Archeologia e Calcolatori
21, 2010, 211-228

CORRESPONDENCE ANALYSIS IN R FOR ARCHAEOLOGISTS: AN EDUCATIONAL ACCOUNT

1. Introduction

We have previously published papers that involve the use of a statistical technique, Correspondence Analysis (CA), for comparing assemblages of finds across different sites. The positive response from archaeological colleagues with similar concerns to those we addressed has been encouraging, but it is apparent that many of these colleagues – particularly those not located within a university (and therefore without access to effective but costly statistical software and easy access to expert statistical advice) – have no problems in understanding how CA works, but do have problems implementing it. The main purpose of this article is to try to indicate how CA can be implemented in a software package, R, by an archaeologist prepared to invest some effort in exploring the technique.

R is an open source (that is to say, free) statistical package. It has been developed by and for experienced statisticians so is not necessarily easy to use by those without some statistical training or guidance. We aim to provide the latter.

A brief and non-technical account of CA is given in the next section, but the ideal reader of this paper will already know about CA and will be more interested in how to use it. Expositions of CA written for an archaeological readership include Shennan (1997, 308-341), Baxter (1994, 100-139) and Baxter (2003, 136-146). Greenacre (1984) provides a thorough mathematical account. Many articles on CA applications to archaeological case studies have also been published in this Journal (for a theoretical account see lastly Djindjian 2009).

The next section includes a brief description of the aims of CA and the way its use has developed in archaeology. Following this, some information about the software, R, is given. The remainder of the paper illustrates how R can be used for CA, using archaeological data, and concludes with advice on some practical issues of implementation.

We have trialled the instructions using a PC running Windows XP Pro with a 2.2 Mb Broadband connection.

2. Correspondence Analysis

At its simplest, CA can be viewed as a statistical technique for visualising a table of non-negative numbers. As a concrete example, suppose information has been collected from *r* sites (or contexts) and a count has been made of the numbers of each of *c* artefact types (or more generally finds) present

within the context. The results can be collected in tabular form, where each of the *r* rows corresponds to a context, and each of *c* columns corresponds to an artefact type or find.

A natural question that arises with this kind of data is to ask how similar contexts are in terms of the profile of finds within them. It is also of interest to ask how similar the profile is of find types across sites.

Essentially CA reduces a table of data to two maps (or plots). In the first map, points on the plot correspond to the rows of the table (i.e. the contexts). Points on the plot that are close to each other can be identified with contexts that have a similar profile in terms of their finds assemblage; points which are very distant correspond to contexts which have very different assemblage profiles.

In the second map, points correspond to the columns of the table (i.e. the finds type), and points which are close together identify finds which have a similar distribution across sites. The two maps can be "superimposed" and viewed together they allow the similarities and differences between contexts to be assessed. Examples will follow.

The way CA has developed in archaeology is quite curious and the story is described in some detail, up to 1992, in Baxter (1994, 133-139). The use of CA for the purpose of seriation (not something discussed much further in this paper) is now common. An early application, appearing in a statistical journal and having little immediate influence on archaeological practice, was Hill (1974). The «World Archaeology» paper by Bølviken *et al.* (1982) is often credited in the archaeological literature for introducing CA to archaeology, but this ignores contributions in the French-language literature dating from the mid-1970s, much of it associated with the work of François Djindjian (see Baxter 1994, 134 for details).

Given that CA is an obviously useful method, its diffusion into certain sections of the archaeological community was painfully slow. Orton (1999) identified CA as the most important statistical technique introduced into archaeology in the 1980s. Baxter (1994, 134) credits Ringrose (1988) as being the first fully fledged British application of CA that post-dated Bølviken *et al.* (1982), and Ringrose was a statistician, so regular British use of CA by archaeologists only really dates from about the early 1990s (Baxter disqualified a slightly earlier British contribution on the grounds that it was assisted by an Australian).

Given the contribution of North American scholarship to the development of quantitative methodology in archaeology, the length of time it took for CA to penetrate the North American literature borders on the astonishing. Cowgill (2001) has said that CA was «virtually unheard of» in the US until the late 1980s; Baxter (1994, 135) was unable to identify much in the way of American usage of CA up to about 1992; and Duff (1996, 90), at a comparatively late date, was able to write that CA «was not well established in Americanist literature».

The situation has changed now (to some extent), but why the comparative neglect? For a start, many archaeologists, even if they have been exposed to some training in quantitative methodology, are averse to statistics and – even when they acknowledge the potential usefulness of statistical methods – lack the confidence to use them. Collaboration between archaeologists and statisticians is the obvious solution to this problem, but a lot of archaeologists do not have ready access to a usable statistician.

