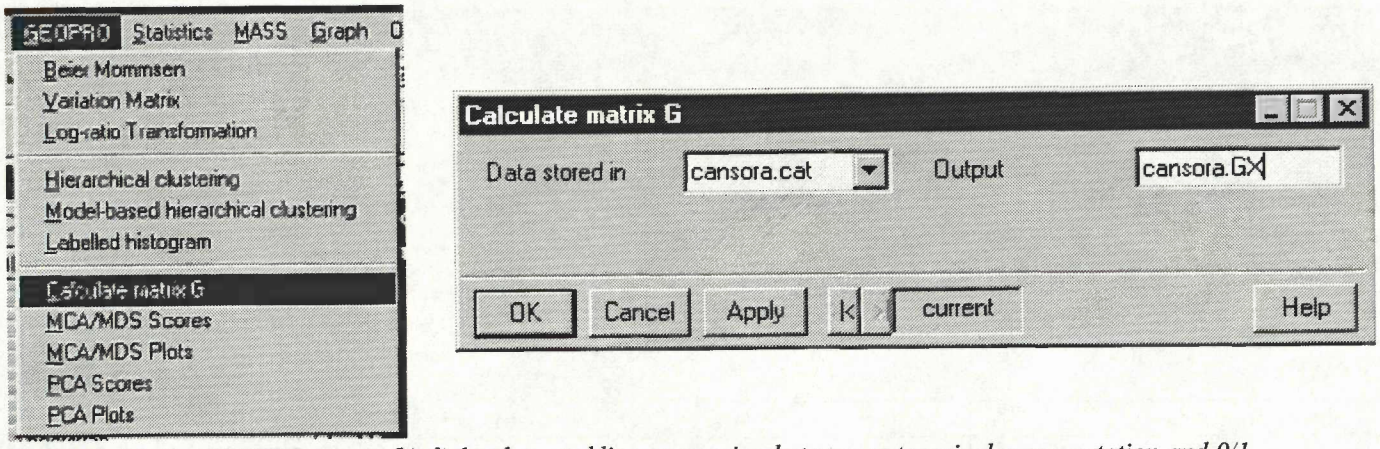

*Figure 3. (a) the GEOPRO menu, (b) dialog box enabling conversion between categorical representation and 0/1 representation (**catdist.mat** routine).*

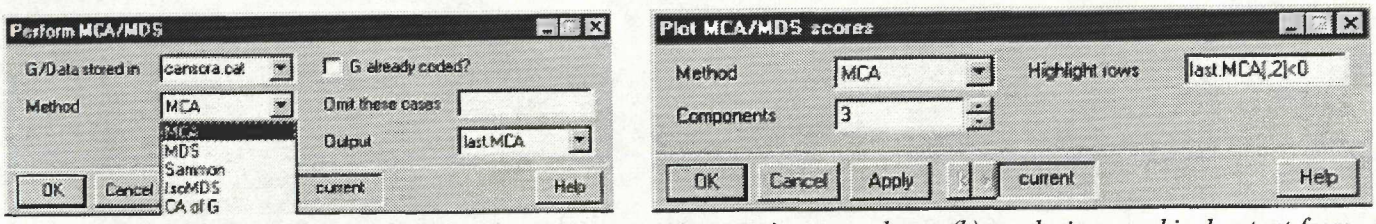

Figure 4. Dialog boxes for (a) application of various automatic grouping procedures, (b) producing graphical output from the calculation in (a).

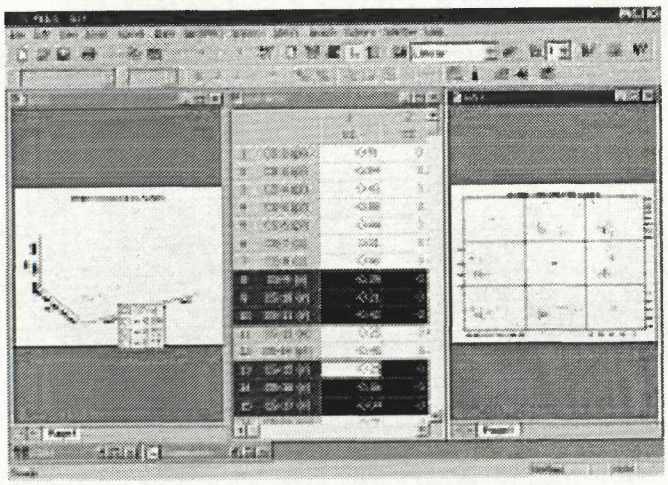

Figure 5. Application of MCA to the Can Sora data set.

