Chapter 9

Correspondence analysis

9.1 Introduction

Many of the ideas that underpin correspondence analysis (CA) are similar to, if not identical with, those for PCA. Differences between PCA and CA, and some technical detail, are discussed in Section 9.2.

Baxter (1994a: 133–39) summarizes the development of the use of CA in archaeology to about 1992; taking the story slightly further in Baxter (2003: 12–13). A brief resume is that Hill (1974), in a statistical journal and using an archaeological seriation problem as an example, described CA as a 'neglected multivariate technique'. Benzécri and colleagues, in the French-language literature in the 1970s and 80s, is widely credited with the modern mathematical development of CA. This literature is equally credited with being a difficult read. Greenacre's (1984) English text is also heavy going. Greenacre (2007) – a thorough revision of Greenacre (1993) – is more approachable. The use of CA for archaeological purposes in the French-language literature was little noticed elsewhere, and it is Bølviken *et al.* (1982) who are usually credited with popularizing archaeological uses of the method.

The edited collection of Madsen (1988a), with many examples, helped CA on its way. The method did not 'catch on' in Britain until the early 1990s, with North America lagging behind. The use of CA in archaeology is now commonplace and CA is now mentioned in the same breath as cluster analysis and PCA, the most widely used multivariate methods in archaeology.

An introduction to CA in `R`, written for archaeologists, is available in Baxter and Cool (2010b). This chapter uses packages not available when that paper was written. More recently, Alberti (2013) has published on the use of `R` for CA with a view to developing scripts for those more comfortable with menu-driven analyses..

9.2 CA and PCA – similarities and differences

This section is more 'mathematical' than the rest of the chapter and some readers may prefer to go directly to the examples of Section 9.3.

Usually CA is presented as a technique for analyzing tables of counted data, in contrast to PCA which is usually applied to continuous data. In fact, CA can be applied to any table of non-negative numbers, and PCA to counted data, but the distinction drawn between them is that usually emphasized. To reflect this, notation is changed here. Let $\mathbf{N}$ be the $I \times J$ table of counted data, with typical element n_{ij}. For later reference the sum of the n_{ij} is n, the sum for row i is r_{i+}, and the sum for column j is c_{+j}.

A frequent selling-point of CA is that it can jointly represent both the rows and columns of a data table, aiding interpretation. This oversells the difference between CA and PCA since a joint representation is possible with the latter, and the joint representation is not compulsory for CA. If a joint representation is used it is called a *biplot* and often takes the form of a plot of the column markers superimposed on those for the rows, The examples to follow present the output as two separate and adjacent plots, one for the rows and one for the columns. This is often easier to read, particularly with large data sets, and does not detract from the interpretation. Mathematically the treatment of the rows and columns is symmetrical, so only the former is treated in detail here. Greenacre (207: 31–32) is a convenient notational summary that provides details for the treatment of both rows and columns[1].

In PCA, with an $n \times p$ data matrix, the idea is to produce a map in which distances between the row markers approximate the true distances in the full p-dimensional space. This is explained in more detail in Section 7.3. The aims of CA can be described in an identical fashion. The way 'variance' and 'distance' are defined differs, however, and is discussed below. In CA the 'variance' is called the

[1]Biplots can be presented in several ways, about which a lot has been written. Treated algebraically, via the singular value decomposition (Section D.2), a 'strict' interpretation of a biplot is that row and column representations can be combined to approximately reproduce the data. When these are used for plotting this can result in row (column) markers being plotted round the periphery of the plot with column (row) markers bunched up in the center. This 'asymmetrical' treatment can make plots difficult to read, so a symmetrical treatment that is more readable is often preferred, although this then loses the property that the data can be reconstructed algebraically from the row and column representations. Since CA is mainly used to produce an interpretable graphical representation of the relationships between rows, columns and each other this does not trouble many users. When plotted separately the relative positions of row and column markers on their respective plots is informative about the relationship between row and column categories, best illustrated in the examples to follow. It is obligatory to issue a warning that an interpretation of the difference between the positions of row and column markers as a 'distance' is not valid.

inertia and *chi-squared distance* is used, in contrast to the use of Eucidean distance in PCA. This can be thought of as a weighted PCA, and it is the introduction of weights that complicates the mathematics.