3. Simple Correspondence Analysis in R

3.1 *Getting the package*

The "Comprehensive R Archive Network" (CRAN) at http://cran.r-project.org/ provides a great deal of information on R. The information summarised below was current in August 2010, but as R is constantly updated details will change. Here version 2.11.1 is used, and it assumed that this is to be installed on a Windows platform. The route

Windows → base → Download R.11.1 for Windows

provides access to the package. Downloading the file places the application R-2.6.1-Win 32 in the folder of your choice. This can be installed by clicking on it and running the Window Installation Wizard as normal. Accepting defaults will load R in C:/Program Files/R/R-2.9.1 and create a desktop icon (R 2.9.1) that can be used to launch the package. The file to be downloaded is c. 33 Mb and the complete download and installation should take only a minute or so with a Broadband connection. Please note that depending on how you have the firewalls on your computer set up, you may have to override them sometimes, for example when installing the library packages discussed below.

The package operates from typed lines of command entered after the > prompt. It is case-sensitive so it is important to use upper case letters where shown. When you start using the package you are likely to encounter error messages because of mistakes in your typing. If you see «Error: syntax error in …», it will often be because of quite simple mistakes, such as either omitting a space or introducing one accidentally. An error message «object ….. not found» often means the double quotation marks have been omitted.

Plots open in separate windows so the window where you type the commands (R console) is best kept in a slightly minimised mode so that you can see them.

3.2 *Data entry*

This can be done in more than one way. For illustration, data from Table 1 of Cool and Baxter (1999) are used. For 18 sites the amounts of six glass

vessel types are quantified by estimated vessel equivalents (EVEs). The sites differ in date and type. The data are given in Table 1 in the Appendix.

If the data is set up in EXCEL (including variable names) highlight the data you wish to use; within EXCEL use

Edit → Copy

and then, within R, create a data file JRA1 using

```
> JRA1 <- read.table(file = "clipboard", header = T)
```

and type

```
> JRA1
```

to see the data[1].

An alternative method of data entry is to create a plain text file (.txt) with a plain-text editor (Windows NotePad in this instance, which can be found in Accessories in the Windows Start menu). Note that column (variable) headings are provided. For illustration purposes the file is named `JRA1` and placed in a folder called `CAinArch` in My Documents. The path for this will be similar to

```
"C:/Documents and Settings/Your Name/My Documents/CAin-
Arch/"
```

where the `Your Name` element relates to whatever name your system is set up under. The file can be read in after the R prompt, >, using the `read.table` function. This contains two components within the brackets; the first specifies the path to the file and the second, `header = T`, indicates that variable names are to be expected in the first line of the file.

```
> JRA1 <- read.table("C:/Documents and Settings/Your
Name/My Documents/CAinArch/JRA1.txt", header = T)
```

Note that the path is enclosed in double quotation marks and uses forward slashes (/) (this is essentially what was done to import the file from EXCEL, where the copy command within EXCEL placed the data on the `"clipboard"`).

An error message will occur if there are problems. If there are none type

```
> JRA1
```

to see the data. For full help on the function type

```
> ?read.table
```

[1] The writers of the `R` manual for data import/export prefer you to write the EXCEL file to a Tab or comma-separated file and use `read.delim` or `read.csv` (see below).

These notes mainly provide information on what is needed to get started. It is worth getting into the habit of using the help facility [`help(read.table)` is an alternative to `?read.table`] to see what else is available. For example, `read.csv` and `read.delim` can be used with comma separated variables and Tab delimited variables respectively.

3.3 *Packages in R*

By default R comes with a "base" statistics package, but to make the most of it you need to be able to access packages of functions that are either bundled with R or contributed by users and accessible from R.

In the first instance we shall use the bundled `MASS` package, associated with the book *Modern Applied Statistics with S* by Venables and Ripley (2002). Load this either by typing

```
> library(MASS)
```

(R is case-sensitive so it is important to use capital letters) or, from the menu

Packages → Load Package → MASS

This latter route will show you other available packages. To get help on the library function use `?library`; to get help on a particular package, `MASS` for example, `> library(help = MASS)` will provide basic information. Access to non-bundled packages is discussed later.

3.4 *Simple Correspondence Analysis*

At this stage a data set has been created, and the `MASS` package loaded. For simple CA the function `corresp` is available, and `?corresp` provides help on this.

Using

```
> JRAca1 <- corresp(JRA1, nf = 2)
Warning message:negative or non-integer entries in table in: corresp.matrix(as.matrix(x), ...)

> biplot(JRAca1)
```

gives Fig. 1, which may be compared with Fig. 3 in Cool and Baxter (1999).