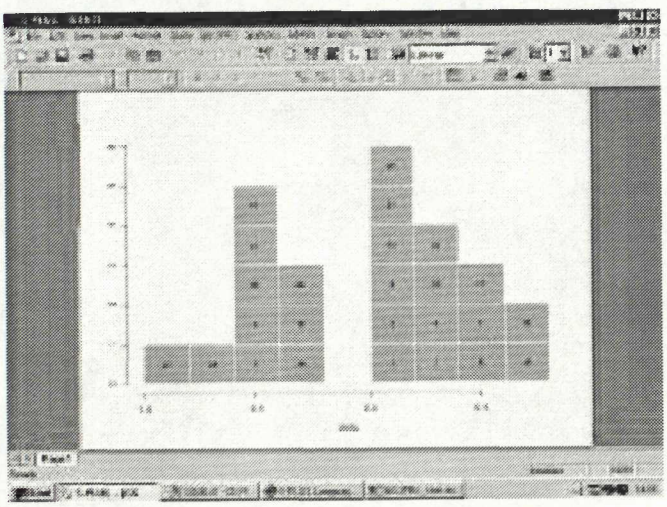

Figure 6. A labelled histogram (second component of MCA output for Can Sora dataset).

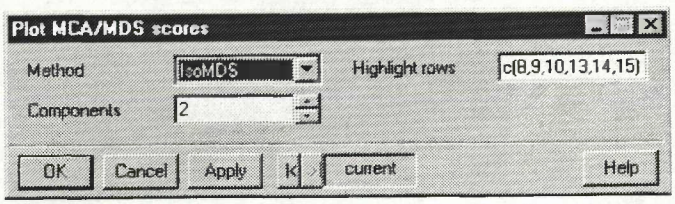

Figure 7. Generating the graphical output resulting from Kruskal's non-metric MDS. Rows 8, 9, 10, 13, 14 and 15 will be highlighted (this was the subgroup identified earlier using MCA, see figure 5).

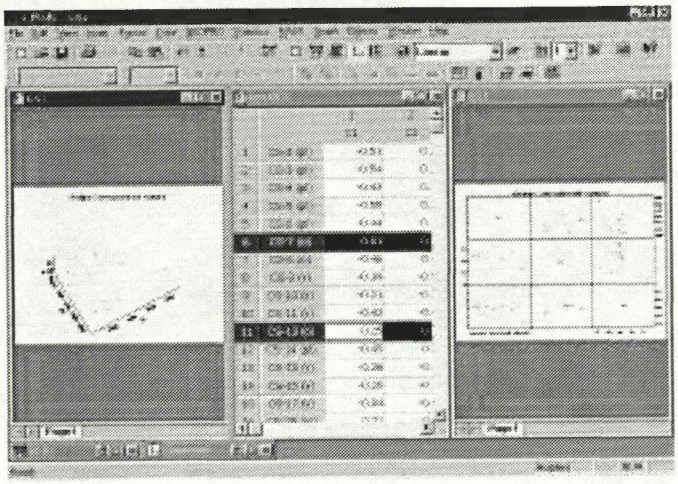

Figure 8. MCA output revealing that rows 6 and 11 (cases CS-7 and CS-13) of the Can Sora data set are fairly clear outliers.

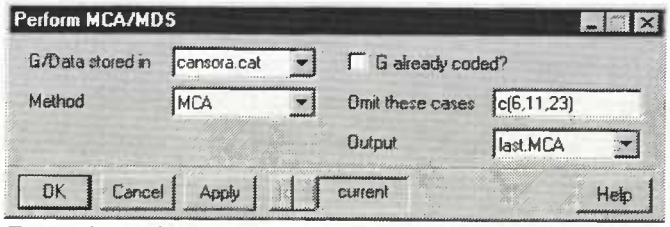

Figure 9. Applying MCA to the Can Sora dataset minus rows 6, 11 and 23 (cases CS-8, CS-24 and CS-25).

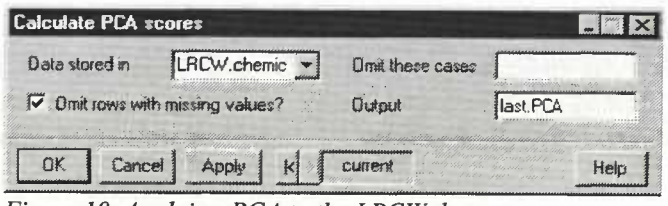

Figure 10. Applying PCA to the LRCW dataset.

Tables

Object number	Category
1	1
2	4 (= 1 + 3)
3	3
4	1

Table 1. A simple example of a 4 by 1 table of categorical data. The single variable is divided into categories 1, 2, 3 and 4, the last of which consists of the presence of both category 1 and 3. Also note that category 2 is missing in this particular data set.

Object number	**1**	**3**	**4 = 1 + 3**
1	1	0	0
2	0	0	1
3	0	1	0
4	1	0	0

Table 2. Coding method 1.

Object number	**1**	**3**
1	1	0
2	1	1
3	0	1
4	1	0

Table 3. Coding method 2.

Method	Acronym	Variable name	Components (maximum)
Multiple Correspondence Analysis	MCA	last.MCA	4
Multi-Dimensional Scaling	MDS	last.MDS	4
Sammon's Non-linear Mapping	Sammon	last.SAM	2
Kruskal's Non-metric MDS	IsoMDS	last.ISO	2
Correspondence Analysis of G	CA of G	last.CAG	4

Table 4. Summary of available methods.