The chi-squared test statistic for no association between the rows and columns of a data table is often written as

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Here $O_{ij}(= n_{ij})$ is the observed value in cell (i, j) and E_{ij} its expected value. The latter is defined as

$$E_{ij} = r_{i+}c_{+j}/n.$$

Thus, X^2 can be written as

$$X^2 = \sum_{ij} \frac{(n_{ij} - r_{i+}c_{+j}/n)^2}{r_{i+}c_{+j}/n}.$$

The *total inertia* is *defined* as X^2/n, the division by n removing the sample size effect that X^2 is subject to. It is a measure of the variance in the data, the contributions of individual cells being $(O_{ij} - E_{ij})^2/nE_{ij}$. These can be summed across rows or columns to get row and column inertias. The *mass* of row i is defined as $r_i = r_{i+}/n$ and of column j as $c_j = c_{i+}/n$, which can be collected together as $\mathbf{r} = (r_1\ r_2\ \ldots\ r_I)$ and $\mathbf{c} = (c_1\ c_2\ \ldots\ c_J)$. The elements of $\mathbf{r}$ and $\mathbf{c}$ sum, by definition, to 1; $\mathbf{c}$ defines the average profile of the rows and $\mathbf{r}$ that for the columns. The masses play an important role in defining the weighting used in CA.

Define $p_{ij} = n_{ij}/r_{i+}$ for $j = (1, 2, \ldots, J)$; the profile for row i is then given by $\mathbf{p}_i = (p_{i1}\ p_{i2}\ \ldots\ p_{iJ})$. The aim in CA is to represent the profiles on a 'map' where the Euclidean distances between the row markers on the map approximates the chi-squared distance between profiles[2].

Greenacre (2007: 32) shows that the contribution of cell (i, j) to the total inertia is proportional to

$$(p_{ij} - c_j)^2/c_j$$

where the numerator is just the square of the difference between observed contributions to the profile and their expected values, c_j. What this formulation makes clear is the introduction of weights dependent on the c_j. In effect this amounts to weighting profile contributions by $1/\sqrt{c_j}$. Remembering that $c_j = n_{+j}/n$, larger values correspond to columns with more observations. Thus, relative to Euclidean

[2]This definition of p_{ij} departs from the notation used in Greenacre (2007: 31) who defines it as $\tilde{p}_{ij} = n_{ij}/n$. This involves replacing p_{ij} in the following equation with $\tilde{p}_{ij}/r_i$.

distance, categories with low frequencies receive a higher weighting than those with larger frequencies. This 'evens out' the influence that rows with small and those with larger frequencies have on the CA. It is analogous, in a way, to the use of standardized data in PCA, where the standardization $(x_{ij} - \bar{x}_j)/s_j$ can be thought of as weighting of the x_{ij} to remove the undue influence of high-variance variables.

9.3 Romano-British glass assemblages

9.3.1 First- to third-century vessel glass

This is an extended example, designed both to illustrate one approach to the use of CA and to elaborate on aspects of interpretation. It involves an examination of the use of Romano-British vessel glass in the first- to third-centuries AD. The data are given in Table B.13 and are a slightly modified version of tables from Cool and Baxter (1999). The table is based on estimated glass vessel equivalents (glass EVEs) for 25 sites.

For the purpose of assemblage comparison numbers need to be directly comparable. When material is fragmented several commonly used measures (e.g., fragment count, minimum number of vessels) lack this property. Orton (1975) developed estimated vessel equivalents (EVEs) for pottery data in response to this situation. Diagnostic sherds are needed, and typically the proportions of the rim that survive on rim sherds are counted. Later the ideas were extended to other materials. Moreno-Garcia *et al.* (1996) developed a zonal method for quantifying bone data. A complete bone is defined by a number of zones (which can vary with bone type) and the proportion of zones recognizable in a fragmented bone are counted. The first author of Cool and Baxter (1999) drew inspiration from this to develop glass vessel EVEs. The resultant data are fractional but comparable and CA can be applied in the usual way.

An analysis, not shown here, was undertaken on first- to fourth-century glass. The interpretation was obviously chronological, with the fourth-century separating out completely. Accordingly the fourth-century data was removed and the analysis here begins with the first- to third-century data. The last two rows in Table B.13 are not used in the first instance. Two chronological groups have been defined according to whether occupation on the site ends before or after 150 AD. This is an example of what Cool and Baxter (1999) call 'peeling' the data, where the more obvious structure is removed from the analysis to reveal more subtle aspects of patterning, if any.

Figure 9.1 shows the outcome of a CA using the `ca` function from the `ca` package. The inertias are shown on the axes. The obvious interpretation is, again,

chronological. Three of the later sites sit within the region occupied by the earlier sites, or are on the same side of the plot. This is explored in the paper, where the decision was made to analyze the two chronological groups separately, as is done in the next two subsections. Some of the diagnostic information available is illustrated in Table 9.1. Code for carrying out the analysis is given in Section 9.6