The analysis can be done in a single line using

```
> biplot(corresp(JRA1, nf = 2))
```

The warning message can be ignored as it is simply alerting us to the fact that some of the data is non-integer, without stopping calculations. In the present context there is no problem with non-integer numbers, but some

software will not allow this and requires data manipulation (e.g., multiplication by some power of 10) before analysis can proceed.

To save a figure, from within R use

File → Save as

and select from the file formats available.

Numbers in the figure label the rows of the data set from 1 to 18, and column names are also given. Before interpreting the results, note that we have not used any information about site date. To do this, create a new variable, `date`, as follows

```
> date <- c(1,1,1,1,1,1,1,1,2,2,4,4,4,4,4,4,4,4)
```

where 1 = sites of the 1st/2nd century AD, 2 = sites of the 2nd/3rd century, and 4 = sites of the 4th century, and use

```
> biplot(JRAca1, xlabs = date)
```

which gives Fig. 2. The addition of `xlabs = date` results in points corresponding to rows being labelled by the numbers in `date`. The plot shows quite nicely that later assemblages have a composition distinct from earlier assemblages, and that this is largely attributable to the relatively higher proportion of cups present in later assemblages. Note that there is no need to re-do the CA, since the results from this are held in the "objectQ" `JRAca1` previously created.

The appearance of the plot is not complicated here, and the message is quite clear. For larger data sets and/or longer labels for the variables, plots like those of Figs. 1 and 2 can become overcrowded and difficult to read, and we often prefer to present plots for rows and columns separately. This is now done, where the opportunity is also taken to simplify labelling of the types. The latter is not really necessary here but, for illustration, can be done using

```
> type = c("C", "Bw", "Ja", "F", "Ju", "Bt")
```

Note that, because the labels are names rather than numbers, they have to be enclosed in double quotation marks.

To get the row plot use

```
> biplot(JRAca1, xlabs = date, ylabs = rep(" ", 6))
```

where `ylabs = rep(" ", 6)` makes 6 copies of a blank label.

For the column plot use

```
> biplot(JRAca1, xlabs = rep(" ", 18), ylabs = type)
```

As given, two separate plots are produced. If they are sandwiched between the directives

```
> par(mfrow = c(1,2))
```

and

```
> par(mfrow = c(1,1))
```

Fig. 3 results.

The commands would thus look like this on the worksheet

```
> par(mfrow = c(1,2))
> biplot(JRAca1, xlabs = date, ylabs = rep(" ", 6))
> biplot(JRAca1, xlabs = rep("", 18), ylabs = type)
> par(mfrow = c(1,1))
```

The first `par()` usage produces a plotting region of 1 "row" and two "columns", and the second usage restores things to normal. Such multiple plots are not always satisfactory and some manipulation of plot parameters (beyond the scope of the present article) may be needed before getting aesthetically pleasing and informative results.

3.5 *Other packages*

A large number of user-written packages are available for R. To access those not automatically bundled with R, go to Packages on the tool bar and select a download site after

Packages → Set CRAN mirror

then

Packages → Install package(s)

and select the package of choice. Here we select `ade4`.

Once the package is downloaded type

```
library(ade4)
```

to access the functions within it. The quality of documentation and transparency of use for different packages is variable. Some come with extensive and helpful documentation; others with little at all. Apart from information available via CRAN judicious use of `Google` is often very helpful.

The sequence

```
> JRAca2 <- dudi.coa(JRA1, scannf = FALSE)
> scatter(JRAca2)
```

produces Fig. 4, which may be compared with Fig. 1. The shaded bars in the plot to the top left shows the relative importance of the first two CA axes compared to the rest, while d = 0.5 in the top-right defines the scale, the length of the side of a grid square being 0.5.

Replacing the `scatter()` directive with

```
> s.label(JRAca2$li, label = date)
```

produces Fig. 5, which shows the row plot labelled by date. Using

```
> s.label(JRAca2$co, label = type)
```

produces the variable plot (not shown). The `$li` and `$co` parts in the above code identify plotting coordinates held in the object `JRAca2` originally created.

You can do some quite clever things with a little experimentation (suggested by the help facilities and material found via `Google`). For example,

```
> bet <- between(JRAca2, as.factor(date), scannf = FAL-
SE)
> s.class(bet$ls, as.factor(date), xax = 1, yax = 2)
```

produces Fig. 6 in which the ellipses emphasise the separation of the date groups. As with many R functions there is a lot of control over the appearance and labelling, and what is presented here is basic. Use the `help()` directive described earlier to see what is available.