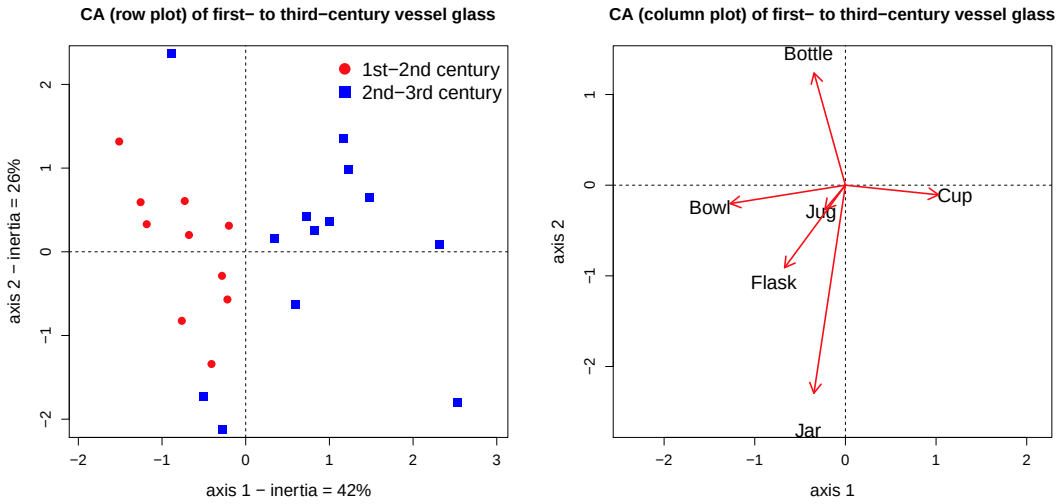

Figure 9.1: *Row and column plots for a correspondence analysisof the first- to third-century AD Romano-British vessel glass assemblages of Table B.13.*

Table 9.1 and those to follow are presented as they appear in the relevant `ca` output. The total inertia, 0.256, is analogous to the total variance in PCA, and the inertias for the individual axes are analogous to the variance of the individual components in PCA. The first two axes account for 68.5% of the total inertia.

dim	value	%	cum%
1	0.10771	42.1	42.1
2	0.06781	26.5	68.5
3	0.03897	15.2	83.8
4	0.02324	9.1	92.8
5	0.01834	7.2	100
Total:	0.25607	100	

Table 9.1: *Inertias from a correspondence analysis of first- to third-century vessel glass assemblages from Table B.13.*

9.3.2 First- to second-century vessel glass

This analysis is based on the first 10 rows of Table B.13 to which have been added the last two rows of that table (i.e. sites where occupation terminated before 150 AD). The CA, at first sight, does not reveal any obvious pattern, but other information is to hand that allows a more suggestive interpretation. This is that the sites can be classified as military or civilian. If this is used to label the row plot it can be seen that, with one exception for each site-type, the military sites plot to the right and the civilian sites to the left. In Cool and Baxter (1999) this is interpreted as showing that by the Flavian period (roughly the last-third of the first-century AD) the civilian population had developed a pattern of vessel-glass use that differed from that of the military.

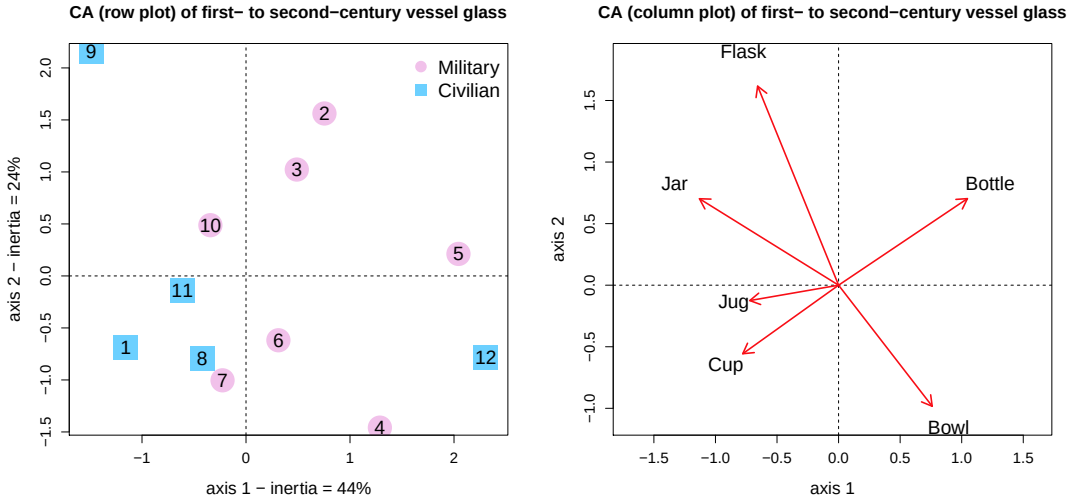

Figure 9.2: *Row and column plots for a correspondence of the first- to second-century AD Romano-British vessel glass assemblages of Table B.13.*

Comparison with the variable plot suggests that military sites are, relatively speaking, characterized by a higher proportion of bottles and bowls, with civilian sites having more of the other types. The outlying civilian site in the top-left of the row plot has a higher proportion of flasks than other sites. Numerical flesh can be added to this interpretation and serves to illustrate further aspects of the `ca` package. Diagnostic statistics are presented in Table 9.2 for the variable plot.

The masses, which are the column totals divided by n, are scaled to add to 1000. The quality (qlt) shows how well each vessel type is represented in the plot; numbers are scaled to a maximum of 1000 to aid comparisons. The larger the values are the better the representation. Thus cups (933) and bottles (901) are the two types best represented, with jugs and jars the least well-represented.

type	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
Cup	281	933	144	-213	727	236	-113	205	121
Bowl	214	803	181	208	418	171	-200	385	286
Jar	46	299	146	-309	246	81	142	52	31
Flask	109	711	175	-180	164	65	328	547	392
Jug	106	271	126	-197	267	76	-25	4	2
Bottle	245	901	228	286	722	371	143	180	167

Table 9.2: *Diagnostic statistics for columns for the CA of the first- to second-century AD Romano-British vessel glass assemblages of Table B.13.*

The entries for `k=1` and `k=2` are plotting positions in terms of *principal* coordinates. For `k=1` negative values plot to the left (west) of the plot; positive values plot to the right (east). For `k=2` negative values plot in the lower-half (south) of the plot; positive values in the upper-half (north)[3].

The entries labelled 'cor' are squared correlations between the columns and each of the first two axes and the quality is defined as the sum of these two terms. That is, they measure how the quality is decomposed between the first two axes. Thus, of the four vessel types where the overall quality of represntation is good, cups and bottles, with squared correlations of 0.73 and 0.72 with the first axis (numbers in the table are multiplied by 1000), different signs for `k=1`, and fairly small values for `k=2`, might expected to lie roughly along an east-west axis at some distance from the origin. Bowls make a significant contribution to both axes, while flasks are particularly dominant on the second axis.

Most of this is probably more evident from the figure than the table; what the latter does do is warn against over-interpretation of the prominence of jars in the figure, since the quality of representation is relatively poor. The columns *ctr*, scaled to add to 1000, are the contributions to the inertia of the associated axes of the different types. This highlights, once again, the importance of bottles and cups in defining the first axis and bowls and flasks in defining the second.

9.3.3 Second- to third-century vessel glass

The CA plots for the second- to third century glass are shown in Figure 9.3. The row plot was examined in Figure 7 of Cool and Baxter (1999) by labeling sites according to type. This was perhaps most interesting for what it didn't show,

[3]There is a complication in that two coordinate systems are avaialble, *standard* as well as principal coordinates (Greenacre, 2007: 62). The former are scaled to have zero mean and unit variance and are the values extracted from the `ca` object used for plotting in the figures. They thus differ numerically from the output of Table 9.2, though not in import, because of the different scaling of the two coordinate systems.

since there were no clear patterns with respect to site-type. The civilian/military distinction observed for the first- and second-centuries largely disappeared. The two separated sites to the right (3, 4) are small urban settlements but plot opposite two similar sites (1,8). The most extreme points at the bottom and top of the plot (7, 9) are auxiliary forts; that is, they are the same type but don't occupy the same region of the plot.

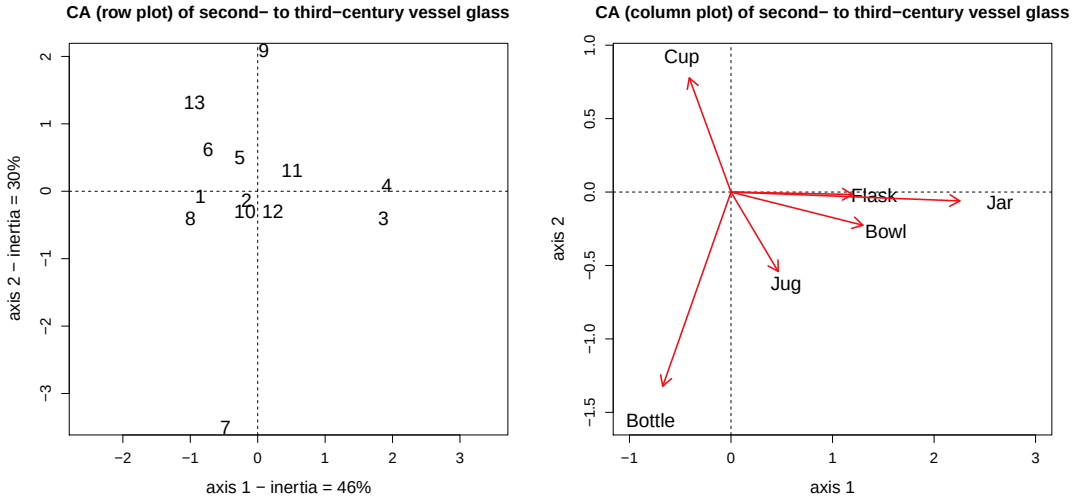

Figure 9.3: *Row and column plots for a correspondence analysis of the second- to third-century AD Romano-British vessel glass assemblages of Table B.13.*