3.6 *Seriation and detrended CA*

A common use of CA is for seriation (Madsen 1988, provides numerous, and in some cases idealized, examples). Usually results from a CA are presented as a two-dimensional graph from which it is hoped that a one-dimensional ordering, interpretable as a temporal "gradient", can be read off. In some instances, as in the example used here, a "gradient" can be interpreted as a spatial one.

To fix ideas, Table 2 reproduces Table 5 from Cool and Baxter (1999). This shows EVE values for seven glass drinking vessel types, from contexts dating to the later 1st century AD. The contexts are ordered from north to south – Carlisle to Fishbourne – with the first three from the north, the next two from the Midlands, and the remaining five from the south (we are aware that the numbers are rather small, and discuss the more general issue of sample size in a later section.)

Calling the data set `Flavian`, and using `corresp` from the `MASS` library

```
> Flavianca1 <- corresp(Flavian,nf = 2)
```

```
> biplot(Flavianca1, ylabs = rep(" ", 7))
```

Fig. 7, showing the plot for contexts, results. Labels correspond to the order of contexts in Table 2.

The figure has an approximate "horseshoe" shape, which is what is usually hoped for. We can read round the horseshoe from 10 in the bottom left to 3 in the bottom right to get a one-dimensional ordering which in this case corresponds, more-or-less, to the ordering on the first axis (the positioning of context 1, which lies off the horseshoe, is a little ambiguous). With the exception of context 2 (York), which is a bit out of order, the ordering corresponds to a south-north gradient and, in conjunction with the plot for vessel types, Cool and Baxter (1999, 90-91) interpreted this as evidence of regionality in the assemblages. (It was argued that the southern sites were characterised by newer Flavian forms, with the northern and Midland sites characterised by older Claudio-Neronian forms – these differences not being related to site type.)

CA as used here suggests *relative* chronological or spatial ordering. Ecologists, and others, who have used and developed CA extensively, would sometimes like to be able to interpret distances between points on the graph in *absolute* terms. Characteristically, and with larger data sets, there is also bunching at the terminals of the horseshoe that can hamper interpretation. Detrended Correspondence Analysis (DCA) attempts to rectify these problems (for a more extensive discussion of DCA and seriation in archaeology see Lockyear 2000a).

For archaeological applications to seriation problems we do not view the horseshoe as a problem (in fact achievement of the horseshoe effect is often seen as evidence of success of a seriation). Some aspects of DCA methods, which can be thought of as algorithms to "unbend" the horseshoe, have been considered to be arbitrary (see Baxter 2003, 139-40 for a brief discussion) and it is primarily discussed here, both to further illustrate the potential power of R and for the benefit of those archaeologists who may wish to explore it.

The function `decorana` in the library `vegan` may be used. This library will need to be downloaded in the same way as `ade4` was. Then

```
> library(vegan)
> Flaviandca <- decorana(Flavian)
> plot(Flaviandca)
```

with results in Fig. 8. We don't think this adds anything to the previous analysis.

In first using R it will often be the case that, as above, the simplest of analyses (that will often be suggested by examples given in the help for a function) are adequate. To improve graphs – for publication purposes, for example – a little experimentation usually helps. It does not take long, for instance, to work out that

```
> plot(Flaviandca, display = "sites")
```

drops the vessel-type labelling from the previous plot. If you prefer the sites to have names rather than numbers create a variable, for example `sitenames`, as discussed previously, and use

```
> plot(Flaviandca, display = "sites", type = "n")
> text(Flaviandca, display = "sites", sitenames)
```

`type = "n"` produces a blank plot and the `text()` function adds the desired labels.

4. Some practicalities

4.1 *Dealing with small numbers*

By "small numbers" we mean "small" row and column totals. It is possible, though not inevitable, for such small numbers to have an adverse effect on a CA display and interpretation. In the most extreme case, with a zero total, the row or column affected cannot be used at all. With small but non-zero totals, omitting offending rows (columns) is an obvious possibility. It is also legitimate to amalgamate rows (columns) to obtain larger totals, providing the newly defined rows (columns) have a legitimate archaeological interpretation.

Our preferred approach is to retain all the data, in the first instance. This is because, in some applications, numbers are inevitably small and it seems wasteful to throw them away without first seeing if they nevertheless have a useful story to tell. Table 2 is an example. If there are problems (see below) then the courses of action already alluded to are available.

4.2 *Dealing with outliers*

One problem that can arise with small totals is that the associated row (column) marker on a plot appears as an outlier. An outlier is a point that lies at some distance from other points on a CA plot. It can represent a "rogue" data point, possibly arising from small numbers, or may be genuine and associated with a large total but simply very different from other rows (columns).