Diagnostic statistics for the rows are shown in Table 9.3. They have the same interpretation, with respect to rows, as Table 9.2 has for columns. Site 7 (Rochester) is a clear outlier in Figure 9.3 and this is reflected in the diagnostic statistics where it dominates the second axis, with lesser contributions from sites 9 and 13. It is characterized by an unusually high proportion of bottles. Sites 9 (Housesteads) and 13 (York 160-289 AD.), to the north, are, by contrast, characterized by cups. These inferences can be confirmed by examining the corresponding table for columns (not shown).

As far as the first axis goes the two most extreme sites to the east, Towcester (3) and Harlow (4), are differentiated from other sites by the relative proportion of jars. Inspection of Table B.13 confirms this. Site 9 (Housesteads) has a comparable proportion of jars, but is overwhelmingly dominated by cups and is, in consequence, a major contributor to the definition of the second axis rather than the first.

The row plot in Figure 9.3 shows that most other rows cluster fairly close to the origin and are not well represented by the CA. In summary, the overall picture obtained is largely determined by three types and four or five sites. A

Id.	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	194	786	82	-304	782	140	-21	4	1
2	76	118	12	-59	81	2	-40	37	1
3	61	799	126	668	775	213	-118	24	10
4	124	867	240	684	866	453	24	1	1
5	75	226	36	-95	67	5	146	159	19
6	32	706	17	-263	480	18	181	226	13
7	47	910	197	-171	25	11	-1014	885	578
8	76	865	44	-357	782	76	-116	82	12
9	51	569	118	32	2	0	603	567	223
10	67	141	21	-67	52	2	-88	89	6
11	53	513	15	182	412	14	90	101	5
12	72	198	18	80	93	4	-85	105	6
13	72	902	73	-335	393	63	381	509	124

Table 9.3: *Diagnostic statistics for rows for the CA of the second- to third-century AD Romano-British vessel glass of Table B.13.*

detailed archaeological interpretation of the results is provided in Cool and Baxter (1999) – the intention here has been to illustrate the 'peeling' process and the statistics available for interpreting the output. In practice some these may often be superfluous mainly confirming what is evident from inspection of the plots.

The statistics in the tables can be viewed as different ways of assessing the 'validity' of a CA that aid its interpretation. For such purposes Ringrose (1992) suggested assessing the stability of a CA using bootstraping/resampling methods. A large number, N, of 'replicate' tables are generated, each being subjected to a CA, so generating N plotting positions for each row/column of the table. For any given row/column, if these plotting positions cluster tightly the representation of that row is stable; if the points are widely spread it may be unsafe to read too much into their positioning on the original CA plots. Convex hulls or confidence ellipsoids (Chapter 6) of the plotting positions can be used to get an overall impression of plot stability. Ringrose's ideas have largely been neglected in the archaeological literature until recently, but have been exploited fruitfully in two recent papers by Peeples and Schachner (2012) and Lockyear (2013). Both use `R` and readers are directed to available `R` code in the former paper.

9.4 Flavian drinking-vessels

Perhaps the most common use of CA in the archaeological literature is for seriation. Seriation, in the ideal case, typically produces an unambiguous ordering of the data and, most commonly, it is hoped that this can be given a chronological interpretation. If this is reasonably successful the pattern in the data is that of a 'horseshoe' on a plot of the first two axes, and the ordering is read around the horseshoe. Here an example is provided where a clear seriation is obtained of a spatial rather than chronological nature. The data of Table B.14 are used.

These are glass vessel EVEs for seven types of drinking-vessel, current during the Flavian period in England, from 10 sites ordered by their north-south orientation, with three from the north, two from the Midlands, and five from the south. The outcome of the CA is presented in Figure 9.4.

With the exception of Site 1 (Carlisle) an almost perfect seriation of the data is obtained in the row plot. Labeling sites by their north-south orientation Site 2 (York) sits apart from others in its region but, though not perfect, the seriation admits a spatial interpretation. Reading around the horseshoe, sites from the north and Midlands are perfectly separated from southern sites on the first axis.

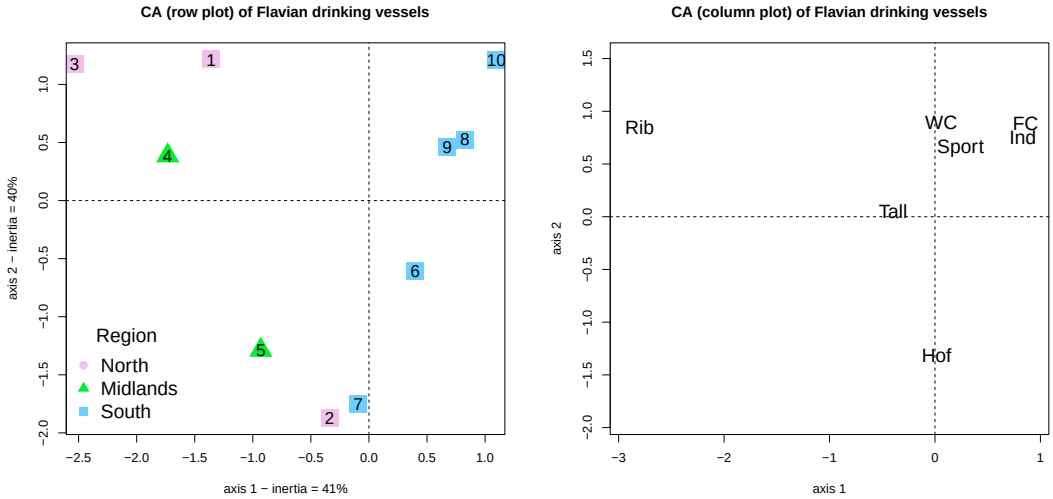

Figure 9.4: *Correspondence analysis row and column plots for the Flavian drinking-vessels of Table B.14.*

Cool and Baxter (1999: 90–91) can be referred to for discussion of these results, which admit more than one interpretation. One is that the north is characterized by earlier vessel forms and the south by later ones, and this may be the 'rule rather than the exception', contradicting 'traditionally' held views that the earlier forms found in the north and Midlands were 'isolated survivals. An alternative interpretation is that drinkers in the north/Midlands favoured low cups while those in the south preferred tall beakers.

If the diagnostic statistics for the variable plot in Table 9.4 are consulted together with Figure 9.4 it can be seen that two cup-types (ribbed and Hofheim) and two beaker-types (indented and facet-cut) are those best represented in the plots. The ribbed and Hofheim cups, in particular, dominate the first and second axes. The quality of representation of the two beaker-types is more evenly split between the two axes; they plot closely together in the north-east quadrant of the plot.

vessel type	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
Sport Cups	91	166	104	148	21	5	394	146	38
Tall Beakers	68	76	56	-239	75	10	34	1	0
Ribbed Cups	97	993	342	-1730	912	759	516	81	70
Hofheim Cups	359	994	246	10	0	0	-796	994	616
Indented Beakers	138	944	75	515	526	96	459	418	79
Facet-Cut Beakers	175	822	132	531	402	129	542	419	139
Wheel-Cut Beakers	73	512	45	37	2	0	543	509	58

Table 9.4: *Diagnostic statistics for columns for the CA of the Flavian drinking-vessel glass of Table B.14*

Ribbed cups only appear in the northern and Midland assemblages (Table B.14), all of which plot, along with these cups, to the left. Hofheim cups are not confined to any particular region, but are particularly prominent in Sites 2, 5 and 7 (York, Caersws, Gloucester) which accordingly plot in a similar region towards the bottom. The indented and facet-cut beakers are particularly prominent in the three most southerly sites, 8 9 and 10 (Caerleon, London, Fishbourne), and also plot closely in the north-east quadrant.

The analysis of these data raises an interesting and more general question. It has been observed (in more than one personal communication) that the numbers (of EVEs) on which the analysis is based are small and asked if this raises questions about the validity of any interpretation based on them. Both the observation and the question are legitimate. The 'generality' of the question arises because in archaeological data analysis one has to work with the data to hand, and the sample sizes involved may often be quite 'small' according to the desiderata sometimes laid down for what constitutes an adequate sample (not, incidentally, a question that necessarily has an easy answer). It can also be argued that most archaeological data analysis is concerned with pattern recognition in some form or other, using the terms 'data' and 'pattern' in a very broad sense.

The thought occurs – and it is difficult to put into words – that *if* there is *obvious* pattern in a data set *and if* a *plausible* archaeological interpretation can be advanced for that pattern (the emphases are important here) then this transcends any purely statistical concerns one might have about sample sizes. That is, intelligent archaeological pattern recognition and interpretation may 'trump' statistical formalism when the two appear to conflict because of sample size doubts.