Whatever the cause, a major problem is that a serious outlier will determine the scale of a plot, possibly causing other points to bunch together, obscuring interpretable pattern in them. In our view it is almost always sensible to re-do a CA omitting obvious outliers, to see what patterns – if any – have been obscured in the remaining data. This is related to, but separate from, the final presentation and interpretation of results. This will be determined by those analyses which tell the most informative story, and could, for example, involve plots both with and without outliers, or only the latter, with

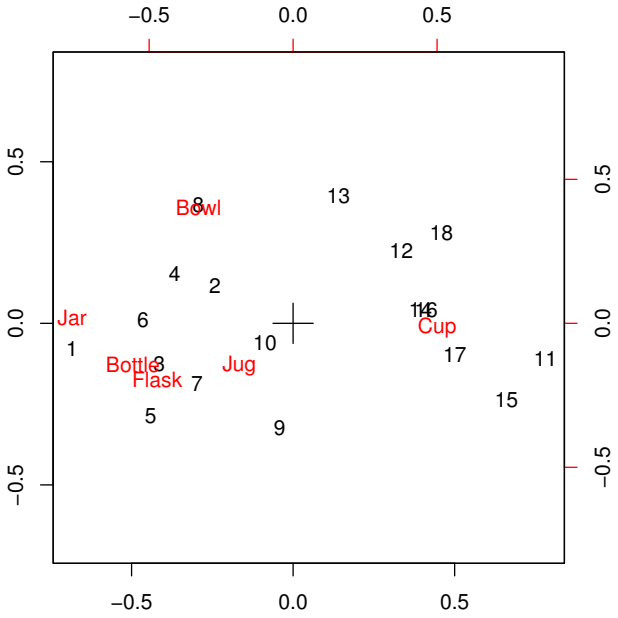

Fig. 1 – Correspondence Analysis of Table 1 – superimposed row and column plot. The `corresp` function from the MASS package was used.

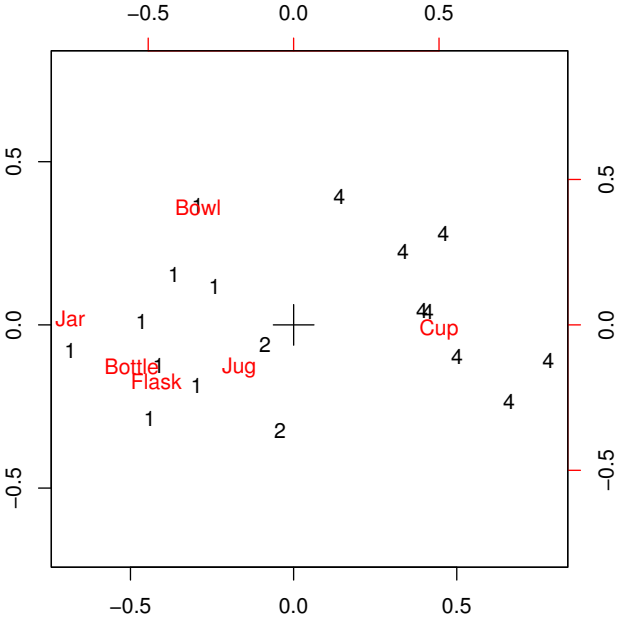

Fig. 2 – Correspondence Analysis of Table 1 – superimposed row and column plot; rows labelled by date (1 = 1st/2nd century AD; 2 = 2nd/3rd century; 4 = 4th century). The `corresp` function from the MASS package was used.

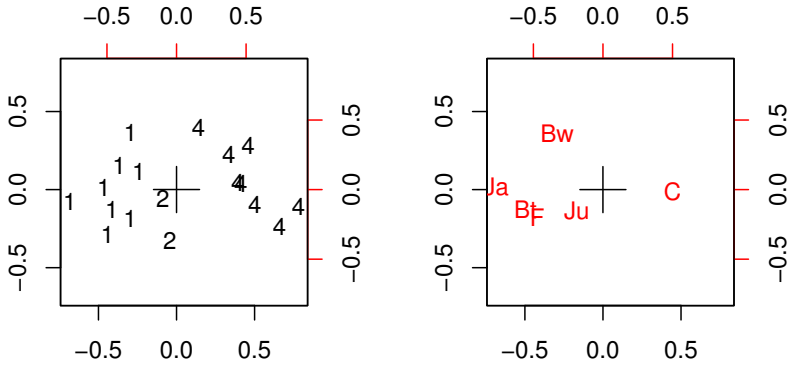

Fig. 3 – Correspondence Analysis of Table 1 – separate row and column plots. The `corresp` function from the MASS package was used.

Fig. 4 – As Fig. 1, using the dudi.coa function from the ade4 package.

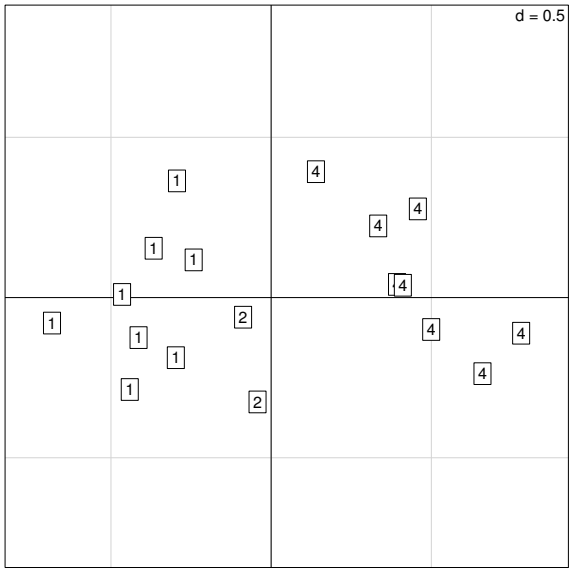

Fig. 5 – Row plot for Table 1, labelled by date, using the dudi.coa function from the ade4 package.

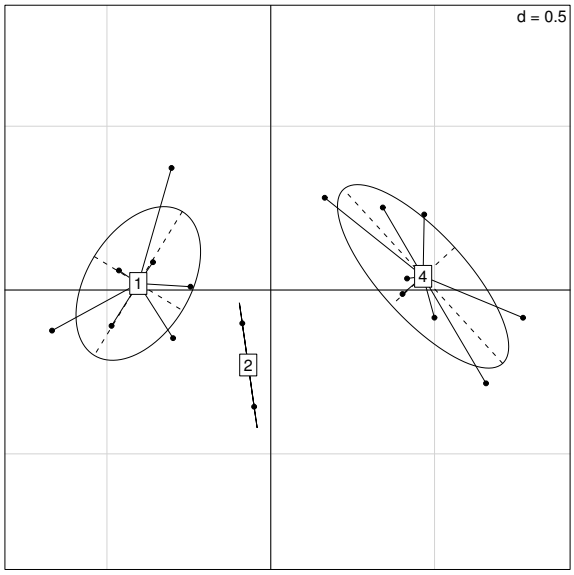

Fig. 6 – An alternative presentation to Fig. 5, associating ellipses with different date groups to emphasise their separation.

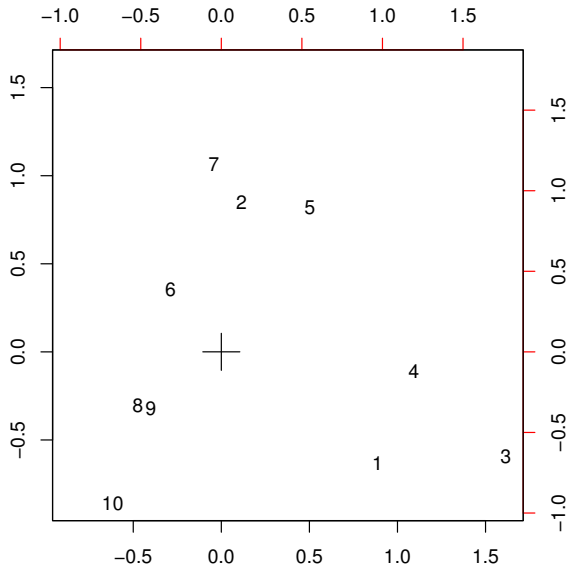

Fig. 7 – Correspondence Analysis of Table 2 – row plot onl. The `corresp` function from the MASS package was used.