The thought, though it can be differently articulated, is not terribly original; something of the kind is expressed in Section 3.15 of Doran and Hodson (1975), and others, around at the time when quantitative methodology began to be widely applied, grappled with the conflicting requirements of archaeological and statis-

tical inference. In the context of debates that were taking place at the time it might, fancifully, be seen as a choice between the Scylla of unquestioning acceptance of the rigor and utility of statistical analysis for archaeological purposes, and the Charybdis of complete rejection of statistical methodology. This depends, of course, on the conception one has of what statistics is 'about', and there was particular concern with the merits of 'classical' methods of statistical inference (Chapter 12). The extremes of both positions had their proponents; the issues raised have not disappeared, though they can be discussed in a more sober fashion that was sometimes evident. The strait between the two is not so narrow that it can't be negotiated.

9.5 Anglo-Saxon male graves and seriation

The final example is at the opposite end of the spectrum from that in the previous section which was a rather small one. A large table of 272 male Anglo-Saxon burials characterized by the presence or absence of 80 types of grave goods, coded as 1 or 0, is analyzed[4]. This is an example of *incidence* data; the previous examples used counted *abundance* data, albeit fractional. These latter analyses were exploratory in nature with no strong expectation about any patterns that would emerge; indeed the seriation in the example of Section 9.4, with the evidence of the regional pattern, proved something of a surprise.

By contrast, CA for the purposes of seriation in Anglo-Saxon burial studies is quite common, and typically there is a clear expectation that a chronological seriation exists. Again, commonly, graves are not stratified so this provides no help in relative dating, which is what is attempted in CA applications. If a cemetery was in use for a reasonable period of time, however, it is expected that changing fashions associated with the grave goods will result in graves that are temporally close to each other having assemblages more similar to each other than to temporally more distant graves. It can be shown that in such situations CA is expected 'to work'.

The data were collected for, and analyzed in, Bayliss *et al.* (2013). Analysis there proceeded iteratively, over 50 pages or so. Rather than entering all the data at the start, only a subset of the types were used initially. Graves and types were then omitted if their representation was unsatisfactory; further types added; and the process repeated until a seriation judged to be satisfactory was achieved. There is some further 'tweaking' at the end that is discussed shortly.

The 'philosophy' that underpins this process seems to follow that espoused in Jensen and Høilund Nielsen (1997), the latter being a co-author of the book un-

[4]Given the size of the data set it is not reproduced here but can, with a little effort, be extracted from the Archaeology Data Service archives at
http://archaeologydataservice.ac.uk/archives/view/aschron_eh_2013/

der discussion. A condensed summary of the 'philosophy' is that an underlying seriation is assumed to exist at the outset; analysis proceeds iteratively as just described, omission of graves and types that do not conform to a 'seriation pattern' being justified by the assumption that such a pattern exists. Ideally this is also rationalized on archeological grounds; for example, some grave goods are of a type of some antiquity (compared to other types in the burial assemblage) at the time the burial took place. As described thus there is a resemblance to the 'peeling' process of Cool and Baxter (1999) that informed the analyses of Section 9.3. There are, however, differences. The Jensen/Høilund Nielsen approach assumes a particlar type of structure in the data, whereas Cool/Baxter do not, and an emphasis is on identifying 'outliers' that 'conceal' the seriation. Cool and Baxter, by contrast, envisage the use of CA for purposes other than seriation, where the emphasis is on identifying patterns in the data, and removing the more obvious structure to reveal more subtle features.

In addition to the kind of information normally used for this kind of exercise, 48 radiocarbon dates were available for 40 of the male graves, indicated in Figure 9.5. This is the preferred seriation in Bayliss *et al.*[5]. The seriation provides a relative ordering of the graves and is used to suggest phasing for the data. What is novel is the way Bayesian modeling is used in conjunction with the seriation to provide date estimates for the phase boundaries. In Figure 9.5 the ordering of the dated graves is determined by the ordering of the graves on the first CA axis. Some of the phasing is rather 'fine' and some phases contain few dated graves.

What the phasing does is provide prior information about the relative chronological position of subsets of the dated graves that feeds into the modeling process. Not all of the original data set contribute to the seriation and some of the graves omitted have associated radiocarbon dates. The 'almost' final seriation provides partial information on the relative chronology of these graves that allows them to be fitted into what becomes the preferred seriation.

The combination of techniques used converts the relative chronology provided by the seriation into an estimated absolute chronology. This provides finer dating evidence for the sixth to seventh-centuries than that previously available, including an estimate of the 'end' of Anglo-Saxon occupation that places it in the later seventh-century rather than the earlier eighth-century, as previously accepted. This is an eye-catching conclusion (for Anglo-Saxon scholars at least) and is a good example of innovative statistical methodology leading to archaeologically important interpretations.

There are open questions, as the authors acknowledge. The preferred seriation is based on calibrated radiocarbon dates that assume fish is not an important

[5]Their Figure 6.49 on p. 286. Figure 9.5 reproduces this closely, but with some embellishment not in the original; `R` was used, rather than the software in the book.