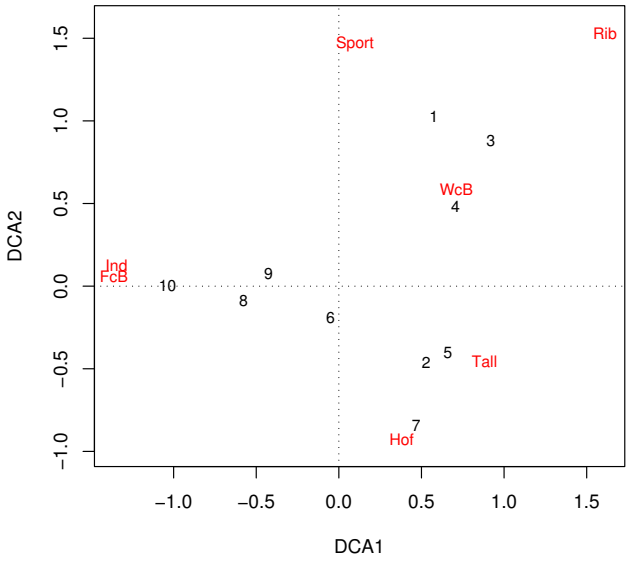

Fig. 8 – A Detrended Correspondence Analysis for the data of Table 2, obtained using the `decorana` function from the vegan package.

outliers and their cause considered separately. We emphasise this point since some archaeologists (mistakenly, as we think) regard this kind of omission as subjective and unjustified data manipulation.

4.3 *Dealing with structure ("peeling" the data)*

The foregoing discussion applies equally well to "small" and distinct groups of data. In some instances, and Figs. 1-4 provide examples, distinct groups that are not "small" may be manifest in a plot and, it is hoped, lend themselves to ready archaeological interpretation. We refer to such grouping, and other forms of obvious pattern, as "structure".

A feature of many pattern seeking multivariate statistical methods is that they will reveal structure in a set of data, but often this structure is (at least in retrospect) rather obvious. It can be argued that our analyses of Table 1 are of this kind.

The sensible thing to do in these circumstances is to strip obvious groups from a set of data, and subject these, and what remains, to separate CA analyses to try and identify more subtle pattern in the data. We refer to this process as "peeling" the data, and its application is illustrated in Cool and Baxter (1999). Although not a seriation, the main message is that there is structure in the data having a temporal interpretation associated with changing patterns of vessel usage. Having established this, a separate focus on temporally homogenous groups may reveal difference associated with site types, or spatial disposition, not evident in the original analysis.

4.4 *Data transformation*

A feature of good CA software is that axes of the plots are equally scaled, so that the configuration of points can be read exactly as one would read a map. It sometimes happens that this results in rather elongated plots that are unpleasing to the eye and possibly difficult to read. Outliers can have this effect as well. Although not widely used in archaeological applications, data transformation can sometimes avoid this problem. Lockyear's (2000b) combination of outlier removal and a square-root transformation in his study of coin finds from Roman sites in Britain provides a pleasing illustration of the effectiveness of this strategy. We do not emulate this here, but note that if one wanted to base an analysis on the square-roots of the data in Table 1, `JRAsqrt <- sqrt(JRA1)` suffices to produce the square-rooted data for operating on.

5. Conclusion

This paper has had the limited aim of providing sufficient information to allow archaeologists with the interest, but not necessarily the statistical

training or software (as they – we hope incorrectly – think) to carry out correspondence analyses on their own data, using state-of-the-art statistical software for free. We suspect that anyone who dips into this and perseveres will be entranced by the power and possibilities of R – if only for publication quality graphics – and will want to learn more. We do not provide detailed references here, but there is a lot available on the web and the CRAN site listed earlier provides references to printed sources (which are increasingly available at all levels).

Michael J. Baxter
Department of Physics and Mathematical Sciences
Nottingham Trent University

Hilary E.M. Cool
Barbican Research Associates
Nottingham

REFERENCES

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

Baxter M.J. 2003, *Statistics in Archaeology*, London, Arnold.

Bølviken E.E., Helskog K., Holm-Olsen I., Solheim L., Bertelsen R. 1982, *Correspondence Analysis: an alternative to principal components*, «World Archaeology», 14, 41-60.

Cool H.E.M., Baxter M.J. 1999, *Peeling the onion: an approach to comparing vessel glass assemblages*, «Journal of Roman Archaeology», 12, 72-100.

Cowgill G.L. 2001, *Past, present and future of quantitative methods in United States archaeology*, in Z. Stančič, T. Veljanovski (eds.), *Computing Archaeology for Understanding the Past. CAA 2000*, BAR International Series 931, Oxford, Archaeopress, 35-40.

Djindjian F. 2009, *The golden years for mathematics and computers in archaeology (1965-1985)*, in P. Moscati (ed.), *La nascita dell'informatica archeologica. Atti del Convegno internazionale (Roma 2008)*, «Archeologia e Calcolatori», 20, 61-73.

Duff A.I. 1996, *Ceramic micro-seriation: types or attributes?*, «American Antiquity», 61, 89-101.

Greenacre M.J. 1984, *Theory and Applications of Correspondence Analysis*, London, Academic Press.

Hill M.O. 1974, *Correspondence Analysis: a neglected multivariate technique*, «Applied Statistics», 23, 340-354.

Lockyear K. 2000a, *Experiments with detrended Correspondence Analysis*, in K. Lockyear, T.J.T. Sly, V. Mihăilescu-Bîrliba (eds.), *Computer Applications and Quantitative Methods in Archaeology: CAA96*, BAR International Series 845, Oxford, Archaeopress, 9-17.

Lockyear K. 2000b, *Site finds in Roman Britain: a comparison of techniques*, «Oxford Journal of Archaeology», 19, 397-423.

Madsen T. (ed.) 1988, *Multivariate Archaeology*, Aarhus, Aarhus University Press.

Orton C. 1999, *Plus ça change? - 25 years of statistics in archaeology*, in L. Dingwall, S. Exon, V. Gaffney, S. Laflin, M. van Leusen (eds.), *Archaeology in the Age of the Internet: CAA97*, BAR International Series 750, Oxford, Archaeopress, 25-34.

Ringrose T.J. 1988, *Correspondence analysis as an exploratory technique for stratigraphic abundance data*, in C.L.N. Ruggles, S.P.Q. Rahtz (eds.), *Computers and Quantitative Methods in Archaeology 1987*, BAR International Series 393, Oxford, 3-14.

Shennan S. 1997, *Quantifying Archaeology: Second Edition*, Edinburgh, Edinburgh University Press.

Venables W.N., Ripley B.D. 2002, *Modern Applied Statistics with S: Fourth Edition*, New York, Springer.

Appendix

Cup	Bowl	Jar	Flask	Jug	Bottle
4.4	6.8	1.1	2.44	4.22	14.44
9.2	5.6	1.8	1.49	3	5.75
6.4	4.2	0.52	3.68	2.72	7.28
4	3.52	0.69	0.94	1.54	3.78
3.6	1.2	1.43	2.94	1.42	2.52
5.6	5.2	1.03	3.02	2.84	6.72
5.2	2.13	0.57	1.84	2.4	4.48
6.4	6	1.44	2.2	0.84	3.08
4.2	0.4	0.4	0.65	1.37	2.31
4.2	1.4	0.36	0.6	1.12	2.31
25.8	0.8	0	0.8	0.7	1.84
9.2	3.2	0	1.02	1.44	0.61
5	2.8	0	0	0.84	1.36
8.73	1.8	0	0.27	1.47	1.45
9.4	0	0	0.07	1.18	1.27
10.2	2	0	0.6	1.26	1.56
24.2	2.6	0	1.8	4.13	2.32
10	3	0	0.2	0.7	0.98

Tab. 1 – Data in EXCEL format, as originally entered into R. Rows correspond to sites, columns to glass vessel types and entries to estimated vessel equivalents. See the text for site dates and Cool, Baxter 1999, 80 for the detailed table.

Sport	Tall	Rib	Hof	Ind	FcB	WcB
0.8	0.2	1	0	0.2	0.2	0.4
0	0	0.2	2.2	0	0	0.6
0	0.4	1.4	0	0.2	0.2	0.2
0	0	0.6	0.4	0	0.2	0.2
0	0.2	0.6	2.2	0	0	0
0.8	0	0	2.4	0.6	0.4	0.2
0	0.6	0	2.8	0	0	0
0	0	0	1.2	1	1.8	0.8
2	1	0	2.8	3	3.2	1
0	0.2	0	0	0.4	0.8	0

Tab. 2 – Data in EXCEL format, as originally entered into R. Rows correspond to sites, columns to glass vessel types and entries to estimated vessel equivalents. Sites are ordered from north to south (Cool, Baxter 1999, 89 for the detailed table).

ABSTRACT

Correspondence Analysis (CA) is a popular tool for archaeological data analysis, appropriate for use with tables of non-negative number. The technique allows the visual display of the associations between the rows and between the columns of a data matrix, and the relationships between them. Archaeologists with this kind of data often have no problem in understanding the ideas behind CA, but with limited training in statistics may have problems in implementing it. Commercial, menu driven, statistical software packages of the type used for service teaching in universities are expensive and restrictive in the way results from a CA can be presented. Archaeologists outside the university sector may not have access to such software. This paper is a guide to how the open-source software R can be used to undertake CA. R is a sophisticated, "state-of-the-art" package that is constantly updated. It is not menu driven and can seem forbidding to new users. The paper provides a detailed account, ranging from installation of the package through to real applications of CA, that has helped, and we hope will continue to help, encourage the use of CA among archaeologists who have previously been discouraged from engaging with it.