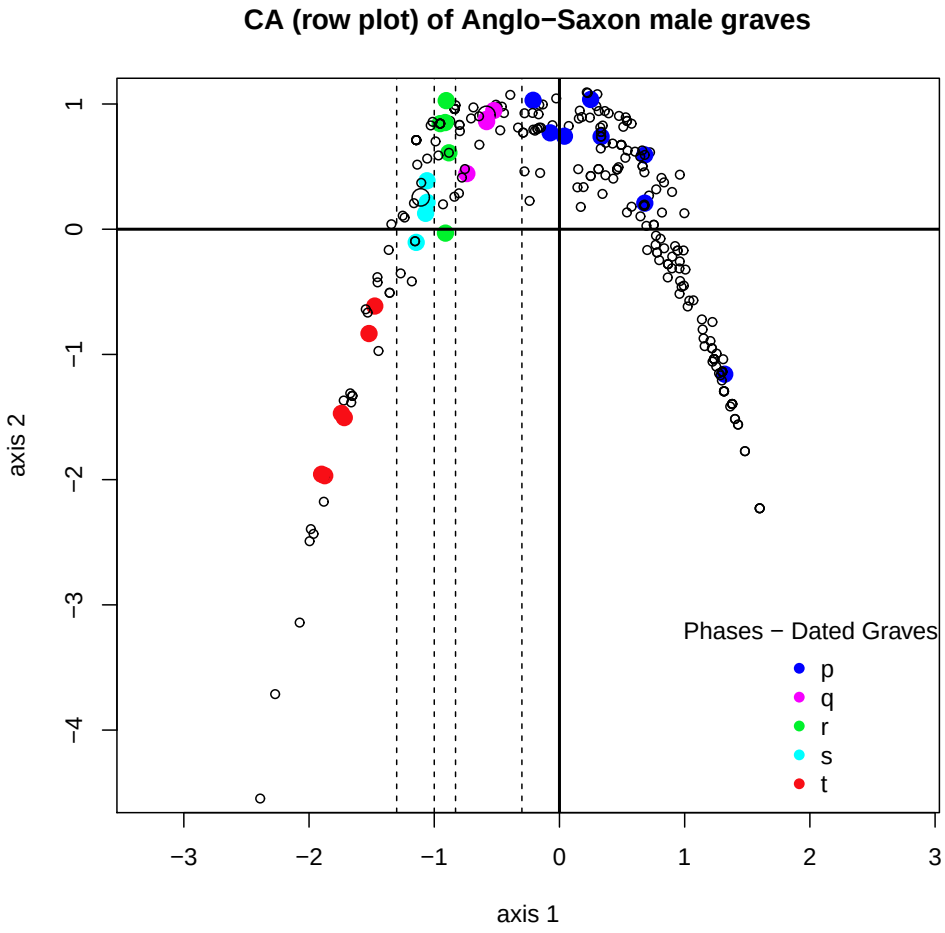

Figure 9.5: *A seriation of 6th and 7th century Anglo-Saxon graves from Britain. Dashed vertical lines are phase boundaries. (Data source: Bayliss* et al.*, 2013.)*

component of diet. Were fish consumption non-neglible – and the evidence in the book allows this reading – the estimated end date would be later, though still earlier than the previously presumed end. For some of the later, and dated, burials, there is also an unresolved conflict with dates derived from numismatic evidence which are later.

Overall the analysis is a thorough attempt to reconcile relative with 'absolute' dates. Previous work on these lines I have seen has been hampered by a paucity of dates, and is not always applied 'in anger'. From the perspective of these notes Chapters 6 and 7 make considerable use of statistics and are central to the book. By common consent they are also a very 'difficult read', particulary for non-

statisticians; Baxter (2014a, b) provides a commnentary that attempts to separate the essential material from the considerable and less essential detail .

9.6 R Notes

Figures 9.1 to 9.5

Coding is covered in earlier chapters. Only a brief note on the basic code for Figure 9.1 is shown. The `ca` and `MASS` packages are needed.

In Table B.13 numbers are in the form of percentages; the raw data used for the CA was in the form of EVEs which. If a table similar to that shown is the source available conversion back to the original data may be required. The subset used in the initial analysis is extracted to the data set `data13`. This has seven columns, six of percentages for the vessel types, with total EVEs in the final column. To convert to EVES use

```
data <- data13[, 1:6] * data13[, 7]/100
```

```
zr <- ca(data)$rowcoord; xr <- zr[,1]; yr <- zr[,2]
eqscplot(xr,yr)

zc <- ca(data)$colcoord; xc <- zc[,1]; yc <- zc[,2]
eqscplot(x,y)
arrows(0, 0, x*.85, y*.85, code = 2, length = .15)
```